Addis Ababa University

School of Graduate Studies

IT Doctoral Program

Integrated Information Architecture in Support of Road Safety Organizations: The Case of Ethiopia

Tibebe Beshah Tesema

A Dissertation Submitted to the IS Track

Presented in Fulfilment of the Requirements of Degree of Doctor of Philosophy in Information Technology (Information Systems)

Addis Ababa, Ethiopia

July 2013

Addis Ababa University

School of Graduate Studies

This is to certify that the thesis prepared by Tibebe Beshah, entitled: *Integrated Information Architecture in support of Road Safety Organizations: The Case of Ethiopia* and submitted in fulfilment of the requirement for the Degree of Doctor of Philosophy in Information Technology (Information Systems) complies with the regulation of the University and meets the accepted standards with respect to originality and quality.

Signed by the examining Committee:

Examiner Assoc. Prof. Alemayehu Molla Signature ________ Date 9/10/2013

Examiner Dida Midekso Signature ________ Date 13/10/2013

Advisor Prof (Dr.) Ajith Abraham Signature ________ Date October 09, 2013

Advisor Dejene Ejigu (PhD) Signature ________ Date 15 October 201

Salehu Anteneh, PhD ________ 21/10/20

Coordinator of the Research Track

ABSTRACT

Integrated Information Architecture in Support of Road Safety Organizations: The Case of Ethiopia

Tibebe Beshah

Addis Ababa University, 2013

Safety in general and road safety in particular, is one of the major national agendas of any nation. Road safety can be affected by different factors road accidents being the main. Thus addressing issues pertinent to road accident will improve road safety. Road safety strategies include accident prevention, reduction and mitigation, which among other things, depend on accident data collection and analysis. Proper accident information management plays significant role in addressing the strategic issues effectively and efficiently. Various research works have been done in the area from different perspectives. This work investigated the use of enterprise Information Architecture as a way of visualizing and describing the various information assets and their interaction within road safety organizations and other stakeholders. It also explained a road safety situation based on data and defined an empirically grounded architecture as an approach to the management of road safety information to support effective information utilization and management at road safety organizations.

The research followed a Design Science research paradigm employing specific methods, including qualitative data analysis, data mining techniques, and critical literature review. Qualitative approach is used to determine information requirements of the road safety domain, while data mining technology is used to support in determining information requirement as well as accident data analysis in explaining road safety situation through identification of determinant factors. Critical literature review is used to define theoretical background and relate the result to the knowledge base in the area. In connection to this, while Zachman Framework was used to guide the development of the Road Safety Information Architecture, the popular data mining model, Cross-Industry Standard Process for Data Mining (CRISP), has been utilized during the data mining experiments to identify important patterns in accident dataset.

Results have shown that the architectural representation guided by the selected framework provides a holistic view to the management of road safety data. This Information Architecture can serve as a strategic guide to the review and development of the road safety data collection and analysis systems. The development of Road Safety Information Architecture (RSIA) in an enterprise view in the study and design of a road safety information management is an original contribution, which improves and expands the conceptual framework of the research in this field. RSIA is represented in six dimensions (content, process, motivation, people, place and time) from three perspectives (contextual, business and system). Moreover, in addition to identification of accident information requirement, the identified patterns through data mining have proved to supplement the previous severity prediction with promising performance.

Evaluation of the research showed that both the process and result of this research are valid and acceptable. The result of this research, mainly the RSIA, will help road safety organizations to revisit their focus of attention in crafting and implementing measures to reduce road safety problems. Finally, result of the research can also be used as a hypothesis and/or can be replicated to other developing countries with similar context.

Keywords- RSIA, Road Safety, Road Accident, Data Mining, Information Architecture (IA), Zachman Framework

ACKNOWLEDGEMENT

First and for most I would like to thank my supervisors Prof. Ajith Abraham and Dr. Dejene Ejigu for their constructive comments and invaluable support throughout the research. Their ability to drive my progress through productive and encouraging comments is highly appreciated. I also wish to thank Prof. Václav Snášel, Dr. Pavel Krömer and Dr. Salehu Anteneh for their support during the progress of this research.

I am very grateful to the management and staff of Ethiopian Road Safety Council, Road Safety Directorate and departments at a national and regional level, the Federal Police and Regional Traffic Offices. My special thanks goes to Commander Aklilu, Vice Commander Tesfaye, Ato Tilahun, Ato Abebe, W/ro Azeb, Ato Admasu, Ato Dereje, Ato Biniyam, Ato Tesfalem, Ato Daniel, Sajin Asefa, Sajin Atalel, Ato Yeneabat, Commander Seid, Commander Dereje, Ato Goitom for their active participation and constant support in providing relevant information to my research work. Ato Thil and Ato Abinew also deserve immense appreciation for their support in the research.

My special thanks go to my beloved family Zewditu and Yostana, for their unreserved backing and encouragement. My dear Zed deserves special thanks for her constant love, prop, understanding and faith in me. I am also very much indebted to my brothers Fekade and Betemariam who stood beside me during this critical time.

I am also grateful to my class mates and colleagues. Special thanks to my friends Zelalem, Yihun, Dr. Fekade, Miftah, Dr. Million, Nassir, Mesfine, Habtamu, Kibrom, Mathias, Yeshinegus, and Ayalew for their unreserved help; they often had a solution to the problems I faced. They have been there whenever I needed someone.

Above all, I thank Almighty God who surrounded me with so many wonderful people.

Tibebe Beshah

Dedicated

to

Alganesh Woldu (Agashe)

Table of Contents

List of Abbreviations

ADT	Average Daily Traffic
CART	Classification and Regression Tree
CRISP-DM	Cross Industry Standard Process For Data Mining
DSR	Design Science Research
EIA	Enterprise Information Architecture
IA	Information Architecture
IS	Information Systems
MARS	Multivariate Adaptive Regression Splines
MVA	Motor Vehicle Accident
MVC	Motor Vehicle Collision
RSIA	Road Safety Information Architecture
RTA	Road Traffic Accident
RTC	Road Traffic Collision
RTI	Road Traffic Injury
UML	Unified modelling language
WHO	World Health Organization

List of Figures

List of Tables

Chapter One

INTRODUCTION

Since the first recorded road accident fatality in 1896, road accident has been a serious problem for almost every country (Zheng, 2007). Consequently, it has led to ensuing attempts to research road safety issues and address various pertinent issues from diverse perspectives. Finding a way to effectively collect, analyze and disseminate accident data is critical in improving the learning process, thereby improving road safety. To this end, a notable concept, Information Architecture (IA), involves describing and modeling how information should be organized in such a way that it serves as a foundation for an informed decision. This research contemplates on the concepts of severity analysis of accidents and enterprise IA in road safety domain.

This Chapter introduces the essence of the research and places it in context. Thus, the aim of the chapter is to discuss background, problem statement and motivation, objectives, and significance of the research. Definition of central concepts to the research work is also presented in the chapter.

1.1 Background of the Research

The transportation sector not only forms a complex system with many opportunities for economic efficiency and competitiveness, but also includes many externalities like accidents, congestions and pollutions (Borzacchiello et al., 2009). Among all transportation systems, road transport is the major one - at least in the context of developing economies like Ethiopia. This has also necessitated for a growing interest in research on transportation information system to increase safety through traffic management, information systems and automation (Baguley, 2001, Bobeva, 2005, Abugessaisa, 2008a, Hall, 1995).

Road Traffic Safety or simply Road Safety is safety on the roads - a notion that means to safely drive on road to ensure that there is no harm done. In other words, it can be said that road traffic safety stands for reduction of causes of accident on roads by facilitating proper driving. Though road safety can be affected by different factors like congestion and pollution, it is highly associated with Road Traffic Accidents (RTA). RTA is also defined as any vehicle accident occurring on a road (i.e. originating, terminating, or involving a vehicle partially on the road). A commonly accepted definition states that it is "A rare, random, multi-factor event which is always preceded by a situation in which one or more road users have failed to cope with their environment" (Baguley, 2001).

Many different terms are commonly used in the literature to describe RTA. The World Health Organization uses the term road traffic injury (RTI), while a motor vehicle accident (MVA) is used by the U.S. Census Bureau. Other terms that are commonly used include auto accident, car accident, motor vehicle traffic collision, car crash, car smash, motor vehicle collision (MVC), road accident, road traffic collision (RTC), road traffic incident (RTI), and road traffic accident. Given the high level association between the concepts of road safety and road traffic accident, both phrases are used contextually as required throughout this research work.

Road safety is a multi-dimensional issue at the top list of any nation's concern. In line with this, the United Nations labeled the years between 2011 – 2020, as road safety decade. It is also stated that road safety is a prime concern to individuals, groups and organizations, all of which requiring data and evidence about accidents (Peden et al., 2004, Moon, 2003). RTAs are considered to be a major threat to road safety and are placed among the top leading causes of deaths and injuries of various levels.

According to Bener et al. (2003), RTAs continue to be major causes of mortality and morbidity. Globally, RTAs result in an estimated 1.2 million deaths and a further 50 million injuries per year (WHO, 2013). It is estimated that 90% of road traffic deaths occur in low and middle income countries (WHO, 2009). It is also predicted to rise by 27% in low and middle income countries while decreasing among high-income countries by 83% by 2020, which will further augment the global gap (Hyder and Bishai, 2012); (WHO, 2009)

The cost of fatalities and injuries due to such RTAs has a great impact on the wellbeing and socio-economic development of a society. In connection to this, it is stated that RTA is an escalating problem in Africa and deserves proportionate research efforts (Lagarde, 2007). Road accidents have been estimated to cost low and middle income countries an estimated USD100 billion every year (Hyder and Bishai, 2012). And Ethiopia is not an exception. Ethiopia is one of the countries in the world experiencing highest rate of such accidents, resulting in fatalities and various levels of injuries. As each day passes the extent of loss of life, property and economy of, for example, Addis Ababa due to RTA has reached a critical level (AACG, 2009). Similarly, the rate of traffic accidents and pollution in Addis Ababa goes up together with the increase of motor vehicles and population size (Samson, 2006).

RTAs are a major public health concerns causing thousands of injuries and premature deaths each year. According to Mathers et al. (2008), projections also show that RTIs will be one of the three leading contributors to the global burden of fatalities over the next two decades.

Even though considerable improvements in road safety have been made, the number of accidents remains high in Ethiopia. According to a recent report by the Office of the Federal Police of Ethiopia, the traffic accidents in Ethiopia claim 2,200 deaths, 8,000 injuries and a loss of over half a billion Birr[1]due to the destruction of property annually. However, the true RTA fatality rate in Ethiopia is greater than what is actually shown by the official statistics. According to National Road Safety Directorate of Ethiopia, it is estimated to be near 70 fatalities per 10,000 vehicles. The

[1] Birr is a local currency in Ethiopia with current exchange rate of 1USD:18.5 Birr

problem is serious even compared to Kenya (one of the developing countries with many similarities to Ethiopia) and United Kingdom (one of the developed nations), where the figure is about 19 and 2 fatalities per 10,000 vehicles, respectively (Persson, 2008). Thus, it is easy to learn the magnitude of the problem even from this simple comparison and the WHO (2013) statistics which indicates the overall global road traffic fatality rate as 18 per 100 000 population.

The above facts correspond to the discussion in the literature, which indicates that many countries in Africa, including Ethiopia, have extremely high accident rates (Peden et al., 2004). Based on a detailed analysis, the WHO made predictions of the situation for the year 2020 and the various changes in a rank order of causes of death and disability throughout the world as shown in Figure 1. RTAs are predicted to rise to third place (after ischemic heart disease and unipolar major depression). Thus, the need to attempt to prevent road traffic accidents using various approaches and means is increasingly important.

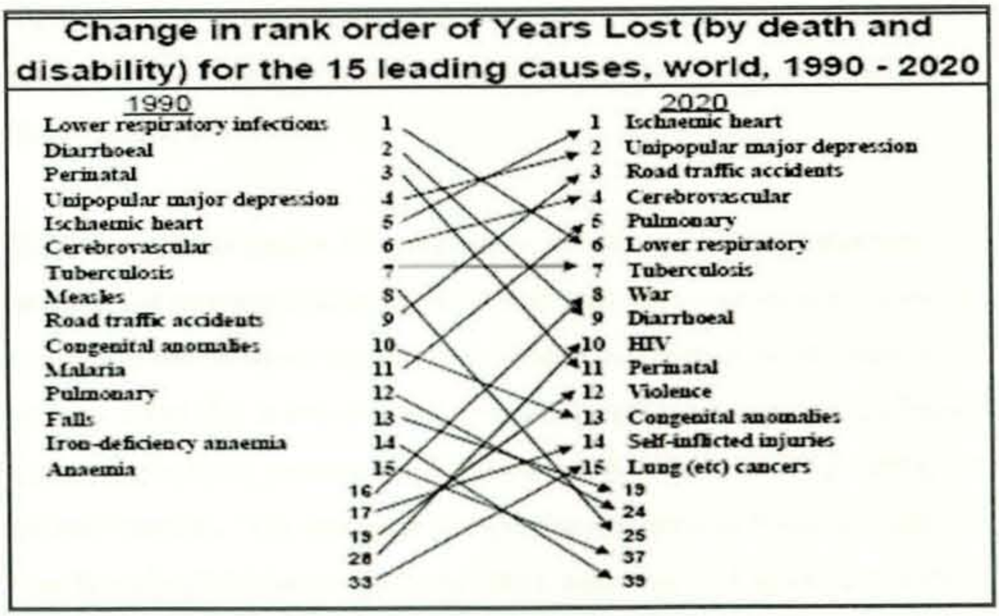


Figure 1: Rank Order Change of Causes of Death and Disability worldwide (Baguley, 2001)

Indicating the magnitude of the problem, it is also said that RTAs are also complicated to analyze as they cross the boundaries of engineering, geography and human behavior, hence becoming very complex events (Sabel et al., 2005). RTAs are very complex events (Vorko-Jović et al., 2006, Sabel et al., 2005). Nonetheless, RTAs are predictable and, therefore, preventable, but it needs close coordination and collaboration through a holistic and integrated approach. In addition the anticipated coordination and road safety policies must be based on evidence as mentioned in (Mohan and Roberts, 2001).

As to the distribution of the problem, although road traffic accidents are major global problems, most of them occur in low and middle-income countries (Persson, 2008). Though, road traffic accidents are global phenomena, the numerous success stories of the developed nations cannot be simply imported and implemented in developing countries and, as a result, tackling the problem requires context-based innovative research (Davis et al., 2003). Developing countries tend to seek assistance from consultants from developed countries, but it is important to recognize that the traffic engineering and safety solutions that work in one country may not necessarily function well in another with totally different transportation systems characteristics (Kar and Datta, 2009).

The scenario is not similar. For example, in developing countries pedestrians and passengers of commercial vehicles are the most vulnerable while drivers of privately owned vehicles are more exposed in developed countries. It is also reported in (Persson, 2008), that in one crash the number of people killed or injured is 30 times higher in developing countries as compared to developed ones. Highlighting on marked disparities, "The challenges of improving road safety in Boston are different from Bangalore," (Mohan and Roberts, 2001). According to (Lagarde, 2007), there is a need to urgently scale up surveillance and research efforts in developing countries in order to determine how to build on these results, taking regional specific issues into account.

Road safety problems can be addressed from different perspectives like education, enforcement, and engineering, all of which requiring learning from previous patterns which in turn is based on data. In connection with this, through a long history of recording, the size of organizational data in business and public sector organizations is getting enormously bigger. Road traffic safety system is one of the various areas where critical data about the well-being of the society is recorded and kept. Various aspects of a traffic system like vehicle accidents, traffic volumes and concentration are recorded at different levels (i.e., huge amount of road safety data/information is recorded every day). Obviously, the reliability, completeness and accuracy of the information and the level of analysis will affect a lot of decision-making and the organization itself, and this is where the need for IA emerges. The IA approach will provide a platform for effective decisions by enabling learning from data. "The information architecture approach focuses, on information followed by technology, therefore clearly making the point that the use and value of information is what provides the competitive resource not the system supporting its use"(Doran, 2007)

IA is an emerging discipline and catches the eyes of IT professionals (Yunus and Rahman, 2008). It is a relatively new discipline that focuses specifically on designing the information itself. Information architecture refers to the information content of a system with a due emphasis on the structure (Robbins et al., 2011). The tools, methods and models, and frameworks are still in the gray area in the IT world. IA involves the design of organizations and navigation systems that support usability and findability (Morville and Rosenfeld, 2006). It includes the logical design for a specific system or a set of systems that involves business models, organizational models, checklists, process models and data models.

The use of the term "information architecture" should not be confused with the way in which a website is designed or the details required in putting together enterprise technology architecture (i.e. components such as servers, networks and desktops). Rather it denotes a high-level detail of information flows and processing within or between applications or systems (Van der Walt and Du Toit, 2007). Through IA, the information is better constructed, organized and maintained. It is much easier for organizations to facilitate communication, update their data and lower the cost of redesigning their system as years go by (Yunus and Rahman, 2008). By putting in place effective information architecture as supporting framework for organizational strategy, it is believed that the risk of information overload would be minimized and could allow for making maximum use of hidden knowledge through efficient analysis of the data within a domain. It will also serve as a platform for effective information dissemination and sharing among stakeholders (Van der Walt and Du Toit, 2007).

Axelrod (2010), in his paper "Quality Data through Enterprise Information Architecture", stated that Architecture is one of the most used (and abused) terms in the areas of software and systems engineering. According to the paper, to get a good feel for the complexity of the systems-architecture topic, it suffices to list some of the most commonly used concepts with "architecture". Further noting, the author mentions some as enterprise; data; application; systems; infrastructure; Zachman; information; business; network; security; and service-oriented architecture (SOA) which he labeled as the "latest silver bullet". This research contemplates on the concept of enterprise information architecture. According to Malloy (2010), an Enterprise Information Architecture (EIA) provides a means to describe and manage information consistently so that it can be accessed, understood, compared, shared and composed in a coordinated and integrated manner across the enterprise at every levels of the hierarchy.

1.2 Statement of the Problem and Motivation

1.2.1 Motivation

The prime motivating factors for this research include the magnitude of the problem across the world in general, and in developing countries in particular. Road accidents as the major factors for road safety problem need investigations from various perspectives. The capability and use of Information Architecture to address the problem of road safety from data/information management perspective is another factor, which is also a novel approach.

In addition, the versatile application of data mining in the analysis and interpretation of cumulated data is worth mentioning. Sensing the absence of a foundational framework in a road safety information management in a local context is among the motivations for this research work though there were attempts dealing with an aspect of road safety data.

Thus, motivated by dearth of a systematic view in accident analysis in a road safety domain and a high-level architectural guide line from integrated enterprise perspective in the road safety information management domain in Ethiopian context, this research focuses on identifying determinant factors for accidents and investigating the potential use of enterprise information architecture as an approach to the management of road safety data.

1.2.2 Problem Statement

Due to the complexity and high traffic on roads, risk of road safety as a major concern is increasing. Road safety can be affected by different factors. Road traffic accident (RTA) being the major, the others are congestion and pollution. It is evident that addressing road accident problems will improve road safety situations. At the centre of any road safety strategy or effort, there is accident data collection and analysis. Road safety is best achieved when all the key groups identified share the same culture about the issue. To this end, reliable accident information and its effective is a prerequisite, as every organization needs to be cognizant of facts on the ground and be information intelligent. The ability to successfully search, assemble, analyze, and effectively use all relevant available information for any decision and initiative is what we call information intelligence (Evgeniou and Cartwright, 2005).

Among the many research related needs for road safety, the most pressing is a means to better collect and analyze data, so as to enable more reliable estimates to be made of the global burden of RTAs, especially in low-income countries (Peden et al., 2004). Current road safety research is characterized for broadly using traffic accident data collected by different bodies with the main objective of producing official statistics and guiding road safety policies and interventions. However, the use of these records for research purpose shows serious limitations, of which the prime cause lies in the instruments and procedures used. In this context, it is reported that a redesigning and modernization of traffic accident data collection and analysis systems are still being considered even in the context of developed nations like Spain (Tormo et al., 2009).

In relation to this, most low-income countries are unlikely to have national injury surveillance systems for several decades. Thus, national estimates of the burden of injuries should be built by collating information from all existing information sources by appropriately correcting for source specific shortcomings (Bhalla et al., 2009). This is because police reports would underestimate the magnitude of vehicle injuries and distort any evaluations of preventive initiatives.

Data completeness and accuracy are also identified as major challenges to make data analysis easier for comparison of data from different countries and to contribute to the general quality of data (Abugessaisa, 2008b). Thus, the most serious concern in a road safety research is how to obtain adequate data about accidents in an effective manner (Larsson et al., 2010). In many low-income countries, systematic efforts to collect road traffic data are not well developed and under reporting of deaths and serious injuries is common (Peden et al., 2004). Bener (2003), also emphasized that lack of reliable accident data is a serious problem in most of the developing countries while Stoop and Thissen (1997) note that biases in information collection due to many reasons may lead to misinterpretation. In this regard, it is imperative to note that the accuracy and validity of decisions regarding traffic safety depends on how well the information is collected and shared, and the quality of the compiled general data (Abugessaisa, 2008a).

It is also argued in the literature that solving the global road safety problem needs innovative research and practical measures (Mohan and Roberts, 2001). Effective research on road safety issues is an essential pre-requisite for better understanding of the problem, thereby providing the framework against which effective policies and counter-measures can be developed (ADB, 1996). Detailed investigation in to accident analysis is essential and will require inter-sectorial collaboration among traffic office, as well as health, law and transport authorities (Bener et al., 2003). The most complete information about causalities in road traffic accident can be obtained from interlinked police, hospitals, and death records (Vorko-Jović et al., 2006). With improved collaboration and management of data, significant reduction can be achieved in the magnitude of road traffic casualties (Peden et al., 2004).

The basic entities that exist and form the information domains in the area of road safety are humans, road and vehicles (Abugessaisa, 2008a). A rapid and significant reduction of road safety risk can be achieved through a scientific approach to the problem; the provision, careful accident analysis and interpretation of good quality data; institutional cooperation across sectors; and the creation of national and regional research capacity (Peden et al., 2004). In line with this, proper information management on road crashes is among the areas of concern identified by the National Road Safety Authority of Israel (NRSA, 2006).

As stated by Persson (2008), the true RTAs fatality rate in Ethiopia is greater than what the official statistics actually show. Therefore, in Ethiopia, there is a high tendency for underreporting, as it is also the case in many low income countries. However, such case is not limited to such countries; this crucial problem is actually common all over the world. In the urban areas, although traffic police and hospitals are available, accident victims are usually evacuated by bystanders who lack the necessary skills and/or equipment in pre-hospital care. As a result, many casualties such as these are not reported. Hence, improving the police report system and establishing a more comprehensive trauma registry that would include data from hospitals and insurance companies should be implemented (Nakahara and Wakai, 2001). This is because lack of harmonization of data between countries, regions and sectors limit the capability of analysis and effectiveness of measures (WHO, 2009). In line with this, one of the major recommendations from African Road Safety Conference held in Accra, Ghana, in 2007, focuses on the accident data and stipulated that:

> "*Data collection is necessary for road safety management but only reliable data can be useful for planning purposes. Accordingly, efforts should be made to address the non-reporting of accidents and to harmonize data that originate from different sources*" (Afukaar, 2007).

Improving data linkages among police, transport and health sectors along with increasing human capacity to undertake data collection, analysis and interpretation are among the recommendations made by different road safety researches and reviews in (WHO, 2009), which could be guided by road safety information architecture.. Likewise, in the Ethiopian context, promoting the need to strengthen the organizational setup of the traffic police and assisting with equipment, training and accident data improvement system are indicated as vital actions in the coming years (NRSC, 2011). Thus, this research is in conformity with the national strategy.

Thus, addressing problems related to accident data collection like incompleteness and underreporting, and accident analysis require architectural guideline. Architecture can be viewed as a strategic plan that defines how an issue of concern is performed and controlled (Hall, 1995). Architecture is something that resides at a higher level than a design, such that while it remains constant, many different designs can conform to it (Bures, 2009). It is a structure and/or guideline as to how things are related. It helps to facilitate communication and describe a system. A system can be software based system, an enterprise or any other system of interest. As one of the concepts related to architecture, IA concerns organization, labeling, navigating in and search for information (Dong and Agogino, 2001).

Information Architecture (IA) principles aim to create an environment with a logical structure that help users to find answers and complete their tasks (Rosenfeld and Morville, 2002). Contextually well defined methods in the field of information architecture, information sharing and web technologies are needed for road safety experts and practitioners (Abugessaisa, 2008a). It is the researcher's belief that effective road traffic accident data/information management will improve road safety as it facilitates road accident information intelligence and provide a way to improve decision-making. This is possible by providing an integrated way of managing accident information, which includes determining the content, structure, process, timing, and participation right from the collection to its dissemination.

In the literature, as discussed in Chapter Three and Four, there are some attempts at both local and international levels, which deal with accident data collection and analysis. Each attempt treats specific aspects of road safety data and presents a specific structure that will be suitable for a specific context. However, practically all of these attempts present a common gap: lack of systemic view of road safety data/information treatment (collection, analysis, dissemination, etc.). Major problems include disintegration, less attention to human factors and variations of determinant factors in accident analysis. This limitation is due to the absence of sound information architecture, methodological steps and technological solution that enable wider view of information management within an enterprise framework. This fact is even more important in developing countries with limited resource.

More specifically, the research hypothesizes that it is not possible to develop an efficient means to improve safety without constructing appropriate platform for effective reporting, storage, analysis and dissemination of road safety information. There is an underlying methodological problem with the collection, analysis and dissemination of road safety information (Abugessaisa, 2008b). This is exhibited with the less attention in researches in a local context to investigate human factors for accident occurrence and severity except drivers. The success of traffic safety and highway improvement programs hinges on the analysis of accurate and reliable traffic accident data (Tiglaco, 1998)

Thus, proper handling and management of the road safety issues depend on better understanding of the situation from the existing cumulated data. Defining well structured information architecture for the collection and analysis of accident data is of paramount importance in order to create conducive environment for understanding road safety situations and taking appropriate measures.

Insight into the effectiveness of injury-reduction technologies, policies, and regulations require a more detailed empirical assessment of the complex interactions that vehicle, roadway, and human factors have on resulting crash-injury severities (Savolainen et al., 2011). As learned from the literature there were attempts to describe and address road safety problems both worldwide and in Ethiopian context. Descriptive analysis of the magnitude and situation of road safety in general and road accidents in particular is important, but understanding of data quality, factors related with dangerous situations and different interesting patterns in a data, is of even greater importance. This gives a big picture of the scenario instead of a fragmented effort to address an aspect of it.

This research is inspired by previous works on the domain and the versatile applicability of machine-learning paradigm and information architecture, which will add on the ongoing effort of reducing road safety problem by providing systemic view. To the researcher's knowledge, apart from fragmented pieces of road safety researches, there is no foundational work to address road safety problem from integrated and enterprise-wide road accident information management point of view in developing countries context to date. This is supported by Lagarde's (2007) argument which states, as compared to the magnitude of the problem, that research in road safety in developing countries is scarce, especially in Africa. Therefore, this work is unique in the dimension it tries to address the problem, machine-learning approaches used, comprehensiveness and the actual observation of the road safety related issues. It is believed that identifying determinant factors for accident occurrence and severity enables road safety experts to make sound decisions.

The desired outcome of this research is to lay a foundation, in the effort of understanding the road safety information complex as an integrated and open system, which then leads to its effective management by developing appropriate information architecture for a road safety system in Ethiopia. This work aims investigated the use of enterprise information architecture as a way of visualizing and describing the various information assets and the interaction of these assets within road safety organizations. In a nutshell, this research work focuses on identifying optimal attributes that determines accident occurrence in explaining road safety situations and developing enterprise information architecture as an approach to the management of road safety information. This is novel in that it draws an integrated enterprise perspective into the management of accident information.

Research Questions

The purpose of this study is to explain a road safety situation and define an empirically grounded architecture of road safety information. The research sought to explain the road safety situation and found a way for improving road accident information management practices. In particular, the study aims to address the following research questions:

- RQ1- What does the road safety situation in Ethiopia look like?
- RQ2- What appropriate analytical models can be developed for analysis of road traffic accident data?
- RQ3- What are the problems of the current road traffic accident data collection, analysis and dissemination practice on the road safety?
- RQ4- How should road traffic accident data reporting, analysis and dissemination tasks be organized under a common architecture?

1.3 Objective of the Research

1.3.1 General Objective

The general objective of the research is to explain the road safety situation in Ethiopia and develop information architecture for road accident information in support of effective information utilization and management at road safety organizations.

1.3.2 Specific Objectives

In order to satisfy the above-mentioned general objective, this research aims at accomplishing the following specific objectives:

a. To explore the status of road safety situation in Ethiopia,
    i. To explain and predict the role of road users' and collision related factors on possible injury risks
    ii. To explore accident analysis trend on factors affecting accident severity
b. To construct analytical machine learning models in the analysis of road traffic accident data,

c. To examine problems related to accident data reporting, data quality and analysis mechanisms in a road safety domain,
d. To identify the structure and requirement of road safety data collection and analysis focusing on road accidents,
e. To investigate a way to define information architecture based on enterprise architecture framework to establish road safety information management,
f. To define a road safety information architecture that will facilitate effective utilization and management of road traffic accident information among road safety organizations, and
g. To evaluate the research process, findings and the resulting road safety information architecture.

1.4 Significance of the Research

Road safety is a critical issue in a society. One of the significant factors that affect road safety is road traffic accident; thus, studies on the data collection and analysis of these road traffic accidents have paramount importance. In this research, an attempt is made to conduct trend analysis, explore road safety situations, and develop integrated information architecture in support of road safety information management. In line with this, the major contribution of this research will be explanations of road safety situations and the design of an integrated architecture for effective road traffic accident information management and utilization among road safety organizations to support informed decision-making and road safety strategies. The information architecture established in the road safety organizations forms the basis of support in delivering future information requirements for the road safety domain.

So far, different approaches like The Haden's Matrix, Vision zero and The Road User approaches were used in addressing some of the problems associated with road safety. Through the use of data oriented approach to road safety problems, this research identify additional insights to the problem, which apply to low-income countries, by extending the existing conceptualization. It is also worth mentioning the different views and theories that will result from such type of study in different context.

Although the study is aimed at addressing road safety in particular, the output of the study can be used as a source of methodological approach for studies dealing with the application of information architecture and standards on other societal problems. In addition to providing an invaluable perspective for road traffic accident data collection and analysis, it will subsequently improve the efficiency of a national data collection system.

The outcome of the study contributes to the body of knowledge on information architecture, road safety, road accident information analysis and management, and might (a) help researchers and road safety organizational leaders base their strategies and plan of road safety on a complete and reliable information guided by a wider view of the problem, and (b) provide empirical evidence on different aspects of road safety situation, allowing for better understanding of the problem at hand. An additional vital contribution is made by the introduction of the enterprise approach for the study of the road accident data management systems.

Thus, results of this research add on the efforts to understand the road safety situation in Ethiopia, evaluate the existing road safety data system and/or design a new one.

1.5 Scope of the Research

The scope of the research is limited to exploring the existing situation of road safety, road accident information management practices and developing information architecture with the aim of laying down a foundational framework for road traffic accident information management. This research would have been more complete if it had included traffic congestion and pollution as well. However, for the sake of manageability and focus, this research is limited to road traffic accident as a major threat to road safety at three selected regions of Ethiopia: Addis Ababa City Administration, Amhara and Gambela. In general, the research will focus only on road traffic accident information management.

1.6 Definition of Working Concepts

Understanding of the ideas and contributions of this research work is grounded in understanding of the key terms and concepts it operates with. To this end, working definitions and explanations of the keywords used throughout the study are presented below.

Road Safety

The concepts 'road safety' and 'road accident' are often used interchangeably in this research, although strictly speaking, road safety problems include not only road accident but also congestion and pollution, whilst road accident refers only to accidents on the road. Where possible, the choice of which term to use is determined by the preferences of the cited/referenced author(s). For good understanding of the issue at hand the definition by Elvik et al., (2006) that illustrates the concept well is forwarded as follows. The authors attest that *"A road safety problem is any factor that contributes to the occurrence of accidents or the severity of injuries."* On the other hand regarding the objectives of road safety analysis, the same authors state that it is the identification of those problems that make the greatest contribution to accidents or injuries and that are amenable to treatment (Elvik et al., 2006)

Framework

A framework helps people to organize and assess integrated models of their enterprises. This organization helps ensure interoperability of systems and control the cost of developing systems (Bahill et al., 2006). The three definitions below best illustrate the meaning of the term 'Framework':

> *"A classification structure for descriptive representation of an object, any object. An object could be an enterprise, an information system, etc..." (Zachman, 2003)*

> *"A systematic taxonomy of concepts and their interrelations" (Zachman, 1987, Sowa and Zachman, 1992).*

The Interoperability Clearing House (2004) also forwarded the following definition: *"A logical structure for classifying and organizing complex information" (Bobeva, 2005).*

Architecture Framework

There are various conceptualizations of architectural frameworks. Some focus at a software level while others try to view at the organizational level. It is also stated that most of the classical enterprise architecture frameworks presented focus on the software architecture, rather than on the total enterprise architecture (Goethals-SAP-leerstoel, 2004). A comprehensive definition of the concept 'architecture framework' is presented by The Open Group (2002) and, is adopted here as follows:

> *"An architecture framework is a tool which can be used for developing a broad range of different architectures. It should describe a method for designing information system in terms of a set of building blocks, and for showing how the building blocks fit together. It should contain a set of tools and provide a common vocabulary. It should also include a list of recommended standards and compliant products that can be used to implement the building blocks" (Goethals-SAP-leerstoel, 2004, Bobeva, 2005).*

Information Architecture (IA)

As the concept is relatively new to both the researchers and practitioners in the area, its definition is still in a gray area. It is noted in (Rosenfeld, 2002), that there is no widely accepted definition. Three definitions are selected here for their comprehensiveness and the fact that they reflect the context of the concept in this research.

> *"IA generally represents a higher level of abstraction, emphasizing an awareness of systems in terms of how critical subcomponents may interact according to semantic aspects of processes, designs, and metrics" (Robbins et al., 2011)*

> *"IA is a professional practice and field of studies focused on solving the basic problems of accessing, and using, the vast amounts of information available today"(Resmini and Rosati, 2011)*

> *Evernden (2002) also put forward the following definition "Information architecture is the foundation for managing information in general as a*

corporate resource. It describes the theory, principles, guidelines, standards, conventions and dimensions that are necessary to design an effective management framework for information. Its purpose is to design information structures that help people to use information in effective, productive and innovative ways" (Bobeva, 2005).

1.7 Structure and Organization of the Thesis

The remaining part of the dissertation is organized as follows. Chapter Two is dedicated to provide details on the research approach and methodology. In Chapter Three, review of literature pertinent to the central theme of the research and conceptual framework is presented while the fourth chapter discuses related works. Chapter Five and Six are dedicated to the presentation of the findings and the resulting Road Safety Information Architecture along with discussions, while evaluation and validation of the results are focuses of Chapter Seven. The last Chapter presents conclusion, contributions and directions for future work.

Chapter Two

RESEARCH METHODOLOGY

The purpose of this Chapter is to present and discuss the approach and methods of the research. Hence, it covers the methodological aspects that have guided the present work. The Chapter starts with general approach. The qualitative approaches of this research are described next, followed by, the machine learning techniques which are mainly used to determine the information requirement of the accident information management process.. The chapter concludes by the presentation of the methodology used for evaluation of the research process and results.

2.1 General Approach and Theoretical Frameworks

2.1.1 General Approach

According to Hevner et al., (2004), two paradigms characterize researches in the Information Systems discipline: behavioral science and design science. The behavioral science paradigm seeks to develop and verify theories that explain or predict human or organizational behavior. The design science paradigm seeks to extend the boundaries of human and organizational capabilities by creating new and innovative artifacts (Offermann et al., 2009, Hevner, 2007). In discussing IS research trends over the past decade it is claimed in (Farkas and Weistroffer, 2012) that "IS research trend is identified as improving the value of IS artifacts for the organization either by tailoring process or improving communication, e.g improved conceptual models, broader applicability of architecture artifacts, etc." *"Design Science research in Information System is important as the dominant behavioral science paradigm is not sufficient for addressing the types of problems that call for human creativity and innovative and novel solutions" (Carcary, 2011)*

This research combines literature review, machine learning and a qualitative research approach, in a Design Science Research Paradigm, in order to explain road safety situations and defines information architecture for road safety information management. Literature on road safety, machine learning and information architecture provides an excellent insight to the concept, evolution, need, guiding frameworks, use and application of information architecture. Machine learning techniques are used to investigate the road safety situations through identification of patterns. A qualitative research approach is used to define road safety information management requirements.

According to Creswell (2009), qualitative research is a means for exploring and understanding the meaning individuals or groups ascribe to problem under study. It is employed in this research as it enables to bring the researchers to an understanding of a complex issue or object and can extend experience or add strength to what is already known through previous research. It is used to identify the information requirement of the road safety domain and define enterprise wide information architecture. This research is descriptive in method, because its main purpose is to explain and define road safety information architecture through conceptual modeling by utilizing road accident data collection and analysis practices. On the other hand, it is fundamental in purpose because it describes road safety information architecture based on theoretically genuine, contextual and empirical models.

Regarding the general process; with background, motivation, problem statement and objectives given, Figure 2 shows major phases of this work. As shown in the figure the research work commenced with domain understanding and problem formulation through literature review, which helped to create a good understanding of the research area. The next task was qualitative data collection and analysis, which was done in parallel with the analysis of road safety situations through machine learning techniques including trend analysis. In the next stage, the research follows and uses the output of the analysis in designing the meta-model of Road Safety Information Architecture and the subsequent detail architectural descriptions. Finally the research findings and architectural artifacts were evaluated based on appropriate validation techniques. Reflections on lessons learned, conclusions, recommendations and indication to future research are the last important tasks of the research process.

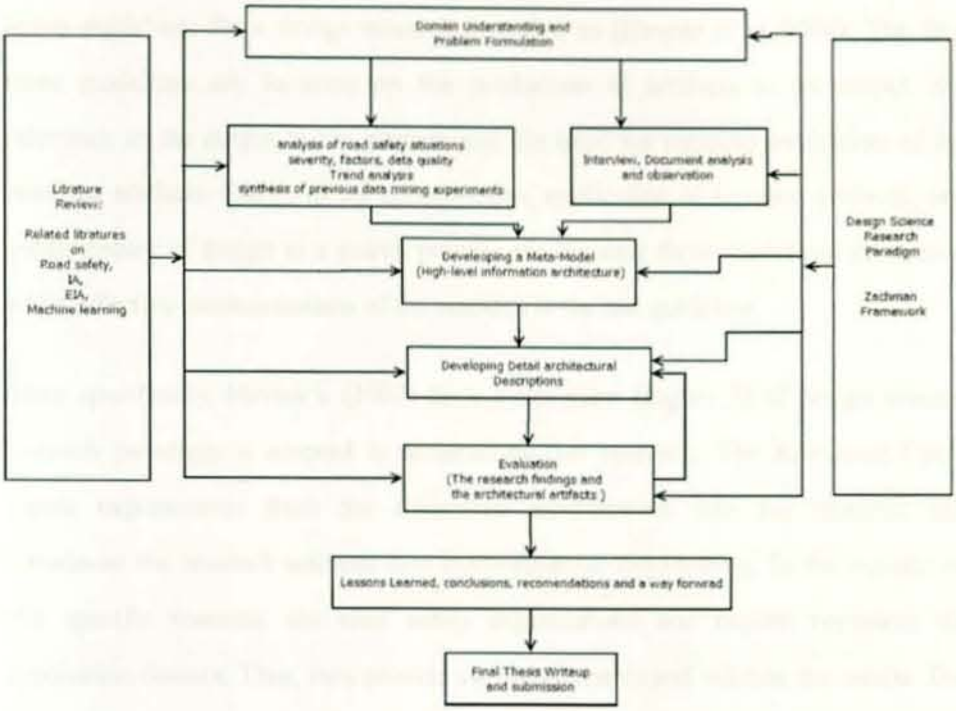


Figure 2: Conceptual model of the research process

2.1.2 Design Science Research Approach

Design Science research is centered on building and evaluating artifacts in order to solve organizational problems (Carcary, 2011). However, Design Science research in the field of Information system is not limited to IT artifacts in the form of computer based systems. According to (Carcary, 2011) and Venable (2006), artifacts or solution technologies include IS development methods, tools and techniques, IS security and risk management practices, and IS planning and management methods. It has to be noted that both behavioral and design science paradigms are foundational to the IS discipline, positioned as it is at the confluence of people, organizations, and technology.

As indicated above, considering its appropriateness to the research agenda of the current work, design science research approach was adopted for the design and evaluation of the proposed architecture. Accordingly, this research is in line with the seven guidelines for a design research provided in (Hevner et al.,2004). The first three guidelines are focusing on the production of artifacts as an output, the relevance of the output to the domain and the need for rigorous evaluation of the resulting artifacts. Clarity of the contributions, application of rigorous methods, and consideration of design as a search process are the next three guidelines mentioned while effective communication of the research is the last guideline

More specifically, Hevner's (2007) three cycle view (Figure 3) of design science research paradigm is adopted in undertaking this research. The Relevance Cycle inputs requirements from the contextual environment into the research and introduces the research artifacts into environmental field-testing. In the context of this specific research, the road safety organizations and experts represent the application domain. Thus, they provide the requirements and validate the results. The Rigor Cycle provides grounding theories and methods along with domain experience and expertise from the foundations knowledge base into the research and adds the new knowledge generated by the research to the growing knowledge base. In the context of this research, use of extensive literature in the study area and the guiding framework along with the process and resulting architectural description represent the knowledge base. The central Design Cycle supports a tighter loop of research activity for the construction and evaluation of design artifacts and processes (Hevner, 2007).

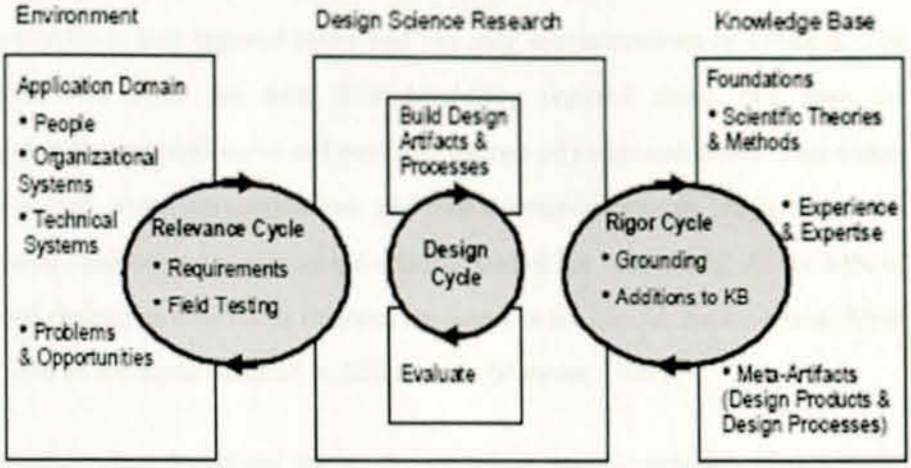


Figure 3: Design Science Research Cycles (Hevner, 2007)

As indicated in Chapter one the central theme of this research is the definition of information architecture for the road safety information management. In line with this, the available architectural frameworks were reviewed and an enterprise-based framework, Zachman Framework, was selected as a guiding framework for the development of the roads safety information architecture. Zachman Framework is selected for its comprehensiveness and wider applicability. While Zachman Framework is used as a way of visualizing the problem and solution domain, CRISP data mining model is also used in the process of data mining experiments to explain various aspects of road safety situations.

2.2 Qualitative Data Collection and Analysis Approaches

As indicated previously, the research is based on road safety departments and traffic office units (road safety organizations) in three selected regional administrations of Ethiopia. The detail of the process is presented in this section.

2.2.1 Study Population and Sampling

The unit of analysis is the road safety data system with road traffic accident data reporting, analysis and dissemination practice and activities being the phenomenon of inquiry. Accordingly, two regional states and one city administration were selected randomly from nine regional states and two city administrations in Ethiopia. The selection was made one from four developing regional states, one from the remaining five regional states and one from the two city administrations. This makes the research more comprehensive and representative, though majority of the registered vehicles and accidents are concentrated at the capital city. About 64% of the total registered vehicles in Ethiopia are found in the capital, Addis Ababa. More than 65% of accidents occurred in Addis Ababa (Persson, 2008).

Within the selected regional states, the sampling frame comprises all stakeholder organizations working on road safety data in Ethiopia. Research population includes road safety experts at the National Transport Authority, Regional Road Safety Department and Traffic office Units. With recognizable variations, literature in the qualitative research area suggests sample sizes ranging from 3 to 30 (Robson, 2002, Creswell, 1998, Marshall, 1996, Padgett, 2008). Though there appear no agreement on sample sizes, based on recommendations and discussions in qualitative research literature 17 respondents were selected from organizations in the area. The purposive convenience sampling was adopted, which is the most frequently used approach in qualitative studies (Burnard, 2004, Creswell, 2009, Kumar, 2005). These individuals were identified by the recommendation by executive management of respective offices and the researcher and highlighted as major contributor to the study.

2.2.2 Data Collection Techniques

In the effort of addressing the research problem, various empirical data regarding different road safety data system aspects needs to be collected. In doing so, there are different qualitative techniques of data collection. Qualitative data sources include observation and participant observation (fieldwork), interviews and questionnaires, documents and texts, and the researcher's impressions and reactions (Myers, 2009). Written data sources can include published and unpublished documents like reports and newspaper articles. In qualitative study, use of different data collection technique is highly recommended in order to get a full picture of the subject at hand, which is road safety in the context of this research.

Accordingly, the preferred data collection techniques for this specific research are interviews, observation, focused group discussion, and document analysis. Particularly, interview and focused group discussion were used to collect the required data from road safety experts in selected organizations while observation and document analysis helps for closer understanding of the case.

2.2.3 Data Analysis procedures

As it has been repeatedly stated in qualitative research literature, it is not easy to clearly draw a line between data collection and analysis tasks. The analysis affects the data and the data affect the analysis in significant ways. Therefore, literature prefers to speak of "modes of analysis" rather than "data analysis" in qualitative research. These modes of analysis are approaches to gathering, analyzing and interpreting qualitative data. The common thread is that all qualitative modes of analysis are concerned primarily with textual analysis, whether verbal or written (Avison and Myers, 2002).

The analysis of this research is based on the principle of the hermeneutic circle, which suggests that a deeper understanding of a text or text-analogue (road safety information management) in relation to its context can only take place through a back and forth movement of renewed understandings (Klein and Myers,1999). In doing so, the interview sessions, repeated observation and focused group discussions helped a lot.

As stated by Darke et al. (1998), the three concurrent activities in data analysis, namely data reduction, data display and conclusion drawing/verification were employed. Data reduction refers to the process of selecting, simplifying, abstracting and transforming raw case data. Data display refers to the organized assembly of information using techniques like narratives, tables, matrices etc, to enable drawing of conclusions. Conclusion drawing/verification involves drawing meaning from data and building logical chain of evidence.

Cross-analysis of the road safety data collection and analysis practices in the organizations was mapped against the Zachman Framework to determine the dimensions and elements of the architecture. With its variety of modeling profiles, UML is chosen as a modeling tool to document architectural artifacts. UML and its business profile could be freely used for EA modeling based on the Zachman framework as discussed in (Fatolahi and Shams, 2006) and (Van der Walt and Du Toit, 2007). Literature providing conceptual frameworks and review of internal practices are also part of the effort and approaches in the process of the study.

2.3 Machine Learning/Data Mining Techniques

As investigating road safety situations through accident data analysis is one of the major concerns of this research, data mining techniques were used in the process of identifying a novel approach in discovering potential knowledge hidden in road traffic accident data accumulated. Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and summarize the data in novel ways that are both understandable and useful to the data owner (Hand et al., 2001). Researchers tried to prove its applicability in many domain areas and organizations. One of such areas could be road safety system, where very large historical accident related data is accumulated.

In connection to this, apart from the analysis previous machine learning researches in the selected areas, data mining experiments were also conducted. Though the main research covers three administrative regions in Ethiopia, this particular experimental study used data obtained from Road Traffic Office at Addis Ababa (Central Part of Ethiopia), Ethiopia. This is mainly because accident data is in long hand written format in Gambela region (South West part of Ethiopia) and it is still in a process of transferring accident data to a computer system in Amhara region (North West Part of Ethiopia). Even in the case of Addis Ababa, though all accident types were recorded on a computer based file at a regional level for about 5 years, fatal accidents were the only accident types recorded in a computer database during the data collection stage of the current research at a central level, while other type of accidents are recorded at woreda (lowest administrative division) level on a flat file format.

To confirm to the industry-standard process, the machine learning methodology used was guided by the CRISP-DM (Cross-Industry Standard Process for Data Mining) process framework. Accordingly, based on the situational analysis on the case study, business and data understanding were the first tasks. Then, follows exploration of data quality issues, pre-processing and feature/attribute selection tasks relevant to the data mining goal identified. Model building and evaluation along with a possible recommendation to integrate the resulted pattern or knowledge with the existing one was the last stage.

With respect to the specific techniques employed, predictive modelling methods were used. More specifically, five classification techniques namely; CART, TreeNet, RandomForest, PART rule induction and J48 Decision Tree were used. They are selected for their wide applicability and advantages that they have, which is discussed respectively. In addition, a parallel configuration of combining models with a majority vote approach is used as an ensemble technique. A brief description of the five predictive methods and the model combination techniques is presented in the next subsections.

2.3.1 CART Method

As explained by Gey and Nédélec (2005) Classification and Regression Trees (CART) is a robust decision-tree tool for data mining, pre-processing and predictive modelling tasks. CART can analyze complex data for patterns and relationships and uncovering hidden structures. Moreover, CART is a nonparametric technique that can select variables from a large data set and their interactions that are very important in determining the outcome variable to be analyzed. Some of the major advantages of CART as described in (Steinberg and Golovnya, 2006), include faster training times, its ability to use non pre-processed data (no need to transform or prepare the data), automatic handling of missing values, automatic handling of categorical (nominal) predictors, handling very large numbers of predictors, and ability to handle very large training data files.

CART analysis include important features like a set of rules for splitting each node in a tree; deciding when a tree is complete; and assigning each terminal node to a class outcome. CART always base on questions that have a 'yes' or 'no' answer to split a node into two child nodes; the yes answers to the left child node and the no answers to the right child node. Thus CART's method is to look at all possible splits for all variables included in the analysis which is followed by ranking the order of each splitting rule based on a quality-of-split criterion. Hence, the common criterion usually used is a measure of how well the splitting rule separates the classes contained in the parent node. Having the best split, CART repeats the search process for each child node, continuously and recursively until further splitting is impossible or stopped. After having the maximal tree grown and derived set of sub-trees, CART determines the best tree by testing for error rates or costs. With sufficient data, the simplest method is to divide the sample into learning and test sub-samples. The learning sample is used to grow an overly large tree. The test sample is used to estimate the rate at which cases are misclassified. The misclassification error rate is calculated for the largest tree and also for every sub-tree. The best sub-tree is the one with the lowest or near-lowest cost, which may be a relatively small tree (Sulaiman et al., 2011).

2.3.2 TreeNet Method

Developed by Jerome Friedman, TreeNet is a robust multi-tree technology for predictive modelling and data processing (Steinberg et al., 2002). In addition to its ability to handle both classification and regression problems, TreeNet offers exceptional accuracy, blazing speed, and a high degree of fault tolerance for dirty and incomplete data. (Friedman, 2002, Steinberg et al., 2002).

Applying TreeNet model indicate improved, or at least competitive, prediction accuracy than CART (Elish and Elish, 2009). TreeNet is an enhancement of the CART model using stochastic gradient boosting (Friedman and Meulman, 2003). Boosting refers to the endeavours to "boost" the accuracy of any given learning algorithm by fitting a series of models each having a low error rate and then combining into an ensemble that may perform better (Elish and Elish, 2009, Schapire, 1999). TreeNet can be seen as a collection of many smaller trees contributing to a final model. A final model prediction is constructed by summing up the contributions of each tree. As explained in (Steinberg et al., 2002), the key features of TreeNet models includes automatic variable subset selection; ability to handle data without pre-processing; resistance to outliers; automatic handling of missing values; robustness to dirty and partially inaccurate data; high speed; and resistance to over-training. It is also worth mentioning that, according to Salford Systems, TreeNet is resistant to overtraining and is over 100 times faster than a neural net.

2.3.3 RandomForest Method

As discussed in Krishnaveni and Hemalatha (2011), random forest is described as a collection of tree structured classifiers (h(x, _k), k = 1. . .) where the _k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x. The algorithm works as follows:

- *Choose T number of trees to grow*
- *Choose m number of variables used to split each node. m<<M, where M is the number of input variables, m is hold constant while growing the forest*
- *Grow T trees. When growing each tree do*
    - *Construct a bootstrap sample of size n sampled from Sn with the replacement and grow a tree from this bootstrap sample*
    - *When growing a tree at each node select m variables at random and use them to find the best split*
    - *Grow the tree to a maximal extent and there is no pruning*
- *To classify point X collect votes from every tree in the forest and then use majority voting to decide on the class label*

A Decision Tree Forest (DTF) is a collection of decision trees, where the combination of predictions contributes to the overall prediction for the forest. A decision tree forest grows a number of independent trees in parallel, and those trees do not interact until after all of them have been built. Decision tree forest models often have a degree of accuracy that cannot be obtained using a large, single-tree model (Sulaiman et al., 2011). Its ability to handle thousands of input variables without variable deletion along with quick learning process and its effective method for estimating missing data and maintains accuracy are major sited attributes of this algorithm.

2.3.4 Hybrid Architecture to Combine Models – Ensemble Approach

Hybrid architecture adopted for the three classifiers mentioned before is described in this section. Literature indicates that, combining classifiers provide better result. This is mainly because patterns misclassified by different classifiers are not necessarily same (Kittler et al., 1998). In connection to this, there are different strategies and configurations of combining classifiers. Cascading, Parallel and Hierarchical are the major configurations as stated by Ranawana and Palade (2006). Similarly, Wanas (2003) recognized two major architectures of ensemble; Cascading and Parallel. Cascading is when the output of one is used as an input for the next in order to reach a final refined classification. Parallel architecture adopted in this research, as shown in Figure 4, is a way of providing same input to a number of classifiers and combine their output using a given decision logic.

The decision logic could be linear, which includes averaging and weighted averaging of the results or non linear, which could be voting, probabilistic or rank based methods, as explained in (Ranawana and Palade, 2006). For this specific experiment a voted approach is selected where different classifiers provide their result for majority vote decision logic to determine the final class. A majority voting technique works very well when all the classifiers are somehow comparable or if there is no any very bad or very good classifier (Wanas, 2003). In case of different results from all classifiers the decision logic will consider the result of the classifier with better overall accuracy.

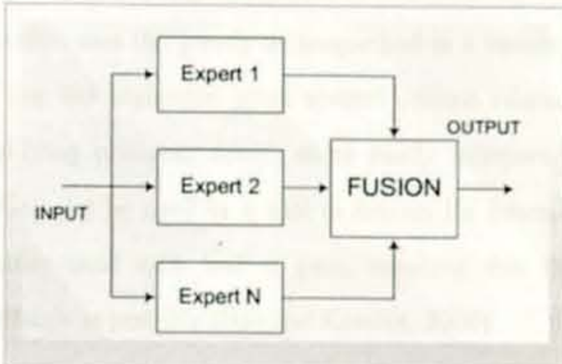


Figure 4: Parallel Ensemble topology

2.3.5 PART (Partial Decision Trees) Classifier

PART is a separate-and-conquer rule learner producing sets of rules called 'decision lists' which are ordered set of rules proposed by (Witten and Frank, 2005). PART is a rule based algorithm and produces a set of if-then rules that can be used to classify data. It is a modification of C4.5 and RIPPER algorithms and draws the divide-and-conquer strategy of RIPPER and combines it with the decision tree approach of C4.5. To generate a single rule, PART builds a partial decision tree for the current set of instances and chooses the leaf with the largest coverage as the new rule. The trees built for each rules are partial, based on the remaining set of examples and not complete as in case of C4.5 (Krishnaveni and Hemalatha, 2011).

Hence, PART is a class for generating decision list in WEKA. It builds a partial C4.5 decision tree in each iteration and makes the best leaf into a rule. Rules or decision lists, which are generated using PART algorithm, are more clear and understandable. Its simplicity and ability to provide sufficiently strong rules are identified as its advantages (Krishnaveni and Hemalatha, 2011) and therefore used for modelling this research.

2.3.6 J48 Decision Tree Classifier

J48 algorithm is the WEKA implementation of the C4.5 top-down decision tree learner. The algorithm uses the greedy technique and is a variant of ID3 (Han and Kamber, 2006). The J48 algorithm gives several options related to tree pruning. Pruning or simplifying produces fewer, more easily interpretable results. More importantly, pruning can be used as a tool to correct for potential over fitting. It recursively classifies until each leaf is pure, meaning that the data has been categorized as perfectly as possible (Han and Kamber, 2006).

The basic Algorithm is presented as follows in (Krishnaveni and Hemalatha, 2011)

- *Check for base cases*
- *For each attribute "a" find the normalized information gain from splitting on "a"*
- *Let a_best be the attribute with the highest normalized information gain*
- *Create a decision node that splits on a_best*
- *Recourse on the sub lists obtained by splitting on a_best, and add those nodes as children of node*

2.4 Evaluation Approaches and Techniques

Evaluation is a crucial component of any research. Both the process and end result of research undertaking needs to be validated. In the context of this research, the utility, quality, and efficiency of a design artifact needs to be rigorously demonstrated via well-executed evaluation methods (Hevner et al., 2004). It is also recommended that the selection of evaluation methods must be matched appropriately with the designed artifact and the selected evaluation metrics. The evaluation method selected from the knowledge base must be appropriate for the artifact in question (Carcary, 2011). Hence, evaluation of relevance to determine the effectiveness in contributing to solve the problem at hand is a preferred evaluation approach.

Literature shows that several different architecture evaluation techniques exist and can be modified to serve in a product line context. As tried to summarize by Abowd et al. (1997), these techniques can be categorized broadly as either *questioning techniques* (those using questionnaires, checklists, scenarios, and the like as the basis for architectural investigation) or *measuring techniques* (such as simulation or experimentation with a running system). This presentation agrees with the classification of evaluation methods in design science researches of Information systems by Cleven et al. (2009) as *action research, case study, experiments, prototyping* and *surveying*

As the study at hand is a design science research and the subject of this study is a model artifact, appropriate evaluation techniques from the paradigm were used. Models are evaluated with respect to their fidelity with real world phenomena, level of detail, robustness, and practical utility (Hevner et al., 2004, Hevner, 2007, Carcary, 2011).

Accordingly; surveying, descriptive evaluation, experiment and the case study itself (Cleven et al., 2009, Hevner et al., 2004) are examined and employed for the evaluation of the design artifacts of this research work. This approach is also in agreement with the generality, novelty and explanation capability measures discussed in (Carvalho, 2012) regarding validation criteria for design science research results. Surveying through questionnaire is an evaluation method to confirm the validity, completeness and utility of a model. According to Cleven et al. (2009), by conducting a survey, information is collected through interviewing representatives of a certain target group in the process of evaluating design artifacts.

A descriptive evaluation method involves informed argument using information from relevant researches to build a convincing argument for the artifact's utility. An experimental evaluation method involves the use of the models or artifact either in a controlled environment for qualities or in a simulation form to test the usability of it. The case study, which means the fact that the development of the architectural models is based on empirical data, itself, is used to validate the proposed approach through empirical data collected.

In line with the above discussion, questionnaire items reflecting relevant metrics are adapted and modified to collect data to confirm the validity of the architectural model, which is completed by road safety experts. It is basically used to conduct interviews with road safety experts who are working directly or indirectly on road safety data collection and analysis. In addition, descriptive evaluation technique which is an explanation based on literatures is also used as an evaluation method (Xie and Helfert, 2011, Hevner et al., 2004).

Thus, while the models of road safety information architecture are evaluated using surveying, descriptive evaluation and the case study itself, accident data analysis models are evaluated using experimental approach through a train-validate-test method in terms of accuracy performance, error rate and the Receiver Operating Characteristics (ROC) curve.

Accuracy is a proportion of correctly predicted instances against the incorrect once. The ROC curve, also known as the relative operating characteristic curve, is a comparison of two operating characteristics as the criterion changes. It can be represented by plotting the fraction of true positives (TPR = true positive rate) versus the fraction of false positives (FPR = false positive rate). An ROC analysis provides tools to select possibly optimal models. The ROC analysis is directly and naturally related to the cost/benefit analysis of diagnostic decision making (Beshah and Hill, 2010).

Another critical issue in validating a research work is a theory support both for the process and the output of a research. Accordingly, reference to literature has been used to validate statements. It is also worth mentioning that during the process of validation of statements, the research has strived to use multiple sources. Member checking through presentations, and repeated observation at the research site, and participatory modes of research were used to ensure reliability and internal validity of the research. Last but not least, peer reviewed conference and journal publications resulted from this specific research as indicated in own work (Beshah et al., 2011b, Beshah et al., 2012b, Beshah et al., 2012a) support the validity of the process and results of the research undertaking.

Chapter Three

LITERATURE REVIEW

3.1 Overview

The literature review was conducted and organized in three major areas; Road Safety and Traffic Accidents, Applications of Machine Learning/Data Mining, and Enterprise Information Architecture. The first sub-section covers the theoretical foundation of road safety and an understanding of road safety and road accidents. The second sub-section includes discussion of the applications of machine learning/data mining in general. The application and theoretical foundation of enterprise information architecture comprises the third sub-section. The final sub-section summarizes the literature review.

3.2 Road Safety and Traffic Accidents

Road safety continues to be one of the nation's most serious public health issues. It affects everyone, whether you drive, walk or cycle. Thus, it is apparent that it is a prime concern to individuals, groups and organizations, all of whom require data and evidence about accidents (Peden et al., 2004, Moon, 2003). In relation to this, it is stated that the major focus of road safety researches should be to study the road and traffic system in any of its aspects to find scientific evidences and suitable solutions for reducing the number of road accidents or their severity (Zheng, 2007, WHO, 2009).

Information relating to road accidents represents an essential means to detect and control road safety problems, to identify priority and to evaluate the effectiveness of the measures used to improve safety. In this context, the study, evaluation, improvement and optimization of the accident information management represent a priority objective in the national and international road safety programs and policies (Tormo et al., 2009).

Road traffic injuries are estimated to be eighth leading causes of death globally (WHO, 2013). In the year 2020, road traffic injuries are projected to become the 3rd largest cause of disabilities in the world. "*Road traffic injuries are increasing, notably in low- and middle-income countries, where rates are twice those in high-income countries*"(WHO, 2013). The problem is increasing in these countries at a fast rate, while it is declining in all industrialized nations of Western Europe, North America, Japan, Australia and New Zealand (Samson, 2006).

As indicated in the previous sections, road safety has been recognized as a major concern long ago. However, a good understanding of system complexity is lacking as the existing approach has been fragmented. In-depth accident investigation is hardly applied in a road traffic safety analysis (Stoop and Thissen, 1997). Estimate of the annual number of deaths vary, as a result of limitation of injury data collection and analysis (Zheng, 2007).

In the literature it is argued, that there is no as such a fundamental research effort on road safety except in few countries like Sweden and the Netherlands. Hence, it is logical to follow an argument by Stoop and Thissen (1997), which states that a shift in attention to a more integrated approach to the notion of 'safety' itself is needed. In addition, as the situations might not be the same, improving road safety requires appropriate measures for particular problems that exist in particular countries or regions (Elvik et al., 2006). As stated in (Kononov and Janson, 2002) one cannot improve safety without successfully relating accident frequency and severity to the causative variables. According to Firestorm (Fridstrom, 1999) factors affecting road accident counts can fall in either of the following categories;

- *Autonomous factors:* this includes factors determined outside the (national) social system; weather, the state of the technology, size and structure of the population
- *Socio-economic conditions;* more specifically issues related with unemployment, taxation, inflation, public education

- *The size and structure of the transport sector;* complexity of the transport infrastructure, overall travel demand, driving license penetration
- *System of data collection;* a case of accident underreporting
- *The sheer randomness of the phenomenon;* difficulty in predicting

In line with the above-identified factors, one of the focuses of this research is to address road safety problem through effective road traffic accident data reporting and analysis.

In dealing with accident analysis, there are a couple of approaches in use so far. One of the dominant approaches in accident analysis is the Haddon's Matrix. It is a conceptual model for the systematic exploration of countermeasures and provides an integrated approach to accident control (Williams, 1999). The Haddon Matrix is a conceptual model used to evaluate the nature of any type of accident particularly those involving repeating occurrences (fire, road traffic, violence…). It provides a compelling framework for understanding the origins of injury problems and for identifying countermeasures to address those problems (Runyan, 1998). It helps to shift what was a nearly exclusive focus on the pre-event/human cell to a broader approach involving event and post event phases. It has been used as a tool to assist in developing ideas for preventing injuries of many types. As a conceptual model it has helped guide research and the development of interventions (Runyan, 2003).

There are also other approaches to road safety analysis. The two widely cited approaches to road safety analysis and prevention are Road-User Approach (RUA), which is road user focused approach and Vision-Zero Approaches (VZA) which assumes road safety as a shared responsibility of different stakeholders. Recently, Systems theory approach is proposed as an enhancement to Vision-Zero Approaches (VZA), to road safety (Larsson et al., 2010). Summarized evaluation of the three approaches to road safety problem, is presented in Table 1. Major issues addressed include definition of the approach, focus and central issue, utility, limitation and implication for the current study.

Table 1 -Evaluation of the Three Approaches to Road Safety Problem

	Vision-Zero Approach (VZA)	Road-User Approach (RUA)	The Haddon Matrix
Definition	Originated in Sweden, it advocates that responsibility for road safety should be shared between road users, designers and administrators	This is an approach attributing cause of accidents to in appropriate use of the traffic system by road users	A conceptual model for systematic exploration of counter measures in three phases of an accident by considering four factors
Focus and Locus	Prevention and mitigation working on Human, Vehicle and environment (technical and social) as a locus of control	Prevention focusing on Road users as a locus of control	Prevention and mitigation working on human, Vehicle and environment (technical and social) as a locus of control
Utility	Views road transport system as a hierarchical system, which enables systemic approach	Emphasis on road users to enable better behavioral change regarding road traffic	Mainly for identification of countermeasures
Limitation and Implication for the current study	The control and learning process between the components is not well established.	As the focus is on road users, less emphasis on other factors. No guideline regarding how accident information can contribute to address road safety issues	More of high-level guideline requiring investigation in context. Though it cites "information" one aspect, no instruction as how the accident information will help addressing the problem
Seminal references	Larson (2009)	(WHO, 2004), Larson (2009)	Runyan (1998); Runyan (2003); Williams (1999)

Assessment of existing practices and literature in the area indicated that police reports attribute more than 90% of road traffic accidents to driver errors worldwide and these lead to incorrect conclusions like changing driver's behavior will enhance road safety. In Ethiopia, this is reported as 81% (Gidey, 2010). Such "simplistic representations" of traffic safety disregard the dynamic interactions among the road environment, the vehicle, and the road user (Zein and Navin, 2003). Systems approach in road safety acknowledges the more complex nature of road traffic accidents where multiple factors interact resulting an accident (Larsson et al., 2010). According to systems approach, accidents occur when the components interact and those interactions violate the constraints (Leveson, 2002).

One of the prominent works cited in the area in line with a system approach is a road safety strategy in (Baguley, 2001). According to the author, contextually appropriate measure should be selected, in the effort of improving road safety. In connection to this, the five ways of road safety improvements known as 5E's (education, enforcement, engineering, encouragement and evaluation) are well known approaches. As illustrated in Figure 5, road safety can be improved through two broad strategies; *accident prevention* and *accident reduction and mitigation*. The research illustrates the fact that the accident data collection and analysis should be at the heart of planning for improvement in all these sectors and should be used ultimately as the fundamental measure in evaluating how effective the various actions that are taken have been.

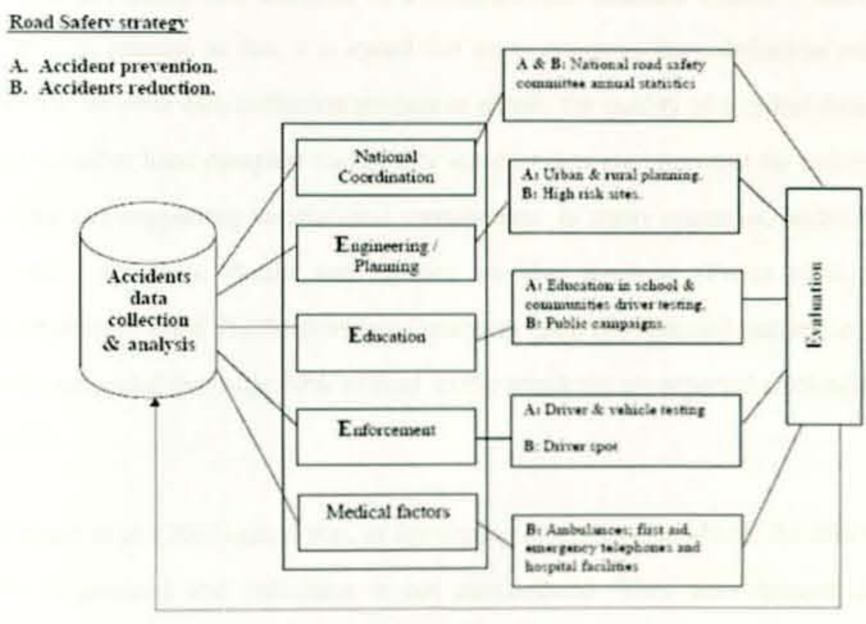


Figure 5: Road safety strategy (Baguley, 2001)

As discussed in literature, (Baguley, 2001) there are a number of areas where road safety data are often problematic, and these include but not limited to source of data (whether from police or health centers), type of data collected, inappropriate use of indicators, non-standardization of data, definition problems, underreporting and poor harmonization and linkage between source of data. As noted in (Peden et al., 2004) improving road safety also involves identifying the risk factors that contribute to crashes and injuries, along with identifying the interventions that reduce the risk associated with those factors. Regular and standard information at a national level would assist in directing the work on improving safety within the government's Road safety strategy (Mosedale et al., 2004).

Information about accidents is a prerequisite for improving road safety. It also helps assess the current situation and give a good indication of its severity (Abugessaisa, 2008a). Data relevant to road safety are collected every day in most countries, but for these data to be useful for informing road safety practice, they must be properly coded, processed and analyzed in a computerized database system (Peden et al., 2004). In relation to this, it is stated that many countries have definition problems, there is no clear data collection process or simply the quality of accident data is low. On the other hand complete and quality accident data are important for auditing road safety and supporting international comparisons. In many countries, underreporting of road accidents, deaths and injuries are also common (Peden et al., 2004). International Road Traffic Accident Databases (IRTAD) reported that earlier studies have estimated that only 50% of road traffic accidents are reported worldwide (Hai, 2009).

Afukaar et al. (2003) noted that, in developing countries like Ghana, the data sources are fragmented and collection is not standardized. They also stressed that the inadequacy in a data analysis affected the efficiency of establishing crash and injury rates. On the other hand Yannis et al. (2009) developed a common road traffic accident data framework (CADaS) for effective harmonization of accident data at both national and EU countries. They used a holistic approach based on thorough examination of both data and system availability, and usefulness to meet the needs of the different stakeholders for accident data analysis.

As stated by Sohn and Shin (2001) the effectiveness of data analysis highly depends on the quality and standard of the accident data collected. With respect to the validation of road accident data's accuracy and reliability, underreporting, missing and low quality of data are identified as major data limitations (Yannis et al., 2009). This implies that contextually framed empirical studies are required for thorough investigation and analysis of the problem.

3.3 Machine Learning/Data Mining approach

Data mining with the intent of prospective and proactive information delivery is the current trend (Thearling, 2003). In this Section attempt has been made to present a review of literatures on the basic concepts of machine learning/data mining technology, techniques available and their applications.

3.3.1 What is Data Mining / Machine Learning

Various definitions for data mining exist in the literature. In (Han and Kamber, 2006) data mining is defined as '*the task of discovering interesting patterns from large amounts of data where the data can be stored in databases, data warehouses, or other information repositories*'. It is also defined as the extraction of implicit, previously unknown, and potentially useful information from data (Witten and Frank, 2005). Though in some literature, the concepts data mining and machine learning often used interchangeably, according to Venkatadri and Reddy (2011) the building blocks of data mining is the evolution of a field with the confluences of various disciplines, like database management systems, statistics, artificial intelligence(AI), and machine learning(ML).

As indicated in (Berry and Linoff, 2000), the three core areas of competency that are needed to be successful in data mining are; data mining techniques, data and the modeling of data. Data mining techniques refer to the conceptual approach to extract information from the data. Data refers to the collection of daily transaction or any sort of information kept from day to day activities. In relation to this coming up with a precise formulation of the problem to be solved and usage of the right data are the two important keys to success in data mining. Modeling is a process of applying algorithms on the data with a certain specifications. Thus, building models is only one step in knowledge discovery. It's important to properly collect and prepare the data and to check the models against the real world (Two Crows Corporation, 2005).

Literature agrees to classify data mining technology into either predictive or descriptive modeling. The goal of a predictive modeling is to predict the value of one column based on the value of other columns. It refers to the process of building a model that will permit the value of one variable to be predicted from the known values of other variables (Hand et al., 2001). Classification and Regression are the two most common tasks in predictive modeling. If the label is discrete (containing a fixed set of values), the task is called classification. If the label is a continuous value, the task is called regression.

Descriptive modeling on the other hand deals with describing all of the data and discovering patterns and segments of the data (Hand et al., 2001). Thus, it is also called unsupervised technique. Being unsupervised task, it helps us to see patterns and segments that behave similarly. The two most common descriptive modeling tasks are association and clustering.

Different problem types in a data mining process in the effort of solving the business problems are often also called data mining tasks. The most widely used tasks are classification, association, clustering, dependency analysis, prediction, segmentation and description (Witten and Frank, 2005).

3.3.2 Applications of Data Mining

Many organizations are using data mining to help manage their specific tasks. Identifying customers who are most likely interested in a new credit product is one instance of data mining application. Detecting fraudulent use of services using data mining technologies is also another area that interests a credit card company (Thearling, 2003). The application of data mining in biomedical and DNA data analysis is also discussed in (Han and Kamber, 2006). In line with this, discovering large-scale sequencing pattern, in identifying and study human gene for the development of new pharmaceuticals and cancer in cancer therapies is a recent focus of data mining.

According to Hong (2001), data mining in general and predictive modeling in particular , which is perhaps the most-used subfield of data mining, has got great attention in insurance, fraud detection and text categorization. In addition, data mining is being used for many other purposes, such as analyzing court decisions, discovering pattern in health care, accident analysis, pulling stories about competitors from newswires, resolving bottlenecks in production process, and analyzing sequences in human genetic makeup. Thus it is easy to learn that data mining can be used in various areas.

In (Han and Kamber, 2006) it is suggest that, since data mining is a relatively young discipline with wide and diverse applications, there is still a nontrivial gap between general principles of data mining and domain specific, effective data mining tools for particular applications.

3.4 Enterprise Information Architecture

This subsection covers researches exhibiting the application of information architecture and enterprise architecture concepts.

3.4.1 Information Architecture (IA)

Needless to say that information is now recognized as a valuable resource in day to day management of any governmental, social, economical or private businesses. As Information Architecture is a theme in this research work, it requires due attention in both understanding the concept as well as pulling it to the domain area of the research. Accordingly, attempt has been made to review relevant research undertakings, viewpoints and whitepapers.

3.4.1.1 Definitions and Evolution of Information Architecture (IA)

There are a number of definitions, views and conceptualizations of information architecture in the literature. So far many researchers and practitioners provide definitions of IA from various perspectives (Morville and Rosenfeld, 2006, Van der Walt and Du Toit, 2007, Evernden and Evernden, 2003, Downey and Banerjee, 2011). However, no single term universally describes an encompassing framework for managing information as a resource. It is defined in (Resmini and Rosati, 2011), that Information architecture (IA) is a professional practice and field of studies focused on solving the basic problems of accessing and using, the vast amounts of information available today. In this context, the focus is clearly on managing information for better enterprise-wide consumption and use. IA includes descriptions of uses, users, physical properties, and constraints. As explained in (Robbins et al., 2011), it generally, represents a higher level of abstraction, emphasizing an awareness of systems in terms of how critical subcomponents interacts according to semantic aspects of processes, designs, and metrics.

In the same line, it is also stated that IA is a high level map of the information requirements of an organization showing how the major classes of information assets are related, to each other and to other components of a system (Pienimäki, 2005). Thus, Information Architecture is used to define the data and information that need to be made available to accomplish the mission and to provide to other agencies (Mahmood, 2006).

According to Xie and Helfert (2011), Information Architecture (IA) is used to organize information about a topic in order to manage it in a structured way. The authors also extend Richard Saul Wurman's conceptualization, that IA is used whenever a high-level overview of interrelated information components wanted to be defined, and when the relationships among them are complex and difficult to understand. It is also stated that IA combines the background theory, design principles, and diagrams representing the meaning of gaining insight from information (Evernden and Evernden, 2003). As can be seen in various literature, an Information Architecture has been described in many ways, each tackling the question from often slightly different perspectives.

White (2004) also recognizes the difficulty in describing what "information architecture" exactly is. Through his discussion of the historical development of the concept the author emphasized the different views in approaching the concept, like from information design, web design and content management point of views. In connection to this, attempts have also been made in literature to present the analysis of all the definitions of IA to get a complete picture of the concept. From the analysis, IA is the science of structuring, organizing and managing the information where the art of labeling and findability and usability is also important. IA is implemented in shared environment with the data collection, analysis, exchange and data industry standards being important to make sure the data flow is transferred with agreement between the users (Yunus and Rahman, 2008).

Basically, information architecture can be seen as the construction of a structure or the organization of information. It defines what data to store, how data is stored, managed, and used in a system. Particularly, information architecture describes how data is persistently stored and how components and processes reference and manipulate this data. Moreover, it describes how external or other systems will access the data, and includes descriptions or diagrams of interfaces to data managed by other systems. According the discussion in (Van der Walt and Du Toit, 2007), it incorporates a variety of techniques drawn from diversity of disciplines such as information science, management theory, artificial intelligence, knowledge management and object oriented modeling.

Regarding the evolution of IA, it is common to get discussions on the change on its focus. One such claim is made in (Evernden and Evernden, 2003) indicating remarkable change on IA, becoming a sophisticated, multidimensional tool for managing information as resource. In connection with this in (Evernden and Evernden, 2003) the three generations of Information architecture is summarized as shown in Table 2. The three parameters used in describing the three distinct generations of IA are the "focus", "drivers" and more importantly the "content" of IAs.

As it can be learned from the summary, the first and second generations emphasized technology solutions rather than at organization's use of Information. In the second generation, as mentioned by the author, the discussion became more about information content and content management. This is mainly because of the emergence of knowledge management during the 1990s as a key business function and the availability of a broad range of information via the Internet. Today, the third generation reflects the need for separate technology and information architectures (Evernden and Evernden, 2003).

In line with this, by reviewing related researches (Xie and Helfert, 2011) stated that the emphasis so far was on data structure aspect, while the behavioral perspective of how information is viewed and used/interacted among all the actors is lacking. In their work, they describe IA as a collection of objects, which are representing/defining the structure of information. Information structure, according to the authors, includes technical aspects as defining and standardizing the information elements, the interfaces and relationships between the various information elements, and behavioral aspects that the relationship of how the information is managed for exchange and share processes among multiple parties (Xie and Helfert, 2011).

Table 2: The Three Generations of Information Architecture (Evernden and Evernden, 2003)

Generation	Focus	Driven By	Content
1st Generation 1970s and 1980s	System as standalone applications within individual organizations	Increasing functionality and sophistication of standalone applications	Explanation of the need for an architectural approach; Analogies with building architecture; Simple2D diagrams or frameworks providing overviews of the architecture
2nd Generation 1990s	System as integrated set of components within individual organizations	Growth in system complexity and interdependence; Demand for software reuse	Extension and duplication of diagrams from the 1st generation architectures; Population of frameworks with industry reference models.
3rd Generation Late 1990s and 2000s	Information as a corporate resource with supporting IT tools and techniques.	Emergence of the Internet, e-commerce, and an increase in business-to-business applications; Adoption of knowledge management, systems thinking, and a more holistic view of information as resource.	Explicit definitions of Principles and background theory; Development of multidimensional architectures; customization of information frameworks to the need of individual organizations; generic information patterns and maps.

Regarding the importance of IA, Barker (2005) mentioned that most people only notice information architecture when it is poor and stops them from finding the information they require. The author further elaborated the discussion by indicating that information architecture is most commonly associated with websites and intranets, though it can be used in the context of any information structures or computer systems (Barker, 2005). Thus, it follows that unless organizations recognize the importance of information architecture, they will face a problem of information access.

It is also worth mentioning that, effective information architecture enables people to navigate through a system knowing that they are getting closer to the information required (Jafari et al., 2009). Besides, information architecture addresses key considerations for both the current and future states of information processing. It makes it possible to articulate the needs of stakeholders, as well as the semantics of behavior of the system and its parts in common terms. It is also mentioned in literature that, sound information architecture allows for full support of distributed processing, interoperability across heterogeneous systems and departments, and internetworking between systems (Barker, 2005).

Extending the idea it is stated, that effective information architecture comes from understanding business objectives and constraints, the content, and the requirements of the people that will use the system or the site (Jafari et al., 2009). Information architecture can also be used to guide application development and to facilitate the integration and sharing of information (Pienimäki, 2005). Some major advantages of enterprise information architectures are summarized in (Pereira and Sousa, 2004). It enables an integrated vision and a global perspective of informational resources (Niederman et al., 1991) in addition to serving as the bridge between the business and technical domains (Young, 2001). It also contributes to having information systems that reflect common goals and performance measures for all managers, to encourage cooperation rather than conflict, and competition within organizations (Stata, 1989).

3.4.1.2 Application and Importance of IA

Though, the concept of IA is usually associated with web design, its application is diverse. Advancements in organizational interdependencies, e-commerce, knowledge discovery, information management and systems thinking have helped drive the view of information as a critical organizational asset, as discussed by Evernden and Evernden (2003). In line with this, there are different attempts exhibiting the application of IA in different domains and context.

Van der Walt and Du Toit (2007) approached the concept of information architecture as a way of visualizing and describing the various information assets and interactions of these assets within an enterprise. The establishment of information architecture is illustrated through a case study with in a large conglomeration of companies requiring scalable information architecture in order to address its information requirements. Another attempt is a research conducted to design information architecture for Malaysian Natural Product Repository by (Yunus et al., 2008). Extensive and critical literature review and a five phase IA development approach were used in the process of designing the said architecture.

With respect to the importance of IA, in (Yunus and Rahman, 2008) it is stated that, the need of IA caused a debate in the information technology world. It is argued, that it is very difficult to calculate the return on investment of IA. In the short term, it is not easy to see the need of IA but they argue that in the long term where maintenance work is needed, we could see the importance of IA. The authors go further in categorizing the importance of IA into 8 major categories that are from the perspective of user, development, cost, organization, guideline, e-commerce, value and time. They reported that all the importances of IA are extracted from literature review of different resources and combined in one table to make it easily understood. The categories are chosen by extracting the main issues that literatures try to point out from the description in each resource.

Similarly, some of the major importances of information architecture are highlighted in (Van der Walt and Du Toit, 2007) as follows:

- *"Information Architecture helps everyone* - information as a resource is the responsibility of everyone within the enterprise and not the preserve only of the technology department. IA provides practical tools, improves efficiency, effectiveness and productivity and supports the organizational strategy, innovation, creativity and flexibility"
- *"Architecture is a necessary foundation* – by making the architectural foundation more explicit, the importance of information as a resource is further reinforced. It is especially useful in the understanding of more complex information structures found in managing large enterprises."
- *"It applies to all types of information*- IA does not apply only to the design and navigation of web sites, nor is it only for the development of information technology or software. It is a universal discipline that applies to use of information in general."

Emphasizing on the importance of IA, a study on health sector, by Brra et al., (2007), recommends focusing on data/information standard is priority than technical standards which will be achieved through Information Architecture.

As to the process of developing an IA, in line with the design science research paradigm, there are different types of research methodologies for the development of IA. Two of the most widely accepted approaches are described here in brief. Evernden and Evernden (2003) provide a guideline consisting of four steps namely; deciding information management requirements, creating a management toolkit, defining an information map and using the information resource. The output of the first stage includes principles, information design guidelines, standards and naming conventions. It also decides which dimensions are in fact parts of the architecture. Requirements determine which architectural dimensions are relevant and in turn govern which tools are needed. A toolkit includes checklists, charts and diagrams derived from using the dimensions. A common tool is a grid structure comparing one architectural dimension with another. But, more sophisticated architectures, based on third-generation design principles, are multidimensional, using a set of interrelated charts and diagrams showing the architecture from different viewpoints.

With somehow similar approach, Morville and Rosenfeld's (2006) methodology tries to address the three major aspects of an Information Architecture (IA); Content, Context and Users. There are five phases in this methodology namely; Research, Strategy, Design, Implementation and Administration.

3.4.2 Enterprise Architecture Frameworks

In the process of developing an IA within a given context, there are a number of reference architectural frameworks including; The Open Group Architectural Framework (TOGAF), Department of Defense Architecture Framework (DoDAF), Federal Enterprise Architecture Framework (FEAF), Treasury Enterprise Architecture Framework (TEAF), Zachman Framwork and The Information Framework.

TEAF is developed in response to the need for establishing common enterprise architecture, consistent practices and common terminology within the various bureaus and offices at the Department of Treasury. *The TEAF describes four 'views' (comparable to the columns in the Zachman framework) and four 'perspectives' (comparable to the rows in the Zachman framework), creating a matrix (comparable to the one suggested by Zachman)* (Goethals-SAP-leerstoel, 2004). TOGAF is a high level and holistic approach to the design of enterprise architecture. It covers four architectural domains; Business, Application, Data, and Technology, which can be seen as views. It also provides architectural development methods with respect to architectural domains mentioned (Urbaczewski and Mardalj, 2006).

FEAF was developed to promote interoperability, the shared development of common federal processes, and the sharing of information. Its focus or dimensions are entities (what), activities (how) and locations (where). With respect to the views it provides five level perspectives in order of details (Goethals-SAP-leerstoel, 2004). DoDAF builds on three sets of views; operational, system and technical standards *Operational view describes/interrelates the operational elements, tasks and activities to accomplish mission operations. Systems view describes systems and interconnections to support the Operational View. Technical Standards describes rules governing the arrangement ,interaction and interdependence of system components to augment the Systems View* (Mahmood, 2006).

Another foundational architectural framework worth mentioning is the Information Frame Work (IFW) created by Roger Evernden during the 1980s. IFW remains the foundation for IBM's strategic solutions in the important financial services sector. IFW is a tool for analyzing and structuring information. It is organized under three views namely; organizational, business and technical. It is used to help structure information change and business flexibility or agility. The deliverables from IFW includes a set of models that provide checklists to business requirements, templates that embody industry standard design principles, detailed designs for specific solutions and application building blocks and code. These all can be combined to provide customized solutions to information-related problems (Evernden, 1996).

Urbaczewski and Mardalj (2006) conducted a study that provides a comparison of enterprise architecture frameworks. The purpose of the study was to present a direct comparison of the frameworks, based on their views and aspects. According to the authors they studied several existing enterprise architecture frameworks which helped them to establish a common ground for the framework comparison. Comparison was made based on the perspectives of their stakeholders and abstractions. Finally the research concluded, that Zachman framework appears to be the most comprehensive using a number of viewpoints (dimensions and perspectives) relate to different aspects. "*Zachman framework is a generic reference model, which serves as the basis for numerous other models eg. TOGAF, FEAF and TEAF*"(Mahmood, 2006).

According to (Goethals-SAP-leerstoel, 2004), many of the classic enterprise architecture frameworks focus on the software architecture, rather than on the total enterprise architecture while Zachman Framework is most comprehensive. The author claims that, the top two rows of the Zachman Framework are neglected too often though, they are very much important especially if integration is an objective.

The Zachman Framework is certainly the most widely known framework in the Enterprise Architecture context. The reason for its extensive use is due to the fact that it is a very flexible framework. The Zachman Framework does not impose a method and it does not restrict any user to a set of pre-defined artifacts (Pereira and Sousa, 2004, Goethals-SAP-leerstoel, 2004).

The Zachman framework provides a logical structure for classifying and organizing the descriptive representation of a given domain. The framework is structured around the view of different users involved in the planning, designing and maintaining an enterprise's information system. It provides a two dimensional matrix representation generic framework (see Table 3) where user's views are represented through Rows: Scope, Owner, Designer, Builder, Out-of-context and Product, while the data (what), function (How), network (Where), people (Who), time (When) and motivation (Why) dimensions are represented through Columns (Zachman, 1999).

Table 3-Zachman Framework

	What (Data)	How (Function)	Who (People)	Where (Network)	When (Time)	Why (Motivation)
Scope (Contextual)						
Business Enterprise (Conceptual)						
System (Logical) Information Systems						
Technology (Physical)						
Component (Detailed Specifications)						
Operations (Functioning Enterprise)						

Though, the Zachman framework is a 6 x 6 matrix representation, in (Bahill et al., 2006) it is stated that "what", "who", "why" and "how" dimensions are the most widely used while the designer and the builder user views get more emphasis by researchers and practitioners. The framework comes with the following set of rules to keep integrity of architectural descriptions; *Rule 1 - The columns have no order, Rule 2 - Each column has a simple generic model, Rule 3 - The basic model of each column must be unique, Rule 4 - Each row describes a distinct, unique perspective, Rule 5 - Each cell is unique, Rule 6 -The composite or integration of all cell models in one row constitutes a complete model from the perspective of that row, Rule 7 -The logic is recursive.*

The Zachman "Framework" is also considered as a taxonomy for organizing architectural artifacts like specifications, and models that take into account both whom the artifact targets such as business owner and builder and what particular issue like content and functionality is being addressed (Zachman, 2003). This entails that, it can be used as a guide in designing a mechanism to effectively and efficiently collect, analyze and disseminate road accident data which is being criticized for its incompleteness and unreliability.

Zachman Framework has been used as a reference model in varied domains for various purposes. Razak and Dahalin (2008) studied Enterprise Architecture (EA) practices in Malaysia through a qualitative approach. Purposive sampling was employed to select 10 enterprises from public and private sectors. Interview and document analysis were major data collection techniques while pattern-matching technique was used to analyze the qualitative data collected. The Zachman Framework was used to evaluate the practice of EA in these enterprises. According to the authors, the study reveals that some aspects of the framework were not addressed at all; whilst other aspects, which were addressed, vary in terms of perspectives. Finally, the authors reported that idea of EA is relatively new to Malaysian enterprises as none of the ten participating enterprises made reference to Zachman Framework or any other EA framework. They also reported on the extent of EA practice based on the mapping of the Zachman Framework, and it is stated that most of the enterprise's work focused on all dimensions except for TIME across all perspective.

An integrated process for developing data architecture views in Zachman framework was investigated and proposed in (Rezaei and Shams, 2008). The authors conducted a case study in one of the Ports and Shipping Organizations of Iran namely shipping and marine affairs authority to build on the credibility of their approach. Some questionnaire items were also used to evaluate the accuracy and correctness of the proposed method for creating integrated data architecture by architects and designers. Finally, the authors concluded that the presented approach using Zachman framework was well acceptable. In another work, (Pereira and Sousa, 2004) proposed a method, focusing on Business and IS perspectives, for achieving an Enterprise Architecture Framework, based on the Zachman Framework. Through their study the authors identified artifacts for each cell, and a method, which defines the sequence of filling up each cell in a top-down and incremental approach.

Fazil et al. (2010) studied application of Zachman framework to content organization in a library environment. The focus of the research was on creating content of semantic theses digital library by applying the Zachman Framework data dimension. In line with this, the goal of the study was to design a formal framework in managing theses and dissertations collections and specifically focused on the content organization based on the Zachman Framework for the Enterprise Architecture. The study involves case study approach with document analysis and questionnaire as a means of data collection.

A Study of Implementing Zachman Framework for Modeling Information Systems for Manufacturing Enterprises Aggregate Planning (Radwan and Aarabi, 2011) is another work to structure aggregate organizational planning using Zachman Framework. According to the authors, a need for adequate tools and methodologies to model the system structure and to define integrated requirements in manufacturing enterprises is a major motivation for the study. And it is stated that such modeling will address a major problem of lack of strategic perspective of information systems which will in turn help to improve response time on orders of customers which is a critical issue to achieve a competitive advantage.

In (Ertaul and Sudarsanam, 2005) an overview of how Zachman's Framework help define, design, and create tools for effectively securing an enterprise is provided and discussed. According to the authors, the research work exhibited that Zachman Framework best fits to plan security architecture for an enterprise as any evolving changes in technology can be implemented onto the Zachman Framework without affecting the direction of the enterprise. They also stated that, the security planning using the Zachman Framework applied to enterprises is helpful for sorting out complex technology and methodology issues that are significant both to general and technology management.

Similarly another study (Mohajerani and Moeini, 2004), "Using Enterprise Architecture Framework to Design Network Security Architecture" presents an approach to use enterprise architecture models as a framework to design network security architecture. The authors discussed network security architecture of academic centers as a case study to show how a conceptual model can be applied to a real organization. The study used the top three levels of Zachman hierarchy along with the data, function and network dimensions to develop descriptive security architecture.

Mahmood (2006) in a study "Architectural Representations for Describing Enterprise Information and Data" discusses a range of architectural representations from various perspectives, and presents a basic model for the development of an enterprise architecture emphasizing on information and data architectures. The author proposed to view each architectural representations from the *what, how, who* and *why* perspectives since EA is basically the *what, how, who* and *why* of the business at every level of the organization. This research work is based on one of the foundational frameworks called Zachman Framework.

Thompson (2006) studied the scalability of Zachman Framework to determine how it can be used to provide a development structure for a non-Enterprise Application. Through the study, the authors confirmed that the Zachman Framework is scalable however as a project becomes more complex, the first three rows of the Zachman Framework become more important. According to the author, this Framework provides a one stop solution and would be a superior methodology for the architect or engineer of systems both large and small.

Jafari et al. (2009) conducted empirical study based on the Zachman framework and developed conceptual knowledge architecture model that can be applied to Iranian organizations. The objective of the paper was to explore the role of knowledge architecture in an enterprise and to provide a model to architect enterprise knowledge based on the Zachman Framework. In the process of the study, the authors discussed several perspectives from the knowledge management (KM) point of view and information technology. A questionnaire was used to poll opinion of knowledge architecture experts as a means of validating the architectural model. According to the authors, the paper is of high value to researchers in the knowledge management field and to practitioners involved with KM adoption in the organizations. Moreover, it is stated that the result of the research gives valuable information and guidelines that hopefully will help the leaders and the senior knowledge management managers to accomplish KM through their organization successfully.

The discussion so far in this sub section provided sufficient theoretical background regarding information architecture and its application along with a guiding framework in the development and application of an enterprise information architecture. The discussion exhibited a wide application of information architecture as an approach, which can also be helpful in a road safety domain. In addition, enterprise view, specifically using Zachman Framework, in using information architecture as an approach is shown appropriate. By extending knowledge in the literature this research is based on the discussion made so far and an empirical data collected from road safety organizations to define road safety information architecture.

3.5 Summary of the Chapter

This Chapter presented research works and viewpoints revolving around the main theme of the research. Accordingly, literature on the very concept of road safety and road accident, machine-learning approaches and the applicability of enterprise information architecture were discussed.

It is learned from the discussion that road safety is an escalating problem for the society that needs multi-sectorial and consistent research efforts. In addition, it is mentioned that there are different efforts from different perspectives so far. Data mining/machine learning approach was also discussed to get an understanding of its application, which exhibits its versatility.

It is also exhibited that efforts to tackle road safety problems hinges on proper collection and analysis of road accident data. Review of the literature presented above provides sufficient theoretical base and evidence for a wide application of Zachman Framework, the selected enterprise information architecture framework. Its robust nature and flexibility lends itself to an application or system of any size in various areas (Thompson, 2006). It is evident that Zachman Framework can be used to define architecture for any object or system of interest. Accordingly, this research strives to employ Zachman Framework to study and describe accident data and its management.

Chapter Four

RELATED WORK

4.1 Overview

This Chapter is dedicated to the discussion of related attempts in addressing road safety problems in general and accident data collection, analysis and dissemination in particular. The process of reviewing related works revealed some interesting facts which provided useful insight in the consideration of the proposed architecture for road safety information management. Key and important findings are presented here in summary.

4.2 Evaluation of Literature to Address Road Safety Problem in General

There exist various attempts to address the road safety problem in different part of the world. Vasconcellos (2000) conducted a research on strategies to improve traffic safety in Latin America. The author started by recognizing the difference in the type and magnitude of the problem in developed and developing countries scenario. According to the author, unlike developed countries - where air pollution has long surpassed traffic accidents as major policy concerns - in developing countries, traffic accident is the worst transport externality and deserves proper priority treatment. It is also stated in the research that prevailing traffic planning approaches to this problem have followed traditional methods utilized in developed countries. However, the persistence of poor traffic safety conditions attests to inadequacies of available techniques and dominant approaches.

Through the study the author constructs a new approach and proposes actions to change current conditions towards improving road safety. Instead of considering isolated factors namely; road users, road and vehicle it is proposed that traffic accidents in developing countries can be better understood if the physical, political, institutional, and technical and enforcement environments are analyzed with respect to their influence on traffic safety conditions. Thus, highlighting on the deficiency of the existing system through critical analysis, the author proposed ten actions, one of which is about the gathering of reliable data and information on accident profile and costs.

In another study, (Larson et al., 2012) in the paper "The Importance of Data for Global Road Safety", presented a discussion on Road Safety in 10 Countries project (RS-10), which is focusing on Brazil, Cambodia, China, Egypt, India, Kenya, Mexico, Russia, Turkey and Viet Nam. The authors tried to compile the first two years of data collection and analysis during RS-10, and they claim that it represents important strides in road safety research.

Similarly (Elvik, 2010) studied why some road safety problems are more difficult to solve than others and presents a discussion on factors that can make some road safety problems difficult to solve. These problems, as stated in the work include the high risk of accidents and injuries involving young drivers; the high risk of injury run by unprotected road users and speeding. The author provides a framework for categorizing road safety problems according to their basic characteristics and serves as the basis for the discussion. In line with this, it is argued that if the main source of a problem is incompatibility of different groups of road users or if solutions involve overcoming social dilemmas, then problems are likely to be difficult to solve. The research also finds out that, problems to which biological factors contribute are also likely to be difficult to solve. The author also stated that, it is unfortunate that these problems all make major contributions to road accidents, suggesting that we may be approaching a point where making further progress in improving road safety will become increasingly difficult.

Another work approaching road safety problem through improvements in information sharing is a dissertation by Abugessaisa (2008a). The thesis presents a three-tiered conceptual model to support the sharing of road safety–related information and a set of applications and analysis tools. According to the author, the overall aim of the research was to build and maintain an information-sharing platform, and to construct mechanisms that can support road safety professionals and researchers which will in turn help the efforts in preventing road accidents. The author used several approaches, including requirement elicitation techniques, iterative prototyping, systems thinking, visual data mining and Technology Acceptance Model (TAM), in both the process and evaluation of the research idea.

In the effort of improving road safety, (Eksler, 2007) conducted a research that aims at providing better taxonomy and description of the different underlying factors hidden behind road safety outcomes at country or regional level. According to the author, the evidence is given on the existence of various structural factors based on two European countries, for which a Full Bayesian (FB) ecological regression model is applied to study different road safety outcomes at regional level. Through this approach, the author combines spatial and time variation within one modeling approach and investigates existing spatial relationships. The model in which a series of structural explanatory factors was tested within road fatality risk model setting, using 2000-2006 data of 3 regions in Belgium, is presented. It is also stated that, improvement on structural factors which could be conveniently influenced by policy related measures would result in more effective road safety management.

A study on 'The Impact of Localized Road Accident Information on Road Safety Awareness' by Yunan Zheng, a thesis submitted for the degree of Doctor of Philosophy to the University of Glasgow, is another related work examined in the area. Zheng (2007) claims that, existing road safety campaigns do little to address this problem; they focus on national and regional statistics that often seem remote from the local experiences of road users. The paper argues that localized road accident information would have better impact on people's safety awareness.

The research used a three stage methodology, where the first stage is the use of a risk perception assessment technique to find out how existing on-line road accidents information affects the general public's safety attitudes. It is followed by the design of a map based accident information system and psychometric techniques to assess the impact of our system on people's safety attitude.

The thesis describes the design and development of a software tool to provide the general public with access to information on the location and circumstances of road accidents in a Scottish city. The results of an evaluation to determine whether the information provided by this software has any impact on individual risk perception was also presented in addition to a route planning experiment carried out. According to the author, the results from the experiment give more positive feedback that road users would consider accident information if such information is available for them.

Similarly, Menon et al. (2010), studied a multi-sectoral approach to capture information on road traffic injuries. The study was undertaken in the hospitals of Bangalore and Pune, to examine the feasibility of gathering information on injuries using multiple sources. The research used a formally arranged stakeholders meeting and training programs for the hospital staff, police personnel, and traffic and transport staff, to identify their roles and responsibilities. In addition, pre-tested questionnaire was also used to collect prospective data on morbidity and mortality due to injuries. It is stated that the information gathered was cross-checked with the hospital and police records to keep the validity of the data.

According to the results of the research, the stakeholders meeting and training programs were able to motivate the departments to provide the correct data. The authors concluded that it is possible to improve the data on injuries by adequate training and a data linking mechanism between the Police, Hospital, and Transport Departments. Finally, the research emphasized the importance of a good data capture mechanism in the effort of addressing the problem of road traffic injuries.

Baguley (2001) studied the importance of road accident data system and its utilization. Through the research work, the author presents road accident fatality statistics of developing countries. The research discusses worrying trends, under-reporting, socioeconomic aspects of road accidents, and also common practices in safety improvement. The main focus of the research, however, is the importance of establishing a reliable road accident database and analysis system; road accidents being the fundamental measure of safety. The author claims that, access to the database is an essential part of identifying, and hence targeting, specific safety problems and in evaluating the effectiveness of any measures introduced. The research also identified and discussed most important data items for accident recording and various examples of analysis systems are presented.

To summarize, it is easy to learn from the discussion above that road safety is recognized as a global issue that deserves comprehensive, multifaceted and detail research. It is also worth mentioning that road safety problems can be tackled from different perspectives. Rigorous researches investigating improvements on the road design and environment, accident information management, behavioral change on road users, revision of road safety policies and measures are required. In line with this, this research focuses on accident information management. Summary of some selected researches addressing road safety problem is presented in Table 4. The summary includes methods used, approaches the researches will fall in (though it not explicitly stated so), key findings, context and perspective of the research.

Table 4-Summary of some selected road safety researches

Author	Approach researches will fall in	Method	Key Finding	Context	Perspe-ctive
Vasconcellos (2000)	Vision Zero Approach	Literature review and critical analysis	Accidents can be better understood if the physical, political, institutional, technical and enforcement are analyzed with respect to their influence on traffic safety	Developing	Policy
Elvik (2010)	Road User Approach	Survey and literature review	Categorizing road safety problems according to their basic characteristics	Developed	Policy and Behavioral
Abugessaisa (2008)	Systems and Vision-Zero Approaches	Design Research and Survey	Three tired conceptual model supporting Information sharing platform and analysis tools	Developed	Information
Eksler (2007)	Systems Approach	Design research	Improvement on structural factors influenced by policy can enable effective road safety management	Developed	Engineering and Policy
Zheng (2007)	Road User Approach	Survey and Design research	Localized accident information can have better impact on peoples safety awareness	Developing	Behavioral and Information
Menon et al. (2010)	Vision Zero Approach	Survey and participatory research	Data quality can be improved by adequate training and data linking mechanism between stakeholders	Developing	Policy and Information
Baguley (2001)	System Approach	Literature Review	Highlights the importance of establishing accident data base and in-depth analysis of accidents data	Developing	Information

4.3 Data Mining Approach in Road Safety Domain

As discussed in Chapter three data mining has versatile applicability. A number of data mining-related studies have been undertaken to analyze road accidents data locally and globally, with results frequently varying depending on the socio-economic conditions and infrastructure of a given location (Beshah and Hill, 2010).

As indicated in (Sohn and Shin, 2001), the trend in accident data analysis is changing from cross tabulation to multivariate analysis and data mining especially variable selection and classification to investigate the relationship among a large number of variables in traffic accident records. Various researchers from different parts of the world have been trying to study different aspects of road traffic accidents or car accidents. Chong, et al (2004) categorizes these researches into three basic areas namely;

- *Prediction of injury severity,*
- *Establishing the most important factors influencing the severity of the injury*
- *Miscellaneous works related to traffic accidents.*

It is also mentioned by Zheng (2007) that, solutions to road safety problems are changing all the time and road safety research involves a very large scientific area. According to the author, it can be divided into the following four fields:

- *The study of occurrence of accidents;*
- *The detailed study of any aspect of the accident process which may be a factor in accident production;*
- *The consequences of accidents, i.e. injury and damage;*
- *The economic aspects of accidents and safety measures*

Attempt has been made to assess the existing accident analysis practice at the three selected regional administrations in Ethiopia. Generally, the result revealed, that currently no such analysis is actually being done in the Gambela region (South West part of Ethiopia) while limited descriptive analysis is practiced in Amhara (North

West part of Ethiopia) and Addis Ababa (Central Part of Ethiopia) regions. In addition, the data quality issues are not yet addressed.

However, though they lack systematic approach, there were some fragmented efforts to show the application of data mining techniques on road safety analysis domain. (Tesema et al., 2005) used adaptive regression trees in their research on rule mining and classification of road traffic accidents, which provides a foundational work on severity analysis in the Ethiopian context. Important variables determining severity in identified in the study were *accident type* and *accident cause*. The results, according to the authors, showed that the developed models could classify accidents' severity within reasonable accuracy. The best accuracy exhibited was 87%.

In (Regassa, 2009) classification algorithms for the study of accident severity and driver characteristics were explored. The study focused on predicting the degree of drivers' responsibility for car accidents. The research used WEKA, data mining tool, to build the decision tree (using the ID3 and J48 algorithms) and MLP (the back propagation algorithm) predictive models and identify important relationships between variables that influence drivers' degree of responsibility such as; age, license grade, level of education, driving experience, and other environmental factors. Accuracies of the models were 88.24% and 91.84%, for decision tree and MLP respectively. In addition, the research reveals that, the decision tree model is found to be more appropriate for the problem type under consideration.

With regard to accident occurrence, Kifle (2009) explored the application of data mining to identify dangerous locations in Addis Ababa. Decision Tree (J48 classification algorithm) was used in conducting a couple of experiments with 10 variables. The research identified school areas, while the weather is rainy as the 1st ranked death locations followed by institution areas during night time and cloudy weather conditions. After providing detail results in identifying dangerous locations, the research recommends exploration of the case for further researches by including property damage as it was only focused on fatal, serious injury and slight injury cases only.

In another study, Mossie (2009) demonstrates data mining models for accident severity analysis in support of reducing road traffic accidents by identifying and predicting the major vehicles and driver's determinant risk factors (attributes) that cause road traffic accidents. The research uses WEKA, version 3-5-8 tool to build decision tree (using j48 algorithm) and rule induction (using PART algorithm) techniques using 11 input attributes. The result of the research proves that the performance of J48 algorithm is slightly better than PART algorithm and it identified that *License Grade, Vehicle Service year, Type of vehicle and Driver Experience* as most important variables to predict accident severity pattern.

Bayesian Network power predictor and constructor was employed by Tabor (2009) for prediction and model construction purpose respectively in the process of two experiments which were made before and after the elicitation of the domain experiments. According to the first experiment, type of accident is directly influenced by four factors namely; license grade, time of accident and cause of accident and driver experience with the accuracy of 87.96%. In the second experiment (after elicitation of domain experts) the best accuracy was 80.28% and type of accident is highly influenced by weather condition, road joint and type of vehicles. Beshah and Hill (2010) utilized Decision Tree (J48), Naive Bayes and K-Nearest Neighbours algorithms to explain the role of road related factors for severity. The result shows that, all the three classifiers perform well similarly with respect to correctly classified cases. A PART algorithm was also used to generate user understandable rule, with the accuracy of 79.94%. The rules presented in PART format indicated that accident severity (fatal, sever injury, slight injury and property damage) varied with different combinations of road-related factors. This is exhibited with example cases were more scenarios for severe injury on straight plain roads than other orientations of roads in the same sub-city. The research also identified that K-nearest neighbor exhibited better accuracy (80.8182%) and ROC (above 0.8) in predicting new instances. The authors propose further investigation by combining different factors like road and driver related variables.

Based on the recommendation in (Beshah and Hill, 2010), Anteneh (2011) studied the role of drivers and road related factors in determining accident occurrence and severity. The aim of the research was to improve public health through identification of determinant factors. Accordingly, through a number of experiments using predictive modelling techniques, *Licence Grade, Sub City, Road Junction, Type of Road, and Light Condition* are identified as determinant factors. The research also shows that J48 slightly outperforms ID3 and PART algorithms.

As these studies are conducted in a similar context with the current research, a detail evaluation is made by considering their objective, methods and techniques, attributes used and key findings. Accordingly, Table 5 presents this assessment.

Table 5- Evaluation of Data Mining Experiments Conducted in a Local Context

Author	Solution Objective	DM Methods & analysis Type	Variables Studied	Key Findings
Tesema et al. (2005)	Accident Severity Analysis	Predictive modelling using adaptive regression trees	16 available variables related to road, driver, vehicle and environment	Accident Cause and accident type as major determinants.
Regassa (2009)	Identifying driver responsibility	Classification and neural network	Driver related factors	Driver age, License grade, level of education, and driver experience are found to be more important
Kifle (2009)	Identifying dangerous locations	Decision tree based classification	Ten Road and environment related attributes	School area during rain is found most dangers location

Mossie (2009)	Identifying determinant factors related to Drivers and vehicles	Decision tree and rule induction techniques	Eleven driver and vehicle related factors	Licence grade, vehicle service year, type of vehicle, and driver experience are found to be most determinant
Tabor (2009)	Determining factors for accident	Bayesian network power predictor	Variables related to road, driver, vehicle and environment	Weather condition, road joint type, type of vehicles, are found to be major factors
Beshah and Hill (2010)	Explain role of road related factors	Classification using DT, Naviebayes, and K-means algorithm	Road related factors (attributes)	Severe accidents on straight plain roads than other type of roads
Anteneh (2011)	Identifying determinant Driver and road related factors	Predictive classification techniques	Driver and road related factors	Licence grade, road junction, sub city, type of road, and light condition are found to be determinant factors

As can be seen from the Table 3, data mining experiments made so far didn't consider road users factors except drivers. They tried to address accident severity from one or two aspects focusing mainly on drivers and road factors, overlooking human factors like pedestrian and victim related attributes. Another important issue worth mentioning is limited number of attributes considered in the experiments. The less the number of attributes used the less the experiment explains the situation well.

While the above were specifically targeted attempts in applying data mining in a road safety domain in a local context, it is also worth mentioning other efforts in employing different methods and tools for better understanding of the domain and improved accuracy worldwide. In (Janecka and Hulova, 2011) an experiment is conducted using spatial data mining to discover the hidden rules in the crime data happened in Czech Republic in the year 2008. Oracle data miner along with Apriori algorithm was used for identifying hidden relationship and association rules in the crime data in the form IF A AND B THEN C. The result shows that the situation about the crime perpetrated by youth differs from region to region.

Daigavane and Bajaj (2009) analyzed road traffic accident data and identified that the causes of accidents stem from different elements namely; the vehicle operator, weather, poor road conditions, age of vehicle, time duration and mechanical failure. The introduction of driver and traffic safety education into the school system was suggested as a major measure to be taken. Highway patrol with a chain of Traffic Aid Centres at intervals of 30-50 Km on highways equipped with ambulance, crane, patrol vehicle and enforcement staff with their equipment to regulate traffic and provide medical assistance to victims of accidents within the first hour of accident was also another recommendation made. Moreover, a suggestion in the design of vehicles so that they include inbuilt warning system for minimum distance between two vehicles to avoid collision is also worth mentioning proposal made by the researchers.

In (Hongguo et al., 2010) the applicability of Bayesian Network in traffic accident causality analysis is explored. In undertaking the research the structure and parameter of the Bayesian network was learnt with K2 algorithm and Bayesian parameter estimation respectively. According to the authors, the results show that the Bayesian Network can express the complicated relationship between the traffic accident and their causes, as well as the correlations among the factors of causes. It is reported that the results of the analysis provided valuable information on how to reveal the traffic accident causality mechanisms and how to take effective measures to improve the traffic safety situations.

Krishnaveni and Hemalatha (2011) also conducted a perspective analysis of traffic accident data using data mining techniques. The study deals with some of classification models to predict the severity of injury that occurred during traffic accidents. Naive Bayes Bayesian classifier, AdaBoostM1, Meta classifier, PART Rule classifier, J48 Decision Tree classifier and Random Forest Tree classifier were compared for classifying the type of injury severity of various traffic accidents.

According to the authors, the final result shows that the Random Forest outperforms the other four algorithms.

Another study (Chang et al., 2012) applied nonparametric multivariate adaptive regression splines (MARS) modeling technique to explore the effects of non-behavior factors in Taiwan. The factors considered in the study include highway geometrics characteristics, traffic factors as well as environmental conditions on freeway accidents. The authors indicated that, that horizontal alignment, vertical alignment, average daily traffic volume (ADT), heavy vehicle ADT and annual precipitation have nonlinear effects on freeway accidents. It is also indicated, that the research revealed underlying relationship between risk factors and vehicle accidents.

An application of Factor Analysis on Road Traffic Accident was explored in (Haixia and Zhihong, 2010). The paper analyzes the causes of 372 traffic accidents that occurred in China by factor analysis. According to the authors, five main factors are extracted through the research process and corresponding explanations are given, which can provide not only strategic support for traffic control department, but also some warnings to perpetrators. Li et al., (2011) analyzed road accident data to partition highway roads to avoid the occurrence of accidents. They employed fuzzy k-means clustering to classify numerical data of accidents for producing numerical clustering membership, and produce categorical memberships using values of corresponding categorical attributes, which was followed by using clustering ensemble to merge all clustering memberships to solve the sole clustering. According to the authors the results showed that cluster ensemble is effective and could be used to avoid occurrence of traffic accidents.

Another study by Nayak et al., (2011) presents a data mining methodology using decision trees for modelling the crash proneness of road segments using available road crash data. The models quantify the concept of crash proneness and demonstrate that road segments with only a few crashes have more in common with non-crash roads than roads with higher crash counts. They also examine ways of dealing with highly unbalanced data sets encountered in the study.

Jinlin et al. (2008) proposed a three-layer analysis system based on spatial data mining of GIS. Through the paper, the authors introduced the method of developing traffic accident analysis system by using ArcGIS Engine and C#.NET and gave the class realization of system main functions. Pakgohar et al. (2011) explored the role of human factors on incidence and severity of road crashes in Iran. The study explains driver's responsibility on an occurrence of an accident. Accordingly, the result of the study indicates the important role of human factor such as "Driving License" and "Safety Belt" in severity of accidents in Iran. The study employed descriptive analysis; Logistic Regression, Classification and Regression Tree. Chang and Wang (2006) used classification and regression tree (CART) to analyze the 2001 accident data for Taipei, Taiwan. More specifically a CART model was developed to establish the relationship between injury severity and driver/vehicle characteristics, highway/environmental variables and accident variables. It is reported that the most important variable associated with crash severity is the vehicle type. Pedestrians, motorcycle and bicycle riders are identified to have higher risks of being injured than other types of vehicle drivers in traffic accidents.

Computational intelligence methods for information understanding and management were presented in (Duch et al., 2005). The major software tool used was DataMiner. In addition to that, a large library written in C++, called InfoSel++, implementing different methods for feature selection, have been developed. As reported by the authors, the methods are based on information theory, distance between probability distribution and statistical approaches. The authors also indicated that dimensionality reduction based on multidimensional scaling (MDS) is another unexplored technique. It is an algorithm basically for data visualization. Besides, feature selection, the authors also experimented different algorithms like support vector for clustering the breast cancer data and Principal component analysis (PCA) for visualization. In (de Oña et al., 2011) the possibility of using Bayesian Networks (BNs) was explored in classifying traffic accidents according to their injury severity. Accordingly, they presented an analysis of 1536 accidents on rural highways in Spain, where 18 variables representing contributing factors used to build 3 different BNs that classified the severity of accidents into slightly injured and killed or severely injured. Finally, the variables that best identify the factors that are associated with a killed or seriously injured accident namely accident type, driver age, lighting and number of injuries, were identified by inference.

Using 5-year data from interstate highways in Indiana, another research (Anastasopoulos and Mannering, 2011) explored fixed and random parameter statistical models. The study used detailed crash specific data and data that include the injury outcome of the crash but not other detailed crash-specific data (only more general data are used such as roadway geometrics, pavement condition and general weather and traffic characteristics). The analysis showed that, while models that do not use detailed crash-specific data do not perform as well as those that do, random parameter models using less detailed data still can provide a reasonable level of accuracy. Another research by Pei Liu (2009) investigated a self-organizing feature maps and data mining based decision support system for liability authentications of traffic crashes in Taiwan. Through the study, the author develops a decision support tool for liability authentications of two-vehicle crashes based on generated self-organizing feature maps (SOM) and data mining (DM) models. According to the author, although with small data size, the decision support system was considered capable of giving reasonably good liability attributions and references on given cases.

In (Delen et al., 2006) a series of artificial neural networks were used to model potentially non-linear relationships between the injury severity levels and crash-related factors. In the process, the authors conducted sensitivity analysis on the trained neural network models to identify the prioritized importance of crash-related factors as they apply to different injury severity levels. According to the authors, the results, mostly validated by the findings of previous studies, provide insight into the changing importance of crash factors with the changing injury severity levels.

A study in Iran (Kashani and Mohaymany, 2011) identifies factors influencing crash injury severity on two-lane, two-way roads using 3 years data. The research used Classification and regression trees (CART), which is one of the most common methods of data mining. According to the authors, the problem of three-class prediction was decomposed into a set of binary prediction models, which resulted in a higher overall accuracy of the predictions of the model. The research reported that the prediction accuracy of the fatality class, which was nearly 0% in some of the previous studies, increased significantly. As per the objective of the research, the findings of the experiment indicated that improper overtaking and not using a seatbelt are the most important factors affecting the severity of injuries. In another study, Savolainen et al., (2011) assessed and summarized the evolution of research and current thinking as it relates to the statistical analysis of motor-vehicle injury severities, and provides a discussion of future methodological directions.

Morgan and Mannering (2011) used a mixed logit analysis to assess the effects that age, gender, and other factors have on crash severities by considering single-vehicle crashes that occurred on dry, wet, and snow/ice-covered roadway surfaces. The results showed that there were substantial differences across age/gender groups under different roadway-surface conditions. Some of the interesting patterns from the research include, for all females and older males, the likelihood of severe injuries increased when crashes occurred on wet or snow/ice surfaces but for male drivers under 45 years of age, the probability of severe injuries decreased on wet and snow/ice surfaces relative to dry-surface crashes, as reported by the authors. The authors argue that, this and many other significant differences among age and gender groups suggest that drivers perceive and react to pavement surface conditions in very different ways, and this has important safety implications. Furthermore, the empirical findings of the study highlighted the value of considering subsets of data to unravel the complex relationships within crash-injury severity analysis.

With respect to data quality, Januzaj (2009) presented an application of data mining technologies based on clustering, subspace clustering and classification in identifying data quality problems. The authors claimed that the proposed approach was efficient in data quality problems in a case study of a financial data. The major quality problems identified were wrong entries, zero and empty fields and doublets. In another study, Chen et al (2009) studied the data quality of Chinese Materia Medica (Cmm) data warehouse by focusing on the problems of data integrity, accuracy and proposed the method of workflow control. As per the authors, data quality control should be carried out from three aspects such as management, workflow and technology.

Farzi and Dastjerdi (2010) examined the use of data mining for measuring the quality of data. The authors introduced a method, which uses data mining to extract some knowledge from database, and then uses it to measure the quality of input transaction. Accordingly, an algorithm with three steps was proposed, which calculates the data quality of transaction; extract association rules, which depend on input transaction (T) and are adapted by the functional dependency, separate compatible and incompatible association rules and finally calculate the quality of input transaction.

In (Xiong et al., 2006) noise removal techniques to enhance data analysis in the presence of high noise levels was studied. Accordingly, they explored four techniques, three of which are based on traditional outlier detection techniques; distance-based, clustering-based, and an approach based on the Local Outlier Factor (LOF) of an object. The fourth technique was hyperclique-based data cleaner (HCleaner). The techniques were evaluated in terms of their impact on the subsequent data analysis, specifically, clustering and association analysis. Through the experiment, the authors reported that all of these methods provide better clustering performance and higher quality association patterns as the amount of noise being removed increases, although HCleaner generally leads to better clustering performance and higher quality associations than the other three methods for binary data.

In another study (Saunier et al., 2011), collision factors by mining microscopic data (road user's trajectories) about all traffic events with and without a collision were investigated. A free and open source tool, TANAGRA, was used to conduct the experiment on video recordings of traffic conflicts and collisions collected at one signalized intersection. Decision trees, the K-means algorithms and hierarchical agglomerative clustering methods were employed to analyze the data. The research revealed that decision tree confirms the importance of the evasive action in interaction outcomes.

To summarize, given the magnitude of the road safety problem, researches on accident data analysis are limited at least in a local context. This is true especially in the case of researches related with data quality and combining models for better result. Empirical studies considering data quality and understanding are still insufficient. On the other hand, there is an understanding that all counter measures should follow from data analysis. In connection to this, again, even the data collected is not complete enough to explain all necessary patterns. Thus, this implies that more works are required on how better to collect accident data and analyze it, which is a research problem that this work addresses.

Thus, from the survey of literatures made and to the knowledge of the researcher, there is no other research made to disclose the role of road user's (drivers, pedestrians and victims) behaviour in accident severity to explain the road safety situation in Ethiopia. Moreover, no model ensemble and trend analysis was conducted so far to explore the variance in the determinant factors and improvements in accuracy of the analysis models in Ethiopia. It is also noted that predictive data mining using classification is accepted and appropriate approach in the domain understudy. Accordingly, this research investigates accident data in explicating issues in the above mentioned gaps. The research also go further defining road safety information architecture in addressing issues illuminated through these experiments.

4.4 Accident Data Collection, Analysis and Dissemination

In addition to researches using a varied approaches and perspectives discussed above, there exist also research efforts addressing road safety through improving accident information management. After a thorough analysis of the existing accident data collection and analysis practice in Hanoi, Vietnam (Hai, 2009) suggested methods to improve the practice. Some of the major problems cited were;

> "*insufficient accident data capture and unestablished analysis criteria, fragmental traffic safety management bodies' functions (distributed to many entities) and not converged resulting in divided responsibilities for safety and accident statistics, underreporting of many of traffic accidents with above 11% in damages, delayed facility establishment.*"

While the above listed problems are major ones identified, improvements suggested and implemented include; establishment of accident database, improving the data collection form by including more details and providing continuous training on importance, understanding and interpretation of accident data for responsible parties. Finally, the author stated that the most important achievement of the research project was to activate a cycle process which includes: scene investigation, data capture, update data among police teams and store data in Traffic Police Division, analyze collected data and implement needed tasks. Thus, at the simplest sense this provides a general framework of accident data collection and analysis processes.

It is also indicated that future work should address mechanisms and frameworks under which accident database can be analyzed in detail and analysis result be shared among relevant agencies as important foundations for traffic safety measures. Moreover, the author stressed that periodical revision of the data collection form and establishment of analysis requirement and criteria should also get more attention.

In another work (Jarvis and Kamal, 2009) proposed a software solution, Crash Data System, for improved national systems based on a crash data reporting framework. The major problems they tried to address includes, under reporting, tiresome accident investigation and data collection process. Regarding the issue of data on road crashes, the authors recommended some basic actions. Establishment of a procedure and methodology for multi-sectoral accident investigation, data collection, reporting, analysis, transmission and define the roles of agencies involved in the process is the major recommendation forwarded. In addition the need for the establishment of a reliable database to define the problem and measure progress through annual performance benchmarking was also emphasized in the research.

Another work (Afukaar, 2007) presented at the African Road Safety Conference, explained road traffic injury data system in Ghana as a key to safety improvement and control. In his work the author identified three (see Figure 3) essential components of a crash/casualty data system as; Standard accident report form for data collection, a means of data storage and retrieval, and a means of data analysis and dissemination. The argument is that attention should be given to these three aspects for improvement in a road safety. According to the author, in Ghana, road traffic crash/casualty data are managed using a Microcomputer Accident Analysis Package (MAAP) developed by the UK Transport Research Laboratory (TRL). The main sources of road traffic crash/injury data in Ghana are, police data, hospital and medical data, Insurance data and special surveys.

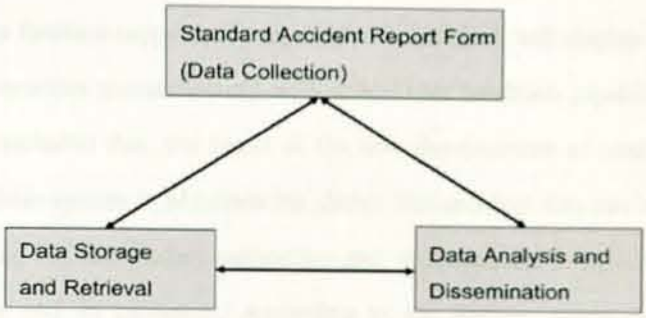


Figure 6 : A Simplified Traffic Injury Data System (source: (Afukaar, 2007))

The data content includes: general information regarding where, when, how did the crash/injury occur; road users involved including who was involved and why did the crash/injury occurs; road environment which consists of where and under what circumstances did the crash/injury occur, and vehicles involved how did the crash occur. With regard to data analysis and applications, basic analysis criteria includes fatality trend, road users involved and at risk, place and time of traffic injuries and major risk factors. Finally, the author concluded that road traffic injury data systems are the key to safety improvement and control in Ghana.

Realizing the absences of road safety data system, at a national level, due to the current system structure where each police station in the 145 districts of the country, maintaining individual database, (Hizal Hanis and Sharifah Allyana, 2009) proposed a comprehensive database system for Malaysian accident data management. The major problem identified through the process were data quality problem, absence of data integrity and checking at the district level thus making the compilation at the national level troublesome and difference in reporting style. Accordingly, the national accident data system which is developed using MySql database management system enables the integration of the data from different districts which actually provide an environment for data quality check. Moreover, a lightweight analysis system was also developed to support the analysis aspect of the data.

The main analysis function supported is cross-tabulation which will display the result of two accident variables simultaneously with optional set condition capability. The authors finally concluded that, the result of the new development of road accident analysis and database system in Malaysia has shown that accident data can be used as the primary source for road safety authorities and professionals to understand the accident situation and its causation. According to the authors, focus or targeted programs can be planned to give effective treatment to specific road safety issues. Moreover, the road accident data can be used as evidence in evaluating the progress of road safety intervention as well as for performance monitoring. It is also mentioned that, by having a well established and comprehensive set of road accident database system, it will aid in the formulation of effective road safety research and interventions thus providing a more promising result in reducing both road accidents and fatalities.

Another study by Bener et al. (2003) reviewed the road safety situations in developed and developing countries compared with the Arabian Gulf countries and suggests a strategy to improve it. The strategy stresses the importance of improving systems for data collection and analysis; establishment of official, national coordinating committee and a small team working on fulltime on road safety; training traffic engineers to analyze accidents, conduct safety studies to identify most important problems; convince officials to take actions and design and evaluate counter measures.

A proposal and formulation of a common framework for road accident data collection across European countries in (Yannis et al., 2009) was another worth mentioning practice regarding accident data management. Mentioning data quality, incompleteness, underreporting and lack of standard, and a need for harmonization of accident data, the authors proposed a common road accident data set (CADas) for the European countries. According to the authors, thorough analysis of practices in member countries and needs of stakeholders were considered in the formulation of the accident data system. The authors stated that national data collection systems from all EC member states were analyzed. It is indicated that, through an iterative process and the feedback received from several national road accident data experts, Common Accident Data Set consisting of 73 variables and 471 values, grouped into four basic categories namely accident, road, traffic unit, person were defined along with the structure of the Data Set. Finally, the authors recommended a pilot phase for the implementation of the CADaS as a further step to be considered in addition to directing CADas as a possible basis for the development of the respective common data set to be used at global level (World CADaS).

As can be seen from the discussion the focus of existing researches and development efforts tends to technological solution through developing databases to the accident data system. Without denying the importance of technological solutions this research contemplates on employing a wider enterprise perspective on accident information management.

4.5 Summary of the Chapter

The survey of literature in this and previous chapter exhibited that systematic and enterprise view is lacking in accident information management. Though, there is good understanding of the importance of road safety information systems by various authors, their focus was limited to a single aspect of the whole system. Systemic and integrated view is lacking towards addressing the problem and it still is short of an architectural framework that guides the development and management of accident data collection and analysis systems. Hence, the current research investigates the use of Zachman Framework to improve the management of accident data in the local context.

Chapter Five

Data Analysis and Experimentation

5.1 Overview

The purpose of this Chapter is to explore and investigate road accident data, and determine information requirement of road safety information management. Accordingly, data mining experiments helps in identifying optimal attributes and exhibit magnitude of data quality problems which informs the content and process dimensions of the proposed architecture. A qualitative data collection and analysis also helps in determining the information requirements of the domain. In line with this, the second section presents machine learning experimentation and results. The third section discusses the result of qualitative data analysis and findings. The fourth section presents interpretation and conclusion formed from the data analysis and experimentation.

5.2 The Road Safety Situation: Experimentation and Results

5.2.1 Introduction

This sub-section discusses findings of machine learning experiments and trend analysis on the accident information as presented in (Beshah et al., 2011b, Beshah et al., 2012b, Beshah et al., 2012a). The purpose of this experiment is threefold. Firstly, it helps to reveal interesting patterns in terms of determinant factor identification in the process of explaining road safety situations and model comparisons. It also empirically supports the assertion that there is a need for integrated road safety information architecture to address the data quality, analysis requirements and related issues. Last but not least, it helps to explicate the variation in the determinant factors of accident occurrence and severity.

Accordingly, two research questions of this thesis, mentioned below, are addressed through the analysis of accident data and synthesis of previous experiments.

RQ1-What does the road safety situation in Ethiopia look like?

RQ2-What appropriate analytical models can be developed for analysis of road traffic accident data?

In addressing this research question, we define two objectives with sub-objectives as presented below.

- *To explore the status of road safety situations in Ethiopia.*
  - i. *To explain and predict the role of road users' and collision related factors on possible injury risks.*
  - ii. *To explore accident analysis trend on factors affecting accident severity*
- *To construct analytical machine learning models in the analysis of road traffic accident data.*

Though the main research covers three administrative regions in Ethiopia, this particular experimental study used data obtained from Road Traffic Office at Addis Ababa, Ethiopia., The total dataset for the study contains traffic accident records from 2004/5-2008/9. Based on the availability of the data, for this specific study a total number of 14,254 accident cases described with 48 attributes are used. According to the variable definitions for dataset, this dataset has information related to road users (drivers, pedestrians and passengers), vehicles and road environment.

Road safety can be explained from various perspectives. Identification of problem locations through the analysis of accident data is identified as a first step in an accident reduction program (Bener et al., 2003). In connection to this, one of the major aspects as mentioned in the introduction part of the thesis is road accident occurrence and severity analysis. In such type of work, the critical issue is identifying determinant factors. Factors could be driver, road, environment, collision, pedestrian or passenger related. It is indicated in literature that, identification of major problems contributing to accidents or injuries, which are amenable should be the major objective of road safety analysis. This assertion is also in line with a discussion in (Elvik et al., 2006) cited in (Eksler, 2007).

In the process of addressing the objectives of the research indicated above, two major phases of experiments were conducted. In the first experiment, attempt has been made to use two of the available predictive modeling methods; classification and regression tree (CART) and RandomForest. The objective of this machine learning experimental research is, to explore and predict the role of road users' related factors on possible injury risks with two target class labels. Moreover, addressing role of collision related factors and getting sense of the data are also other motives for the first experiment. In the second experiment, attempt has been made to use three of the available predictive modeling methods, classification and regression tree (CART), TreeNet, and RandomForest with three target class labels. In addition, a parallel configuration of combining models with a majority vote approach is used as an ensemble technique.

5.2.2 Exploration of Data Quality Issues

Data quality is one of the major concerns in organizational decision making. Especially, from machine learning and data mining point of view, where extracting pattern and knowledge discovery is a major task, it is an issue that needs closer attention. Though data quality can be seen from different perspectives, the focus of this experiment lies on data quality issues at the analysis level, affecting knowledge and pattern discovery. It is explained based on the road safety as a case study. Accordingly, in (Tan et al., 2006) three major problems related to data quality in machine learning environment; noise and outliers, missing values and duplicate data are identified. Noise and outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set like "unknown" values. "Unknown" value is usually used to fill an attribute value with no actual data. On the other hand, while the major reasons for missing values are inapplicability of an attribute for all the cases or due to inability to collect a specific value for an attribute because of different reasons, duplicate data can occur due to the lack of effective design of attributes for data objects.

The process of handling these data qualities in general is referred as data cleaning or preprocessing. However, data cleaning or preprocessing will depend and be restrictive as per the data mining or knowledge discovery task at hand. Thus, this work argues that data quality issues should be addressed at a different level right from the collection to the dissemination. This is reflected on the information architecture proposed as a final deliverable of the main research. However, the magnitude of data quality issues at the analysis level, in a road safety data management are explored and presented. It is easy to learn from the details that the three major data quality problems mentioned above are prevalent in the road accident dataset. In connection to this, though there are different noises and outliers in a given data one of such examples, the "unknown" value is picked as an example to show the magnitude of the problem with respect to noise and outliers. In the process of explaining data quality issues, variables exhibiting 0.5 % and above missing values are presented in Table 6. It can be seen that variables related with road users show considerable missing values, which can affect the size and quality of pattern to be discovered.

Table 6- Percentage of Missing Values

S.N	Variable	% Missing
1	*VEHICLETECHSTATUS*	1.3%
2	*VICTIMAGE*	0.74%
3	*VICTIMCATEGORY*	0.57%
4	*VICTIMHEALTHST*	0.66%
5	*VICTIMOCCUP*	0.64%
6	*WEATHERCONDITIONS*	1%
7	*PEDSTRIANMOVEM*	81%
8	*ROADCONDITION*	0.41%
9	*VEHICLEPLATE*	13%

As to the unknown values, variables with their percentage of unknown values are presented in Table 7. It is easy to understand that by improving the data quality while collecting accident data through quality checks, it is possible to achieve better prediction and more relevant knowledge. It is visible again that properly addressing these issues add on the performance and accuracy of data analysis. In addition to missing values under existing variables, another important attribute missing is use of seatbelt/helmet. Though the use of seatbelt and/or helmet is considered to be one of the important measures in reducing accidents and fatality, it is not included in the accident data. The duplication issue is exhibited with variables related with accident date. There are three attributes; year, month, week, which can be expressed only by proper data structure of date variable itself.

Table 7- Percentage of Unknown Values

S.n	Variable	% Unknown
1	*VICTIMHEALTHST*	0.60%
2	*VICTIMOCCUP*	0.72%
3	*DRIVINGLICENS*	8.82%
4	*PEDSTRIANMOVEM*	0.27%
5	*VICTIMAGE*	0.01%
6	*DRIVINGEXP*	9.11%
7	*VICTIMCATEGORY*	0.01%
8	*ACCUDRIVEHIRELATION*	9.32%
9	*DRIVERAGE*	9.39%
10	*VECHILEMOVEMENT*	0.06%
11	*DRIVERSEX*	9.74%
12	*ACCUDRIVEDULEVEL*	9.30%

In data mining, some level of missing vales are common and tolerable (less than 1%) and manageable though it affects the result (up to 15%), however missing values above 15 % impact the analysis and interpretation (Acuna and Rodriguez, 2004).

5.2.3 First Phase Experiments and Results

Data preparation or preprocessing is always important in machine learning and pattern recognition process. Though, there are various types of preprocessing tasks like handling missing values, minimizing noises, dimensionality reductions, attribute aggregations, feature creation, discretization and linearization, attribute transformation, sampling and feature selection, given the motivation of this particular experiment the first attempt is to expose the data for pattern identification and knowledge discovery with a minimum pre-processing tasks.

Accordingly, the data, which was in a relational database format is first exported in to a single table format of excel sheet. In addition, it is also necessary to translate the data from local language, Amharic, into English for better readability using the filter facility of Ms-Excel application. Moreover, removal of some attributes for ethical reason and their unnecessary nature in the process of pattern identification and attribute creation through aggregation of attribute values of injury severity was done.

5.2.3.1 Model Building - Role of Road Users Factors

The first task of the experiment is to understand and heuristically identify attributes or features related to the goal of the machine-learning task, which will obviously be evaluated by the machine learning process through attribute selection. The best explanation of the data obviously depends on the type of the problem, intention of users, as well as the type of questions and explanations that are commonly accepted in a given domain (Duch et al., 2005). Thus, given the data mining task mentioned above, 12 road user related attributes were selected initially as possible predictors, where accident collision result being a target class. This is to mean that the objective of this specific experiment i.e. identifying role of road users factors, heuristically guides the attribute selection. For better understanding of the result, the values for target class were aggregated into 'injury' and 'non-injury'. Descriptions of the attributes are presented in Table 8.

Table 8- List and Description of Possible Predictors

S.N	Attributes	Description
1	*PedstrianMovem*	Pedestrian movement during the accident
2	*VictimHealthSt*	Health condition of victims
3	*VictimOccup*	Occupation of Victims
4	*VictimAge*	Age of victims
5	*DrivingLicens*	Driving license level of a driver
6	*VictimCategory*	Category of victims
7	*VechileMovement*	How the driver was driving the vehicle
8	*DrivingExp*	Driving experience of the driver
9	*AccuDriVehiRelation*	Relationship b/n a vehicle and a driver
10	*AccuDrivEduLevel*	Educational level of a driver
11	*DriverAge*	Age of a Driver
12	*DriverSex*	Sex of a driver
13	*AccidentResult (target class)*	Whether a collision ended with injury or non-injury

5.2.3.2 CART Analysis Result-Role of Road Users Factors

While running the CART analysis, the attribute selector module identifies *PedstrianMovem, VictimCategory, VehicleMovement, VictimOccup, VictimAge, DrivingLicense, DrivingExp* and *VictimHealSt* variables as important predictors of the target class injury result (risk). With the intent of finding the best prediction, a number of experiments have been done by trying different constraints and parameters. Accordingly, given the purposeful low level of preprocessing done, using these variables with major model specification like 10 fold cross validations for testing, and automatic best predictor discovery, the accuracy of the predictive model is promising. As this experiment uses a real world data where target class imbalance is common, a mechanism called *Priors Equal* is used in this specific experiment (Steinberg and Golovnya, 2006). This is a facility to provide equal probability for all categories. Road user factors like pedestrian movement, victim's category and victim age are found to be determinant whether an accident ends with injury or not, and it can be seen from major splitters as illustrated in Figure 7.

The overall prediction success, which is a percentage of correctly classified against the total data set, is 92.11% for learning set while 91.51% is for testing set. It is also visible that the prediction accuracy for non-injury class is better than the injury class in both learning and testing sets. As can be seen from the details the performance is good as the usual goal is to get as close as to 100%. On the other hand heuristically if the classifier has an accuracy of less than 60%, then it can be said undesirable. The detail is shown in Table 9. It is obvious that in such kind of experiment, the accuracy of learning process is better, which also have been the case for this specific experiment.

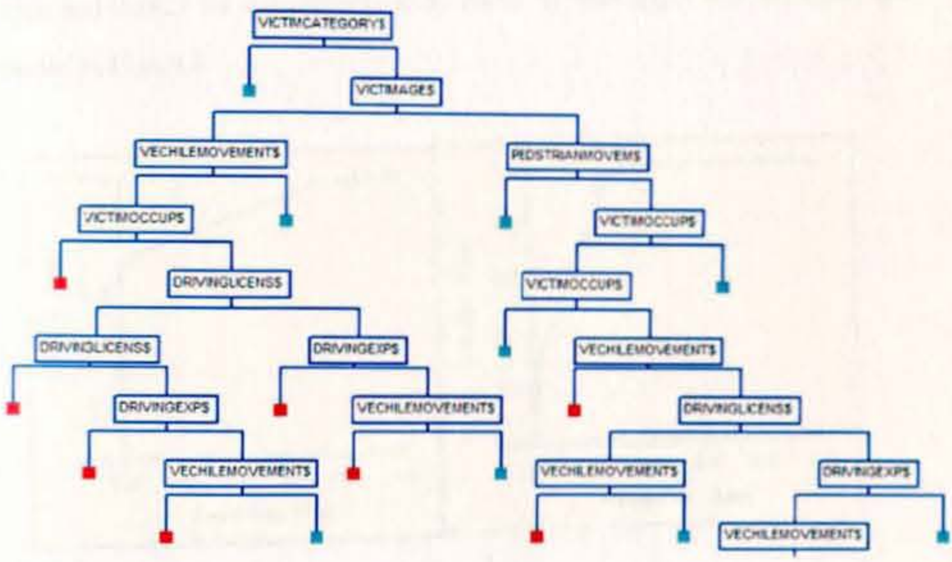


Figure 7: Splitter variables

Table 9 - Learning and testing prediction success

Actual Class	Predicted Class						
	Total Class	*Percent Correct*		***Non-injury***		***Injury***	
	Learn/Test	*Learn*	*Test*	*Learn* *N=11778*	*Test* *N=11716*	*Learn* *N=2476*	*Test* *N=2538*
Non-Injury	10,970	98.56	97.89	10,812	10,738	158	232
Injury	3,284	70.58	70.22	966	978	2,318	2,306
Average		84.57	84.5				
Overall Correct		92.11	91.51				

However, accuracy alone does not completely describe the prediction efficiency, and hence, other means of evaluating our predictive models are necessary. The receiver operating characteristics (ROC) curve, also known as, the relative operating characteristic curve, is a comparison of two operating characteristics as the criterion changes. The area under the ROC curve (AUC) quantifies the overall discriminative ability of a test. An entirely random test (i.e., no better at identifying true positives than flipping a coin) has an AUC of 0.5, while a perfect test (i.e., one with zero false positives or negatives) has an AUC of 1.00 (Beshah and Hill, 2010).

Accordingly, with respect to the ROC in this specific experiment, it scored 0.8873 for training and 0.8827 for test scenario. ROC charts for both injury and non-injury are presented in Figure 8.

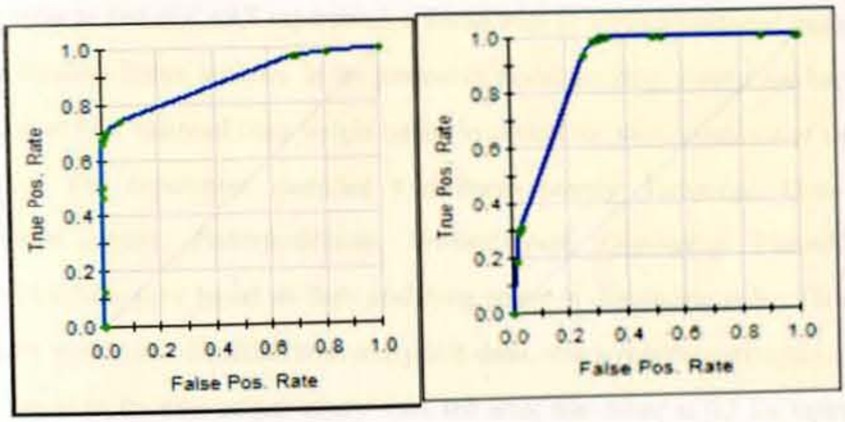


Figure 8: ROC: injury (left) and non-injury (right)

Another important concept regarding the performance of a predictive model is miss-classification rate in terms of error and cost. Misclassification cost is a fraction of product of cases misclassified and cost assigned for misclassification. While the default is one, it is also possible to vary the assignment based on the recommendation of the domain experts.

As it can be seen from Table 10, the model is better in predicting non-injury results than injury. This is clear from the error percentage and cost values in the table against both classes.

Table 10- Misclassification of CART model

Class	Misclassification						
	N *Cases*	*N* *Mis-Classed*		*Pct. Error*		*Cost*	
	Learn/Test	*Learn*	*Test*	*Learn*	*Test*	*Learn*	*Test*
Injury	3,284	966	978	29.42	29.78	0.29	0.30
Non-injury	10,970	158	232	1.44	2.11	0.01	0.02

5.2.3.3 Random Forest Analysis and Result -Role of Road Users Factors

Similar to that of CART experiment, a model with 12 attributes selected was exposed to Random Forest analysis. In the process of model building, attempt has been made to specify a balanced class weight option to control the data imbalance of the target class. The experiment identified 8 attributes namely; *VictimAge, VictimOccup, VictimCategory, PedstrianMovem, DrivingLicense, DrivingExp, VictimHealthSt, VehicleMovement* based on their predicting power in descending order. Thus, using these variables, a RandomForest analysis is done, which results in error rate closer to 0 (zero) in the case of non-injury class and error rate closer to 0.3 for injury class. This implies that, this method is better in classifying non- injury class than injury class. The performance of a predictive model error rate lies between 0 and 1. The average error rate is presented in Figure 9.

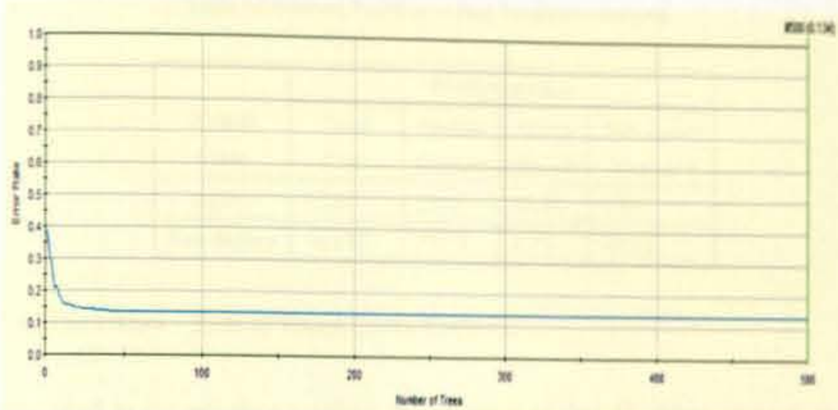


Figure 9: Error Rate Result (overall)

With respect to miss-classification, somehow similar to that of CART analysis result, random forest analysis is also less accurate in classifying injury category, while the miss classification rate is by far less for non-injury category. This is shown with 26.55 % classification error for injury class while it is 0.26 % for non injury class. It is also worth mentioning that the cost, which is a fraction of cases misclassified, multiplied by cost assigned for misclassification in this case 1, is higher for injury class than non-injury. The detail is presented in Table 11.

Table 11-Misclassification Result of RandomForest

Actual Class	Internal Test/Out of Bag			
	N Cases	*N Mis-Classed*	*Pct. Error*	*Cost*
Injury	3,284	872	26.55	872.00
Non-injury	10,970	29	0.26	29.00

Prediction success and ROC results are also important indicators of a given predictive model. Accordingly, percentage of correct prediction for non-injury case is 99.74% while 73.45% is for injury cases. The detail is presented in Table 12. In the same token, the ROC integral indicates, 0.90070, which is closer to one and it indicates minimal zero positives and negatives, which entails good performance.

Table 12-Internal Test/Out of Bag Prediction Success

Actual Class	Predicted class			
	Total Class	*Percent Correct*	*Injury N=244*	*Non-Injury N=11813*
Injury	3,284	73.45	98.81	7.38
Non-injury	10,970	99.74	1.19	92.62

5.2.3.4 Discussion - Role of Road Users Factors

As presented in the previous sections, the purpose of this 1st phase experiment is to identify some new patterns and get an overall understanding of accident data. To this end, one of the patterns that require empirical investigation is the impact of road user's related factors on accident injury risk. It has been learned that road user related factors need more investigation, so as to guide proactive methods in reducing road accident and improving road safety in general. In line with this, the subsequent experiments resulted in more patterns.

Though, the emphasis by domain experts is on drivers as a major contributor for accidents, pedestrian behavior and victims features like their category and age tell significant information about the possible result of a given accident occurrence. This, along with the subsequent experiments, will be used in the design of education and law enforcement measures in road safety domain. Moreover, CART and RandomForest predictive miners are found to be promising in the identification of patterns in a road safety domain. While comparing the two, they both perform less in the case of injury class, while their performance is very good in determining non injury risk of an accident. However, RandomForest predictive modeling technique performs better by exhibiting lower error rate, higher ROC score and greater prediction accuracy than CART.

5.2.3.5 Role of Collision Related Factors - J48 and PART Analysis

In the process of explaining an aspect of the road safety situation, a couple of experiments are conducted. Given the data mining task mentioned above, 10 collision related factor attributes are selected to be used as predictor variables, Accident Severity being a target class with three values ; *'fatal', 'injury' and 'non-injury'*. This is to mean that the goal of this specific experiment i.e. identifying role of collision related factors, heuristically guides the attribute selection. Descriptions of the attributes are presented in Table 13.

Table 13- List and Description of Predictor Variables

S.N	Attributes	Description
1	*AccArea*	*Specific area of an accident*
2	*WeatherCon*	*Weather condition at the time of accident*
3	*LightCon*	*Light condition while accident occurs*
4	*VechMovement*	*How the driver was driving the vehicle*
5	*AccType*	*Accident collision type*
6	*AccCouse*	*Immediate cause of an accident*
7	*AccK/K*	*Sub city where an accident occurs*
8	*AccTime*	*Time of an accident*
9	*AccDay*	*Day of accident*
10	*AccSeverity (target class)*	*Whether a collision ended with fatality, injury or non-injury*

As stated above, with the intent of finding the most interesting patterns and rules in explaining the road safety situation, a number of experiments have been done using J48 Decision tree algorithm and PART rule induction algorithm with different constraints and parameters. J48 algorithm is a top-down decision tree learner that uses the greedy technique (Daud M.N.R., 2007). PART is a separate-and-conquer rule learner producing sets of rules called 'decision lists' which are ordered set of rules proposed by (Witten et al., 1999).

The first experiment is done using J48 decision tree approach with major model specification like 10 fold cross validations for testing, and pruning applied. As a result, a 60 tree size with 55 number leaves were identified best with the overall accuracy of 92.21% which is a percentage of correctly classified against the total data set. It is also visible that the prediction accuracy for non-injury class is better than the injury and fatal classes. It is also worth mentioning that, *accident type, accident cause* and *accident sub-city* are found to be the three top determinant factors in predicting severity. This echoes the findings of previous research (Tesema et al., 2005).

As stated in the methodology section, the second approach selected is the PART Rule induction algorithm. Consequently, experiments have been conducted using the same data set with 10 fold cross validation and pruning. PART has generated 91 rules within 0.64 seconds. The accuracy (correctly classified instances compared to the total data set) of the model is 92.30%. The weighted average ROC area covered was 0.994. A screen shot of the PART result containing some of interesting rules is presented in Figure 10.

```
AccType = pedestrianhit AND
AccCouse = denyingpedestriani%spriority AND
WeatherCon = cold AND
AccTimeF = 6pm-12pm: fatalinjury (7.0)

AccType = pedestrianhit AND
AccCouse = denyingpedestriani%spriority AND
AccK/K = kolfe: Injury (54.0/18.0)

AccType = pedestrianhit AND
AccCouse = denyingpedestriani%spriority AND
LightCon = eveningroadlight: fatalinjury (52.0/20.0)

AccK/K = akaki AND
LightCon = daylight AND
AccCouse = denyingpedestriani%spriority AND
AccTimeF = 6pm-12pm: Injury (17.0/7.0)
```

Figure 10: Sample Result of PART experiment

The finding that highlights denying pedestrian priority as one of the major accident cause in determining fatality and injury is in disagreement with severity analysis research in Iran (Kashani and Mohaymany, 2011) where improper overtaking and not using a seatbelt are the most important factors affecting the severity of injuries. This could be taken as an additional argument for the need of context specific experiments.

As evident from the details above, though both approaches showed good result, PART is better in both accuracy and ROC area. Table 14 shows model comparison using accuracy and ROC area.

Table 14 - List and description of predictor variables

	Accuracy	**ROC area (weighted average)**
J48	*92.2125%*	*0.932*
PART	*92.3071%*	*0.944*

5.2.4 Second Phase Experiments and Results

The main motivation for the second phase experiment is the need to include more number of attributes and address additional issues to the 1st experiment. Expanding the target class labels into three, which is more pragmatic than the first experiment, was also another important factor.

Thus, in this part of the thesis a detailed description of data exploration task and results of different experiments are presented. Accordingly, the data which was in a relational database format was first exported into a single table format of an excel sheet. Moreover, removal of some attributes for ethical reason and their unnecessary nature in the process of pattern identification exposed a total of 38 features for many sided analysis. A screen shot of sample data is presented in Figure 11.

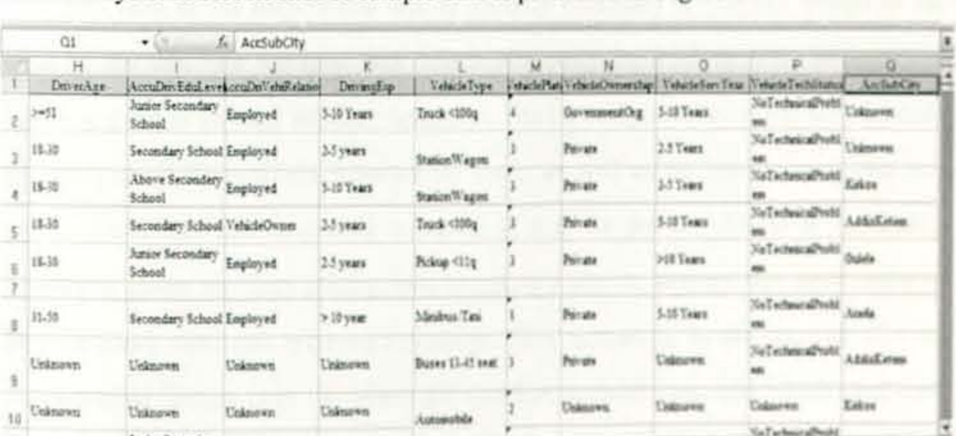


Figure 11: Sample data

5.2.4.1 Pre-processing and Model Building

Data preparation or preprocessing is always important in a machine learning and pattern recognition process. There are various types of preprocessing tasks like handling missing values, minimizing noises, dimensionality reductions, attribute aggregations, feature creation, discretization and binarization, attribute transformation, sampling and feature selection which mainly are guided by the data mining goal at hand. In light of the whole objective of the experiment, the preprocessing task for this research can be considered as lightweight preprocessing. The main reasons are the tool's capability of handling data quality issues like, missing data and the need to expose the actual data as it is.

Preprocessing tasks undertaken for this specific experiment includes; dimensionality reduction by removing records with significant variable values missing and removing of attributes that do not contribute to the analysis like *serial number, date, year* and *month*. In addition, generalization of *serious injury* and *slight injury* into *injury class*, and replacement of blank cells by "not applicable" (N/A) value for variables that do have such features when seen from the target variable respect, are also done. Categorizing some variables like *age* and *hour* into manageable categories was also done for better understandability of the pattern.

The next task of the experiment was to identify attributes or features related to the goal of the machine learning task which obviously be evaluated by the machine learning process through attribute selection. The best explanation of the data depends on the type of the problem, intention of users, as well as the type of questions and explanations that are commonly accepted in a given domain (Duch et al., 2005). However, given the data mining task mentioned above, attempt has been made to include more attributes whenever possible. This is mainly, to see the role of road users related factors over the others on accident severity risk, and which road user related factors are more important in addition to assessing the trend of impacts of factors to severity. Accordingly, based on norm reference criteria which means by improving the number of attributes to be used, 31 attributes are selected as possible predictors, *accident collision result* being a target variable. The target variable has three classes namely; fatal, injury and non-injury. Descriptions of the attributes are presented in Table 15.

Table 15- List and description of possible predictors (2nd phase experiment)

S.N	**Attributes**	**Description**
1	*ACCCOLLISIONTYPE*	Accident collision type
2	*ACCSUBCITY*	Sub city where an accident occurs
3	*VICTIMAGE*	Age of victims
4	*VICTIMOCCUP*	Occupation of Victims
5	*VEHICLETYPE*	Type of Vehicle involved
6	*VICTIMHEALTHST*	Health condition of victims
7	*ACCIDENTCOUSE*	Immediate cause of an accident
8	*VICTIMCATEGORY*	Category of victims
9	*HOURCATEGORY*	Category of accident hour
10	*ACCAREA*	Specific area of an accident
11	*DRIVINGLICENS*	Driving license level of a driver
12	*DRIVINGEXP*	Driving experience of the driver
13	*ACCUDRIVEHIRELATION*	Relationship b/n a vehicle and a driver
14	*ACCDAY*	Day of accident
15	*LIGHTCONDITION*	Light condition while accident occurs
16	*VEHICLEPLATE*	Vehicle Plate Category
17	*ROADSEPARATION*	Road Separation
18	*DRIVERAGE*	Age of a Driver
19	*ACCWEEK*	Specific week of a month
20	*VEHICLESERVYEAR*	Service year of the vehicle
21	*VECHILEMOVEMENT*	How the driver was driving the vehicle
22	*VEHICLEOWNERSHIP*	Vehicle Ownership
23	*ACCUDRIVEDULEVEL*	Educational level of a driver
24	*ROADJUNCTION*	Type of road junction
25	*DRIVERSEX*	Sex of a driver
26	*ROADORIENTATION*	Type of road orientation
27	*PEDSTRIANMOVEM*	Pedestrian movement during the accident
28	*VEHICLETECHSTATUS*	Technical status of the vehicle
29	*ROADCONDITION*	The condition of the road
30	*WEATHERCONDITIONS*	Weather condition
31	*ROADSURFACE*	Road surface type
32	*AccidentResult (target)*	Whether a collision ended with fatal, injury or non-injury

5.2.4.2 CART Analysis Result

The first experiment of the second phase in classifying the class attribute *accidentresult* was using CART technique with 31 predictor variables. While, running the CART analysis, the classification method used was entropy. Entropy is one of the different splitting rules in growing classification trees. Regarding dataset usage, 80/20 percent of the data is used for training and testing respectively. With the intent of finding the best prediction, a number of experiments have been done by trying different constraints and parameters. This includes but not limited to, Gini and class probability were tested as a method for classification while 10 fold validations was also used as a testing mechanism. With respect to best tree selection, the CART default best tree setting, which is a minimum cost tree is employed.

The best model identified indicated that victim related features namely, *VictimAge, VictimCategory, VictimOccup,* and *VictimHealSt* followed by *AccidentCollisiontype AccidentCouse, AccidentSubcity, VehicleType, Hourscategory, AccidentArea* and *DrivingLicens* are the top ten important predictors of the target, class injury result (risk). On the other hand, road and environment related factors like *RoadSerface, WeatherCondition and RoadCondition* are from the least significant factors compared to human related factors. Accordingly, given the purposeful low level preprocessing done, using these variables with major model specification and automatic best predictor discovery, the accuracy of the predictive model is promising with a general classification error of 0.300. Road user factors are found to be determinant whether an accident ends with fatal, injury or non-injury, and it can be seen from major splitters as illustrated in Figure 12.

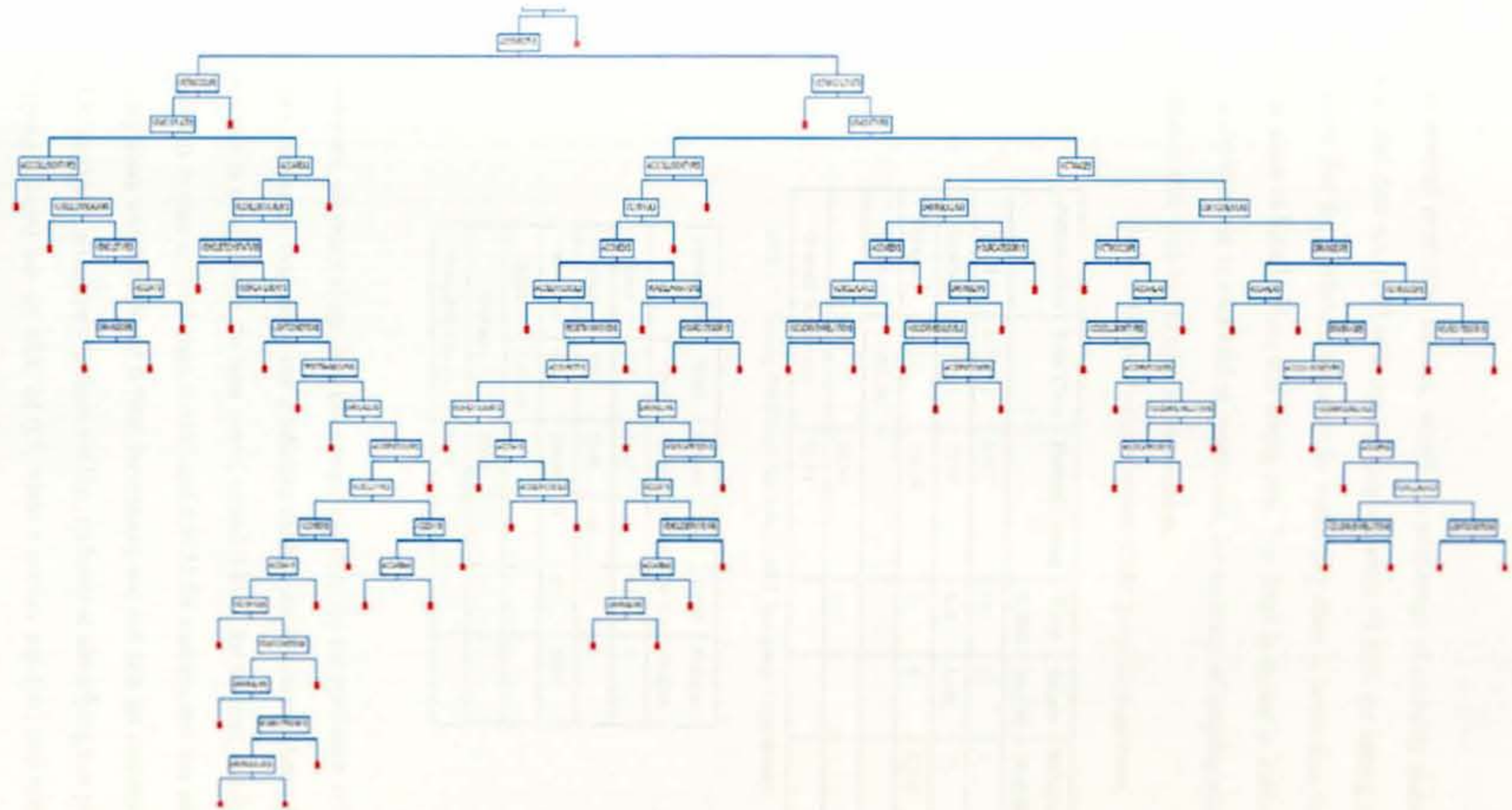


Figure 12: Splitter variables (2nd phase expert)

The overall prediction success, which is a percentage of correctly classified against the total data set, is 95.61% for learning set while 93.52% for testing set. It is also visible that the prediction accuracy for non-injury class is better than the injury and fatal class in both learning and testing sets. The detail is shown in Tables 16 and 17. It is obvious that in such kind of experiment, the accuracy of learning process is better, which is also seen in this specific experiment.

Table 16-Training Prediction success (CART 2nd phase Experiment)

Actual Class	Total Class	Percent Correct	Fatal N=865	Injury N=1747	NoInjury N=8789
Fatal	518	87.07	451	67	0
Injury	2,113	79.51	414	1,680	19
NoInjury	8,770	100.00	0	0	8,770
Total:	11,401.00				
Average:		88.86			
Overall % Correct:		95.61			

Table 17- Testing Prediction Success (CART 2nd phase Experiment)

Actual Class	Total Class	Percent Correct	Fatal N=236	Injury N=412	NoInjury N=2205
Fatal	168	66.67	112	56	0
Injury	485	73.40	124	356	5
NoInjury	2,200	100.00	0	0	2,200
Total:	2,853.00				
Average:		80.02			
Overall % Correct:		93.52			

However, accuracy alone does not completely describe the prediction efficiency, and other means of evaluating our predictive models are necessary. Accordingly, ROC analysis is used where, the best model scored 0.9772 for training and 0.940 for test scenario in case of fatal class, 0.9887 and 0.9721 for training and test sets in case of Injury class and 0.9964 and 0.9962 for training and test sets for non-injury class. In ROC analysis , an entirely random test (i.e., no better at identifying true positives than flipping a coin) has an AUC of 0.5, while a perfect test (i.e., one with zero false positives or negatives) has an AUC of 1.00 (Beshah and Hill, 2010). ROC charts for all the three classes containing both training and test cases are presented in Figure 13.

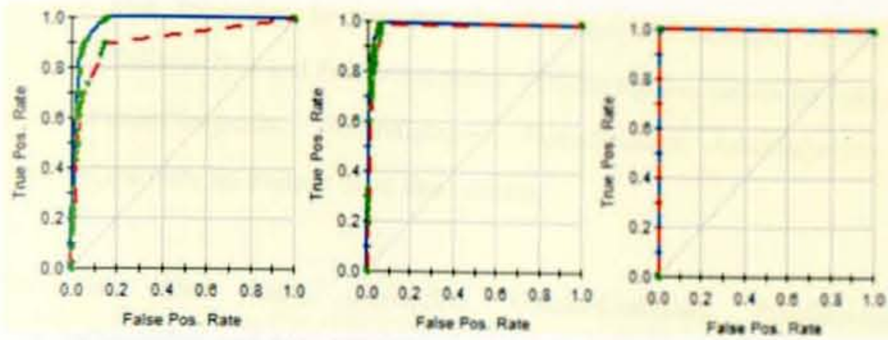


Figure 13: CART ROC charts (fatal, Injury and non-injury)

Another important concept regarding the performance of a predictive model is miss-classification rate. As it can be seen from Table 18, the model is better in predicting non-injury results than injury and fatal.

Table 18-misclassification rate for test data (CART 2nd phase experiment)

Class	N Cases	N Mis-Classed	Pct. Error	Cost
Fatal	168	56	33.33	0.33
Injury	485	129	26.60	0.27
NoInjury	2,200	0	0.00	0.00

5.2.4.3 TreeNet Analysis and Results

The second experiment was done using the TreeNet analysis method. In this experiment 31 predictor variables were used to predict the target class, *accidentresult* (risk). The analysis was done by specifying basic parameters like balanced class weights to up weight small classes to equal size of largest target classes, 80/20 percent for training and test sets respectively, and cross entropy or likelihood as a means of selecting optimal logistic model. Out of the total 31 variables, this method identified 29 of them as important predictors by excluding *roadsurface* and *roadcondition* which scored 0.00 importances.

It was also interesting to see that *AccidentSubcity*, *VictimAge*, *VehicleType*, *AccidentCollisionType* and *VictimOccupation* were the top five factors for Fatal class while *VictimOccupation*, *VictimeCategory*, *VictimHealthSt*, *AccidentSubcity* and *VehicleType* were for average of all three classes.

On the other hand, *DriverSex*, *WeatherCondition*, *RoadSeparation*, *VehicleOwnership* and *VehicleTechStatus* were found to be the least important to determine fatality, while *weatherCondition*, *DriverSex*, *VehicleTechStatus*, *RoadSeparation* and *RoadOrientation* were the least important factors for average of all classes. In the process of the experiment the total trees grown were 200 and the optimal number of tree was found to be 59 with classification error of 0.204 and cross entropy of 0.399. The TreeNet result in terms of entropy and classification error is presented in Figure 14 and 15.

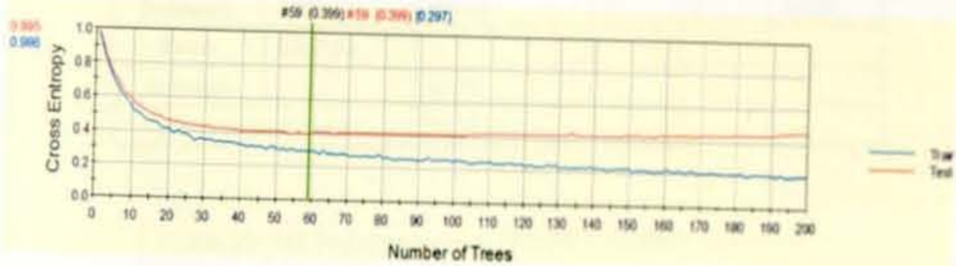


Figure 14: TreeNet Cross Entropy (2nd phase experiment)

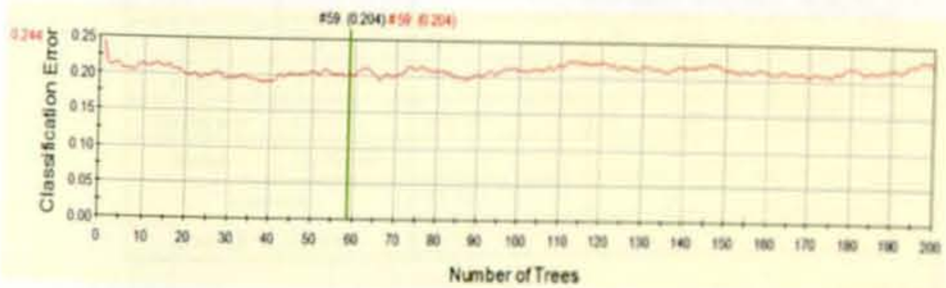


Figure 15: TreeNet Classification error (2nd phase experiment)

Entropy is a measure of dispersion in a matrix of information. Cross Entropy is a version of entropy that incorporates the modeled nature of the information content. The process of building a good model, hence can be seen as initializing a model with a random parameters followed by measuring the cross entropy and then a successive adjustment and measurement of the mode until the cross entropy is low enough.

Accordingly, the TreeNet model exhibits cross entropy of 0.399 at 59th tree, which is referred as optimal number of tree. Similarly, a classification error is a percentage of wrongly classified instances from a total number of predictions, in which is TreeNet model showed a minimum result, 0.204, which is closer to 0. As to the prediction success, TreeNet method exhibits an overall performance of 95.40% for training and 94.15% for testing sets. The detail is presented in Tables 19 and 20.

Table 19-Training Prediction Success (TREENET 2nd phase Experiment)

Actual Class	Total Class	Percent Correct	Fatal N=215	Injury N=438	NoInjury N=2200
Fatal	168	64.29	**108**	60	0
Injury	485	77.94	107	**378**	0
NoInjury	2,200	100.00	0	0	**2,200**
Total:	**2,853.00**				
Average:		80.74			
Overall % Correct:		94.15			

Table 20- Test Prediction Success (TREENET 2nd phase Experiment)

Actual Class	Total Class	Percent Correct	Fatal N=863	Injury N=1768	NoInjury N=8770
Fatal	518	82.63	**428**	90	0
Injury	2,113	79.41	435	**1,678**	0
NoInjury	8,770	100.00	0	0	**8,770**
Total:	**11,401.00**				
Average:		87.35			
Overall % Correct:		95.40			

Misclassification rate is another parameter considered to measure the performance of the model. Accordingly, the misclassification rate is presented in Table 21 for training and testing sets respectively. As it can be seen from the tables in both learning and testing scenario misclassification in case of non injury class is none.

Table 21 -Misclassification rate for test data (TreeNet 2nd phase experiment)

Class	N Cases	N Mis-Classed	Pct. Error	Cost
Fatal	168	60	35.71	60.00
Injury	485	107	22.06	107.00
No-Injury	2,200	0	0.00	0.00

When it comes to the ROC measure, TreeNet analysis method showed 0.96372 for training and 0.95097 for test scenario in case of fatal class, 0.98905 and 0.97823 for training and test sets in case of Injury class and 0.99374 and 0.99395 for training and test sets for non-injury class. ROC charts for all the three classes for test cases are presented in Figure 16.

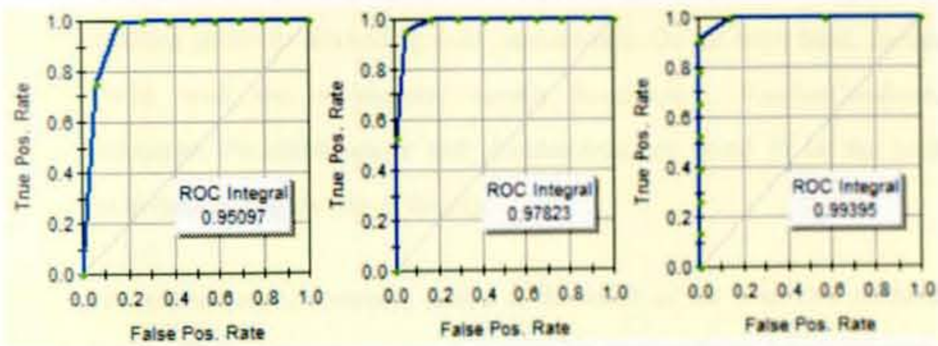


Figure 16 : TreeNet ROC charts (fatal, Injury and non-injury)

5.2.4.4 RandomForest Analysis and Results

Similar to that of the other two models with 31 attributes selected, the accident data was exposed to RandomForest analysis. The analysis was done by setting basic parameters like balanced class weight to up weight small classes to equal size of largest target class and testing out of bag data technique for testing the models. Accordingly, with 500 trees grown, the method exhibited over all error rate of 0.224, while the error rate for fatal, injury and non- injury are 0.226, 0.446 and 0.000 respectively. The performance of a predictive model error rate lies between 0 and 1. The overall error rate is presented in Figure 17.

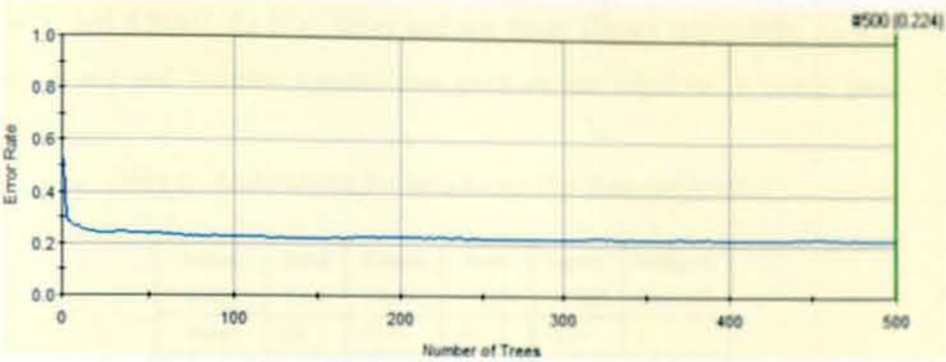


Figure 17: RandomForest Error rate (2nd phase experiment)

With respect to variable importance, *VictmOccup, VictimHealthSt, VictimCategory, VictimAge* and *Accidentcollisiontype* were the top important factors selected based on their predicting power in descending order, respectively. On the other hand, factors related with road and environment namely *Roadsurface, WeatherCondition, RoadOrientation, VechileMovement* and *AccidentArea* are found to be the least important in determining the risk of fatality.

As to miss-classification, somehow similar to the result of the other two methods result, RandomForest analysis is also less accurate in classifying injury category, while the miss classification rate is by far less for non-injury category. This is shown with 22.59%, 44.61 % and 0.04% classification error for fatal, injury and non injury classes respectively. The detail is presented in Table 22.

Table 22- Misclassification Rate (Randomforest 2nd phase experiment)

Class	N Cases	N Mis-Classed	Pct. Error	Cost
Fatal	686	155	22.59	155.00
Injury	2,598	1,159	44.61	1,159.00
NoInjury	10,970	4	0.04	4.00

Prediction success and ROC results are also important indicators of a given predictive model. Accordingly, percentages of correct prediction for fatal, injury and non-injury case are 77.41%, 55.39% and 99.96% respectively. The detail is presented in Table 23. By the same token as shown in Figure 18 the ROC integral indicates, 0.94260, 0.97671, and 0.98941 for fatal, injury and non injury classes respectively. As it is closer to one and indicates minimal zero positives and negatives, it entails good performance.

Table 23- RandomeForest Prediction Success (2nd phase experiment)

Actual Class	Total Class	Percent correct	Fatal N=1634	Injury N=1591	NoInjury N=11029
Fatal	686	77.41	531	152	3
Injury	2,598	55.39	1,099	1,439	60
NoInjury	10,970	99.96	4	0	10,966
Average		77.58			
Overall		90.75			

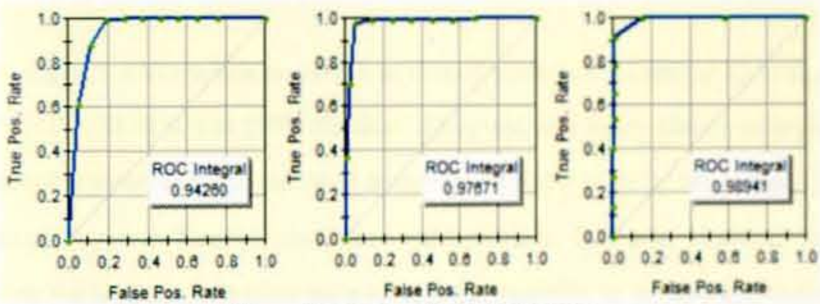


Figure 18: RandomeForest ROC charts

5.2.4.5 Scoring and Ensembling

According to (Steinberg and Golovnya, 2006), predictive modeling process would not be complete without the ability to apply the model to a data. The data could be a new data or an existing data used for training and testing the models. This process is termed as scoring. It can be done externally by translating the model into any other supported language or internally by a built in facility of the tool being used. In line with this, internal scoring is employed on the whole training and testing data to predict the target class of injury risk. The result shows that the models work well and its result is used for the final stage of combining classifiers. This is exhibited with the 95.19%, 94.55% and 95.15% over all prediction success of CART, RandomForest and TreeNet, respectively.

The final stage of the experiment is to combine models. There are different configurations and techniques to combine classifiers. As discussed by (Ranawana and Palade, 2006) cascading, parallel and Hierarchical are the major configurations. In this specific experiment a parallel combination of classifiers, where the result of each classifier is exposed to a given decision logic, voting techniques, is employed.

According to Hall et al., (2000), voting is an aggregation technique used to combine decisions of multiple classifiers. In its simplest form, which is based on plurality or majority voting, each individual classifier contributes a single vote. The aggregation prediction is decided by the majority of the votes, i.e., the class with the most votes is the final prediction (Paris et al., 2010)

Accordingly, it was possible to exhibit an overall prediction success of 95.47%, while it is 87.61%, 78.41% and 100% for fatal, injury and non injury classes respectively. The detail is presented in Table 24. It is easy to see that it is by far better than CART, RandomForest and TreeNet predictions independently. It is also important to note that TreeNet is closer than other techniques to the ensemble or combined classifiers in terms of accuracy especially in case of non injury, injury and overall accuracy. This exhibits that combining different classifiers outperformed other single classifiers for predicting injury risk.

Table 24 -Performance of Ensembling over other Models

Model comparison in percentage of prediction Success for test data				
	CART	**RandomForest**	**TreeNet**	**Ensemble**
Over all	93.52%	90.75%	94.54%	95.47%
Fatal	66.67%	77.41%	64.29%	87.61%
Injury	73.40%	55.39%	77.94%	78.41%
Non Injury	100.00%	99.96%	100.00%	100.00%

5.2.4.6 Models Comparison, Discussion and Lessons learned

As mentioned in previous sections, the experiment was done using CART, TreeNet and RandomForest predictive modeling techniques. With the intent of finding the best model, the predictive performance of these techniques are compared based on three important performance measures namely; prediction successes, prediction error rate and ROC.

One of the major objectives of this empirical research is identifying human related determinant factors for accident severity. Accordingly, the search and testing methods along with the top 10 important variables identified by the three modeling techniques are presented in Table 25. It is learned that, road user related factors are found to be more important in determining accident fatality or injury. On the other hand, factors related to road and environments are found to be the least important. This can be seen from the table as they are listed against "least important attributes" under each method employed.

Looking in to the determinant factors, in both phases of experiments, road users features like their category, occupation and age tell significant information about the possible result of a given accident collision. Thus, this research argues that the focus of attention in reducing accidents and risks should not be only drivers.

As can be seen from the experiment, for an accident to end in either fatal, injury or non injury scenario the most determinant factors are the nature of victims involved. On the other hand, time, road and environment related factors are found to be irrelevant in determining the result risk of an accident. A good example of this could be the splitter variables and variable importance results of the above experiments. It is learned that road user related factors need more investigation so as to guide proactive methods in reducing road accident and improving road safety in general.

Table 25 - Models comparison based on results

	CART	TreeNET	Randomforest
Evaluation	Prediction Success, ROC, Error rate	Prediction Success, ROC, Error rate	Prediction Success, ROC, Error rate
Search and testing Method	Entropy with 80/20 percent for training and test sets respectively	cross entropy or likelihood with 80/20 percent for training and test sets	out of bag data technique
Top ten Evaluating attributes	*VICTIMOCCUP* *VICTIMHEALTHSTS* *VICTIMAGES* *VICTIMCATEGORYS* *ACCCOLLISIONTYPES* *ACCIDENTCOUSES* *ACCSUBCITYS* *VEHICLETYPES* *HOURCATEGORYS* *ACCAREAS*	*VICTIMOCCUPS* *VICTIMCATEGORYS* *VICTIMHEALTHSTS* *ACCSUBCITYS* *VEHICLETYPES* *VICTIMAGES* *ACCCOLLISIONTYPES* *ACCIDENTCOUSES* *VEHICLEPLATES* *HOURCATEGORYS*	*VICTIMOCCUPS* *VICTIMHEALTHSTS* *VICTIMCATEGORYS* *VICTIMAGES* *ACCCOLLISIONTYPES* *VEHICLEPLATES* *DRIVINGEXPS* *DRIVINGLICENSS* *ACCUDRIVEDULEVELS* *ACCIDENTCOUSES*
Least important attributes	*PEDSTRIANMOVEMS* *VEHICLETECHSTATUSS* *ROADCONDITIONS* *WEATHERCONDITIONSS* *ROADSURFACES*	*VEHICLETECHSTATUSS* *DRIVERSEXS* *WEATHERCONDITIONSS* *ROADSURFACES* *ROADCONDITIONS*	*WEATHERCONDITIONS* *ROADORIENTATION* *ACCAREA* *VEHICLEMOVMENT* *ROADSURFACES*

With regard to the performance of the models, all the three models perform less in case of fatal and injury class, while their performance is very good in determining non injury risk of an accident. In addition, they all exhibit better ROC scores for non injury class than the others. However, TreeNet predictive modeling technique performs better by exhibiting lower error rate of 0.204, which is closer to 0, 94.54% over all prediction success and better ROC score which is closer to 1, than CART and RandomForest. The detail is presented in Table 26.

It is to be recalled from previous sections that another major purpose of this experiment is to get an overall understanding of accident data and getting sense of data quality issues. In connection with this, testing the data for identification of patterns without making much preprocessing provides a good insight in to the nature of the data. This guides the subsequent analyses and selection of better tools for this specific domain in a specific context.

Table 26 -Over all Models comparison

Criteria per Target class		Predictive Methods/Techniques		
		CART	RandomForest	TreeNet
Prediction Success (Test Set)	Fatal	66.67%	77.41%	64.29%
	Injury	73.40%	55.39%	77.94%
	Non-Injury	100.00%	99.96%	100.00%
	Overall	93.52%	90.75%	94.54%
ROC (Test set)	Fatal	0.940	0.94260	0.95097
	Injury	0.9721	0.97671	0.97823
	Non-Injury	0.9962	0.98941	0.99395
Error rate	Overall	0.300	0.224	0.204

Moreover, though, all the three techniques are found to be promising in identification of patterns in a road safety domain, TreeNet is shown to be the best method to be used in the domain under study, if the decision is to use a single method. However, ensemble result proves to be the best of all individual models.

5.2.5 Reflections on Accident Data Analysis Trends and its Implications

As discussed in Literature Review section, there were research attempts on accident data analysis in the past few years. A closer look to these selected data mining researches reveal that road safety analysis from accident occurrence and severity perspective requires due attention as variations in results are exhibited. Partly, variation is due to the techniques and data input used.

As per the objective (specific objective "a") set for this specific sub section, experiments in the current research also explains the variations, as road users factors are identified as optimal in determining accident occurrence and severity. The identified attributes are used as an input to the content dimension of the road safety information architecture. Magnitude of the data quality problems was also addressed to empirically show the extent of the problem and the need for architectural guide line in managing accident information. In addressing the second specific objective (specific objective "b") , prediction models using TreeNet decision tree technique and ensemble classifiers through voting technique are found to be effective in analysis of road traffic accident data.

The insight gained so far in the process of the research, revealed that road accident data should be complete with main emphasis on road users factors for the analysis to reflect important patterns and knowledge. There should also be a mechanism for data quality checks about each accident that requires architectural guideline. Once data is organized, machine-learning approaches to periodic analysis of the accident data is required to see changes through time and adjust the counter measures accordingly.

These findings and lessons are reflected on the information architecture proposed as a guideline for accident data collection and analysis. The design of the architectural guideline will create a platform for the analysis of road accidents in supporting efforts to improve road safety. This research perceives accident data collection and analysis as a system that requires a special view towards understanding the whole and making sense out of it for improved decision makings in the effort of reducing the problem of road safety ultimately. That is why the issue of data quality and understanding gets attention in addition to predictive modeling of some interesting patterns not addressed so far.

5.3 Analysis of Accident Information Management in Ethiopia

5.3.1 Introduction

This sub-section presents findings of qualitative data analysis into the accident information management. Accordingly, one of the research questions of this thesis, mentioned below, is addressed through the analysis of qualitative data collected.

> *RQ3. What are the problems of the current road traffic accident data collection, analysis and dissemination practice on the road safety?*

In addressing this research question we define two objectives as presented below.

- ✓ *To examine problems related to accident data reporting, data quality and analysis mechanisms in a road safety domain*
- ✓ *To identify the structure and requirement of road safety data collection and analysis focusing on road accidents*

5.3.2 Data Collection and Analysis Process

In dealing with the specific research objectives mentioned this research involved research participants from road safety domain. Hence, the study is conducted at organizations operating on a road safety data collection and analysis system to investigate practices, contents, tools and techniques in road accident data collection, analysis and dissemination. Accordingly, road safety experts, traffic polices, and traffic data analysts who are considered to be central in the overall road safety systems were interviewed. Purposive convenience sampling approach was adopted, which is the most frequently used approach in qualitative studies (Burnard, 2004, Creswell, 2009, Kumar, 2005).There appears to be no general agreement about sample size in qualitative studies as the data collection continues until reasonable saturation. It is stated in (Kumar, 2005) that; *as the main focus is to explore or describe a situation, issue, process, or phenomenon, the question of sample size is less important in qualitative research.* Literature in the qualitative research area suggest sample sizes ranging from 3 to 30 (Robson, 2002, Creswell, 1998, Marshall, 1996, Padgett, 2008).

Literature also suggests minimum sample size for qualitative studies based on the research design. The major ones includes, 3-5 participants for case studies (Creswell, 2002) ; less than or equal to 10 interviews (Creswell, 1998); 15-20 participant (Creswell, 2002) and 20-30 participant (Creswell and Clark, 2007) for Grounded Theory (Onwuegbuzie and Collins, 2007). However, as stressed in (Onwuegbuzie and Leech, 2007) sample sizes should not also be too small for it is difficult to achieve theoretical and/or data saturation

It was felt that 17 respondents should be able to supply varied and detailed accounts for the purposes of this study with a reasonable saturation. It is stated in (Marshall, 1996) that; "*quantitative researchers often fail to understand the usefulness of studying small samples. This is related to the misapprehension that generalizability is the ultimate goal of all good research... An appropriate sample size for a qualitative study is one that adequately answers the research question. For simple questions or very detailed studies, this might be in single figures*

The data collection protocol incorporates procedures and general guideline that should be followed in conducting the data collection phase of the qualitative study. The design of the data collection protocol was divided into 3 tasks; preparing interview and group discussion questions based on Zachman framework dimensions, determining the subject to be interviewed and initial scheduling of field visit. The data collection protocol was constructed to ensure consistency across multiple discussions and interviews.

The research followed the Zachman Framework as a theoretical base in a structured way for acquiring the necessary knowledge about road safety organizations practice with respect to the accident information management.

Thus, in the process of collecting data for the study and in order to answer the questions triggered by the intersecting factors (cells) from the top three rows and six columns of Zachman Framework, qualitative data-gathering techniques mentioned in the methodology section were used. Accordingly, the interview questions (see Annex B) were designed in such a way that they enable to collect relevant facts regarding the three aspects of the research; accident data collection and reporting, data handling and information sharing, data analysis and dissemination from the six dimensions of the Zachman framework namely; why, what, how, who, when and where.

Questions asked probed data content and process requirements, people, network and time of the process, difficulties faced by the road safety experts and traffic data analysts, and the supporting architecture for road safety data collection and analysis.

The interview questions were open ended to gather as detailed information as possible. The questions were categorized into general and specific. The general questions were about the magnitude of the problem in general and practices related to accident information management. In the second category, six issues related to the six dimensions of the Zachman framework were covered through a set of motivational questions. Accordingly, the first set of questions was related to the structure/content/data "*what*" requirement of the domain. The second set of questions guided the discussion about the required functions and process, "*how*", of the domain under study.

The third set of questions was about the prime motivation, "*why*", of accident data management. In the same line, the fourth set of questions focused on the geographical location "*where*" of accident data reporting, analysis and dissemination while the fifth set of questions were about the timing "*when*" of the processes and services in the domain under consideration. The last set of questions were related to the stakeholders or people "*who*" in the accident information management.

In the process of data collection, the researcher first explained the purpose of the research and got the respondents' agreement to participate in the research. All road safety experts were interviewed by the researcher, for between 30 and 45 minutes. All interview responses were short noted, with the consent of the respondents. After the interviews, the short notes were transcribed into computer files. Care was taken to consider all ethical issues.

Accordingly, respondents were interviewed on a one-to-one basis and group discussion format. A focus group discussion was conducted for three hours with six road safety experts who were drawn from the pool of participants identified for this research. The purpose of a group discussion was to support and cross-check with the data collected through interviews. Profile of the interview and focused group participants is presented as Annex D. The interview focused primarily, on issues related with road accident data reporting, analysis and dissemination. Interviews were recorded on a notebook and transcribed.

As indicated in the methodology section interview results are also supported by repeated observations. The type of observation adopted is non-participant or passive. The purpose of the observation was to best understand the central theme of the research i.e accident information management practice in line with the discussion made in (Creswell, 2009). In addition to repeated observation made on research sites, relevant documents namely strategic plan (National Road Safety Strategic Plan-2011-2020), accident data collection and analysis forms were also consulted. The documents were collected up on request from the National Road Safety Council and Traffic Office at Addis Ababa. These official documents were important to the study as they deal with accident information related issues. Samples of documents collected and used to support the results are attached as Annex F.

It is exhibited in qualitative research literature that, analysis of the data starts right from the first day of data collection. It basically, involves major steps like preparing and organizing data for analysis, reading through all data, arranging in themes along with description, interrelating theme's description and interpreting them. These activities can be categorized in to three namely; data reduction, data display and conclusion drawing. Regarding data reduction attempt has been made to transform the raw data from all data source in to a table format based on the interview questions and themes identified. A portion of it is attached as Annex E. With respect to data display narrative presentation along with supporting quotations was followed as shown in this section. Conclusion is then drawn to substantiate the design of the RSIA. Pattern-matching technique was also used to analyze the data. This is to mean that analysis of a theoretical pattern originated from the ideas or "hunches" of the researchers and the observed or operational one has been done.

Through the research process, the road safety information requirement, content and management practice was examined and investigated to see how integrated information architecture might be designed to ensure improvement. Zachman Framework for Enterprise Architecture was used as a guide to investigate the case and define the road safety information content, organization, resources, processes, and information flows.

As indicated above, in applying Zachman Framework, and to holistically control the study, a qualitative case study approach with multiple data collection techniques mentioned in the previous sections were adopted. Information obtained through these data gathering techniques helps to populate the requirements of the top three layers (18 cells) in Zachman Framework to ascertain the design details of the road safety information management context, business and system model.

The framework abstracts the characteristics and features of the road safety information based on six dimensions, Motivation, Data, People, Process, Place and Time, and explains their structures and processes from the perspectives of the planner, owner and designer of the road safety information. As per the recommendation of Yin (1994), attempt has been made to rely the analysis on all relevant evidences, to discuss all rival interpretation, to address most significant issue of the study and bring prior expert knowledge to the study.

Data analysis was carried out using Yin's pattern matching (Yin, 1994) and thematic coding methods (Creswell, 2009). The thematic coding is based on the dimensions the selected framework. Where applicable, the identified patterns or themes and relationships were compared with existent literature for validity and generalizability of the research findings, which is in line with the idea of comparing empirical patterns with theoretical patterns to expand existing theories (Yin, 1994). Findings were then used to populate the Zachman Framework at the top three layers, namely the scope, business model and the system model.

Particularly the analysis results are used to define the six dimensions of the first row in the Zachman Framework which will be then used to derive the lower level cell contents. All of the interview transcripts were read by the researcher and coded in the style of a grounded theory approach to data analysis (Creswell, 2009). Our research identified a number of accident data, process, people, motivation, place and time issues that shape or influence accident information management. Three category headings were generated and under these all of the data were accounted for.

The reporting style followed is an interpretive qualitative research reporting model where a summary of findings are presented followed by illustrative quotes of respondents and interpretations of the researcher. Findings from the data collection process are presented in three themes.

With the overall objective of the research project in mind, to define an IA that forms the basis of road safety information management system, the general observation from the data collection process revealed that there are definitely shortcomings in accident data collection, analysis and dissemination at various levels requiring an immediate attention and architectural guideline. The interview process, as well as observation made during group discussion with the various stakeholders within the road safety system, gave the researcher the necessary confirmation of the problems and information requirement of the sector. The detail is presented below in three themes preceded by a general issues regarding accident information management practice.

In all the research sites the interview processes were started with two general questions, *How do you explain the road safety situation in Ethiopia? How do you describe the effectiveness of accident data collection and analysis practice?* In responding to the first question, it was easy to learn the respondent's agreement on the magnitude of the road safety problem. They explained the situation as very problematic and top in the priority list. Apart from investigation and enacting legal actions on the accused party, lose attention and disintegrated effort were cited as major problems in the effort of road safety problem reductions. A structural organization of responsible parties like traffic offices and road safety departments is mentioned as a point of discussion. One respondent commented that: *since traffic office mainly focuses on the responsibility identification, the learning process is somehow disabled.* This fact is also observed by the researcher in the process of the research. It was also suggested that road safety departments and traffic offices be organized under same structure as their main agenda is road safety.

5.3.3 Content and Process of Accident Information Management

The first question in relation to the content and process of accident information management was regarding accident data sources for which all respondents confirmed the fact that there was no integration of data from the different sources. Traffic Police record was mentioned as the only data sources. While this is common case in most of the developing economies, it is limiting the completeness and up-to-datedness of the data ((Krishnaveni and Hemalatha, 2011); (Khan et al., 2004), (Baguley, 2001). In their study "Road Traffic Injury in China: A Review of National Data sources" (Krishnaveni and Hemalatha, 2011) mentioned that *"Police-reported data represent the main source of crash data for road safety research worldwide"*.

One respondent expressed his view that the data incompleteness is very serious problem at a national level. It is because that though there is an agreed upon national accident data format, regional states do not strictly follow it. A good instance of such a case could be an accident report containing only number of accident, accident place, time and date in a memo format in one of the research site.

It is forwarded by the interviewees that health sectors and the citizen in large should participate in improving the completeness of the accident data. The suggestions from the respondents are in agreement with literature like a study on multi-sectoral approach to capture information on road traffic injuries made in (Menon et al., 2010).

In responding to another question, general concerns regarding the integrity completeness and usefulness of road accident information provided have been expressed by the interviewees. In line with various literatures in the area (Jarvis and Kamal, 2009), (Puvanachandra et al., 2012), (Nakahara and Wakai, 2001), (Tormo et al., 2009) the process of accident data reporting, which is characterized by underreporting and poor quality, for which there is no standard, is time-consuming, inefficient and frustrating to all. This is exhibited by one interviewee as *"There are many cases where road users agree not to call a traffic police after an accident, just to avoid procedural steps required in resolving the case"*. Mixing up of accident data with "criminal data" and 'miss report' are also other mentioned problems that is said to have their own impact on both the quality and analysis of the data.

One other question in the interview process allowed respondents to comment on the data content about an accident. A number of these responses outlined 'things" or data items that the road safety data system should contain as *nature of an accident, road users, vehicles, road and environment*. This is in a general agreement with a nationally developed accident data format and various literatures in the road safety domain (Baguley, 2001);(Khan et al., 2004); (Yannis et al., 2009)). In (Baguley, 2001) these concepts were described in a form of six basic questions as where, when, why/how accident occurs; who was involved, what is the environmental conditions and the result of a collision, which an accident data system should respond. However, it was interesting to learn from the discussion that pre-accident scenarios like '*slight misses*', '*evasive actions*' and '*congestion*' before an accident were mentioned as relevant data elements to be included. This echoes the findings of (Saunier et al., 2011) who found that recoding 'slight misses" will improve the learning regarding accidents.

'Road user's reaction' after an accident, *'use of seatbelt'*, *'use of stimulant'* like "khat" and *'photograph of drivers'* were also data element indicated in the discussion as relevant to be recorded in the effort of creating a complete and comprehensive learning bed. One respondent suggested that *"whenever possible including photograph of the driver and/or involved road users in accident data will improve the completeness of data"*. Another respondent noted that taking photographs of drivers will mainly help in the legal process. Actually a complete agreement is not expected in defining accident data content as noted by (Baguley, 2001)*"it is unlikely that a completely unified report form would ever be accepted internationally, and indeed it is even difficult to draw up a definitive list of factors that will be required in all cases"*.

Regarding the process of accident data collection and analysis respondents were asked to describe the accident reporting and analysis process. Accordingly, it is learned that the collection of road accident data in Ethiopia is the responsibility of the Traffic Police. In all the three selected regions of Ethiopia, accident data collection and reporting is done at Woreda level. The process starts immediately as a police officer arrives, either by calling or if there is any around. However, data might not be collected if both parties participating in an accident agree on not to call a police officer. Though there are attempts to develop a standard accident recording format at a national level currently accident data collection is in a long hand written format which will be recorded on a daily recording file back at office. With the exception in Addis Ababa, where at least accidents resulting in fatal injury are recorded in an excel format file, in the other two regions all the data recoding is done manually on paper. Despite the relevance of different information at the accident scene, not all the data and information is recorded, resulting in incomplete and also in accurate data.

Once accident data is collected for a couple of years, all accident information, were given to data entry operators back at office to enter onto an excel and/or relational data base using Ms-Access database management software, at least in Addis Ababa case. However, by the time of this research it can be said that almost all the data handling and storage is being done on a paper based by recording on a daily accident record file, with the exception of fatal accident in Addis Ababa case which is being recorded in flat file database format.

Though, it is apparent that there are many problems emanated from manual data handling, the major problem is that data and information is not updated in a structured manner when it needs to be. In responding to a question about accident data updating, respondents mentioned the updating of a causality severity if it changes as an example. Whilst the definition for a fatality is one which occurs within one month of an accident (Puvanachandra et al., 2012), there is no proper mechanism to update this data item. This is another indication of the possible underreporting scenarios.

A question about the level and type of accident analysis was another important issue raised. Given the inefficacy of data handling system and unavailability of complete quality data, rigorous analysis and reporting is impossible. Basically, the analysis task is focused on descriptive analysis, which enables, usually, to respond to the "how many" question. One respondent expressed that; *"minimal statistical analysis limited to percentage and comparisons were the usual tasks in accident analysis"*. This is also supported by the onsite observation made by the principal researcher on the selected research sites and descriptive statistical forms being used at the traffic office. It is easy to learn that very high level, summarized analysis of accident data prohibits from making maximum use of the knowledge hidden. It is also learned that severity analysis, identification of determinant factors, dangerous location identification and predictive analysis were mentioned as analysis requirement of the domain. These entails the requirement for multidimensional new way of accident data analysis that provides maximum use of the knowledge accumulated.

However, it is worth mentioning the research and development effort made by the Road Safety Directorates, especially in Addis Ababa, and a National Road Safety Coordinating office in studying the magnitude and prevalence of the road accident to take counter measures.

A graphical representation of the current architecture for data collection and analysis is presented here in Figure 19.

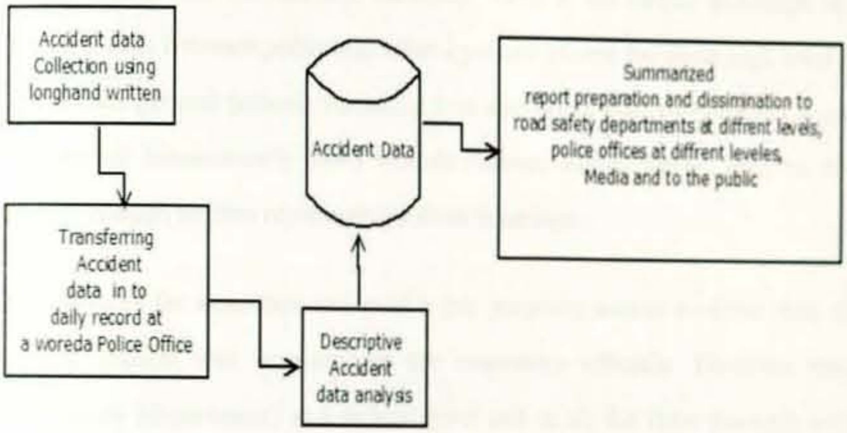


Figure 19: Current architecture of accident data collection and analysis system

5.3.4 People and Motivation in Accident Information Management

Questions regarding people and motivations in accident data collection and analysis include; *What is the prime motivation of managing accident data/information? What sort of information is required by stakeholders regarding road accident? Who do you think is responsible in accident data reporting? Who do you think should participate in accident data management and analysis? Who are the primary users of road accident information?*

In response to the above questions, interviewees clearly stated that, the need to develop an evidence based national road safety measures/plan is the prim motive for the management of accident data. A respondent from road safety department stressed that this will happen by coordinating concerned organizations as mentioned in the national road safety strategic plan (NRSC, 2011). Another respondent forwarded another important strategic motivation as mentioned in the national plan "*Promote awareness and participation of the public for sustained and drastic improvements of road safety*". But, it is noted from the discussion that the fact that the data is incomplete and high level analysis prohibited the effort to achieve this objective.

It has been identified that there are various stakeholders, who directly or indirectly need the accident information. However, there is no formal exchange of detailed accident data between police and other agencies except for some high level reporting to produce general periodic statistics. It is also noted that, summarized accident data is reported hierarchically from woreda (lowest administrative unit) to the federal police through written reports and official meetings.

Researchers for academic and public use purposes access accident data through a formal request and approval of the respective officials. However road safety directorate (department) at a federal level and in all the three research sites do not have a direct access to the detail accident data which in the researcher's view affects the effort of setting appropriate measures based on the actual accident scenario. The researcher observed a reasonable argument regarding responsibility of traffic accident information management between traffic polices and road safety departments (units).

Regarding the responsibility for reporting accidents (accident data collection), currently, it is solely the task of traffic police. In relation to this, some respondents stressed the need for capacity building of traffic polices to improve their awareness and ability on accident data collection. However, it is identified that citizens nearby can also participate in taking record of accidents that happened in the absence of traffic police. In support of this idea, an expert in a road safety department mentioned that "kebele road safety committee" should be considered, as they can educate their community member to record and/or report accident data. Of course, most of the respondents are in favour of this idea, as it is not practically possible to assign traffic polices everywhere on the roads. However, this approach requires a well planned and organized awareness creation for the public. Obviously the use of camera technology also helps in improving data collection, though it requires again huge investment. In relation to this, one interviewee suggested that citizen's participation is very much helpful especially in the case of recording accident with property damage and 'slight misses'. This agrees with the assertion made in (Zein and Navin, 2003). According to the authors, the simplistic representation of traffic safety, which is a result of traffic police report, attributes more than 90 % of all road traffic accidents to drivers. This, in turn, will lead to incorrect conclusion that improving driver behaviour is the only effective road safety strategy (Larsson et al., 2010).

5.3.5 Time and Network of Accident Information Management

Geographical location and network along with the timing issue is another aspect or theme identified. To investigate the issue the following questions are forwarded; *Can you mention a specific place (like organizational units) where data reporting, analysis and dissemination should happen? Would you comment please on the timing for accident data reporting, analysis and dissemination?* A number of respondents mentioned that these issues are critical and can significantly affect the learning process.

In responding to the first question, there was considerable confusion over who should take on the primary responsibility of accident data handling. As has been in many countries traffic police offices at different administrative levels are currently responsible in accident data collection and analysis. However, experts from road safety departments under transport offices were complaining about proper and detail data access. This claim sounds logical as road safety departments are mainly responsible in policy and countermeasures which is difficult without proper and multi-type analysis.

In connection to this, the research at hand argues that though there might be other administrative or structural solutions, which are out of the scope of this research, this problem could be solved by designing an integrated architecture. The architecture will allow keeping one master data source at the traffic office and enable other relevant offices, the road safety departments, to get access to the raw data so that they can make their own analysis and learning that guides the policy and countermeasures. The type of analysis at this site is expected to be a detail one and to make use of the predictive and exploratory machine learning capability.

Regarding the timing, it was found that accidents should be reported as soon as possible so that relevant information that will affect the learning will not disappear and appropriate health surveillance happens. Periodical update of accident data from hospital records was one of the recommendations by respondents. It was also noted that daily, weekly, monthly and yearly descriptive analysis with on-demand exploratory and predictive analysis are fundamental and the dissemination will follow through various channels like media.

5.4 Summary of the Chapter

This chapter presented the foundation to the road safety information architecture defined in this research. It started with a general overview which is followed by the results of experimental machine learning approach in addressing a portion of the objectives set for the whole research. Drawing from previous data mining researches in analysing road safety situations, the current research is focused on determining role of road users and collision related factors on accident occurrence and severity. This is then used to populate the content dimension of the proposed information architecture Exploration of data quality issues and accident analysis trends also inform the function dimension of the architecture. Moreover experimenting on ensemble and model comparisons are also part of the task.

A qualitative analysis of accident information management practice in Ethiopia was also presented in the third sub-section of this chapter. It's focus was on the problems and information requirements of the accident information management. By presenting this result, in addition to providing a ground for the definition of an information architecture presented in the next chapter, a couple of objectives set for this research were also addressed.

Chapter Six

The Road Safety Information Architecture

6.1 Overview

Drawing from the learning gained from literature review and combining this with the insights gained from the analysis of a qualitative empirical data and machine learning experiments, this Chapter proposes integrated information architecture for road safety domain. More specifically the optimal attributes identified through experiments, insights regarding data quality problems and results of qualitative data analysis are used in developing architectural description under the six dimensions. Attempt has been made to relate and refer to the extant literature in the proposed approach. Accordingly, one of the research questions of this thesis, mentioned below, is addressed as presented in (Beshah et al., 2012a) and (Beshah et al., 2011a) using theories and empirical results from previous chapters.

> *RQ4- How should road traffic accident data reporting, analysis and dissemination tasks be organized under a common architecture?*

In addressing this research question, we developed two specific objectives as enumerated below to guide the process.

- ✓ *To investigate a way to define information architecture based on enterprise architecture framework to establish road safety information management.*
- ✓ *Define a road safety information architecture that will facilitate effective utilization and management of road traffic accident information among road safety organizations*

Assessment of the existing scenario and practices in accident information management and a review of an international practices and technologies in the area provide a ground to the formulation of information architecture guided by known Enterprise architecture framework. Moreover, the empirical result of knowledge discovery experiments conducted on the existing data from one of the research sites, Addis Ababa, also revealed some important issues including analysis requirement and data quality that needs to be addressed. The experiments also identified effective predictive models that integrated to the accident data system as per the architecture. Accordingly, description of the defined general architecture followed by the proposed improvements is presented below.

6.2 The Proposed Road Safety Information Architecture

A review of international and existing practices in a local context and result of successive data mining experiments made so far ((Beshah et al., 2011b), (Beshah et al., 2012b), (Beshah et al., 2012a), (Beshah and Hill, 2010)) exhibited a need for an integrated approach to the road accident data collection and analysis which will be addressed by an enterprise view. Therefore, in order to define an information architecture model, two dimensions of human and operative perspectives are mentioned based on the Zachman framework (Zachman, 1999). The first dimension embodies the viewpoint of the people, who are the actors of road safety data collection and analysis process. The second one, operative perspectives, defines the several requirements, constraints and operations that should be created or mentioned to architect the road safety data collection and analysis. In the process of illustrating the proposed architecture, the essence of these dimensions is described. The container of each cell that is deduced from the crossover of columns and rows is also defined, focusing on the anchor cells as described in (Pereira and Sousa, 2004).

Since Zachman framework is more general, including different views and dimensions at different levels, it is relevant to delimit its use as per the research scope under consideration. It is essential to get a clear definition (at meta-level) of the contents of every cell. Thus, the first step should be defining the focus. If a combination of the columns and the rows is considered, it should be explicitly stated, which rows and columns are being addressed and are useful. Moreover, the focus of the framework should be made clear and understandable (Goethals-SAP-leerstoel, 2004). The Zachman framework is simply a framework. It is not a process, a method, a notation or a tool (Bahill et al., 2006). It is also mentioned in (Kilpeläinen and Nurminen, 2007) that Zachman Framework requires tailoring.

Accordingly, computation (row 1 and 2) and platform independent (row 3) representations are selected because of the scope of the study and to keep the flexibility of the implementation of the architecture. Similarly, based on their relevance to the domain and scope of the study, the six dimensions or columns of the Zachman Framework are emphasized (Bahill et al., 2006, Nikolaidou and Alexopoulou, 2008).

In (Mahmood, 2006) four dimensions namely, the *what, why, who* and *how* are suggested as the minimum set of descriptions for each representation, though, it may be necessary to add more descriptions depending on the business domain under study. Thus, reconsidering the idea given in (Zachman, 1999) for the case of road safety data collection and analysis, a road safety information management could be represented in a 3 by 6 matrix containing the top three view points (rows) and six dimensions (columns) as partially presented in (Beshah et al., 2011a).

The rows of this architectural model describe different viewpoints to approach the subject at different layers. They are described in Table 27. Similarly, the columns of this architectural model describe several dimensions of the model, which are referred to as abstractions. They are explained in Table 28.

Table 27- Rows of the Road safety Information Architectural Model

VIEW POINTS	Description
SCOPE (CONTEXT)	Describes the strategies, content and constraints of the road safety data collection and analysis.
BUSINESS (CONCEPT)	Define goals, structure and processes that are used to support mission of road safety organizations with respect to accident information as an enterprise.
SYSTEM (LOGICAL)	Contains system requirements, objects, activities and functions, network that implements the business model. The system model states how the system is to perform its functions.

Table 28-Dimensions of the Road Safety Information Architectural Model

Dimensions	Descriptions
WHY (MOTIVATION)	It translates road safety strategies and objectives into specific meaning.
WHO (ENTITIES)	It defines who is related to road safety data and information management
WHAT (DATA)	It is data to define and understand road safety requirements.
HOW (PROCESSES)	Processes to translate road safety requirements into more detailed implementation and operation definitions.
WHERE (PLACE)	It is related to physical distribution of places where accident data collection and analysis will be implemented and operated.
WHEN (TIME)	It describes how time influences accident information management.

In line with the above discussions, a 3X6 grid representation of the architecture is presented in two tables (Table 29 and 30), by identifying relevant dimensions and user views for road safety data collection and analysis practice. Accordingly, a higher level Road Safety Information Architecture (RSIA) is offered based on the Zachman's architectural framework.

As indicated on the architecture (Table 29 and 30), the first row or contextual, defines higher level view of the data, people, function, time, place and motivation of road safety data collection and analysis. This creates an umbrella under which the conceptual and logical design of the road safety data collection and analysis is to be crafted. Returning to the columns, the "Why" enables to clearly state the motivations at the three levels (rows), the 'what" will describe, what to collect, analyze and disseminate, the "how" help in defining how to collect, analysis and disseminate, the "When" will guide the timing, the "where" illustrate the geographical distribution of activities and lastly the "Who" portray who should collect, analyze, use and disseminate. The next level of architecture produces various drawings, checklists, diagrams appropriate in representing the above-mentioned content with in their context for their respective users.

Table 29-Road Safety Information Architecture (RSIA) based on Zachman Approach

	Why (Motivation)	**Who (People)**	**What (Content)**
Scope (Contextual)	Social, health and economical impact case in a form of goals	Essential Road safety organizations and their role.	List of important road safety (accident) information.
Enterprise and Environment (Conceptual)	Citizen's benefit and road safety data collection and analysis objectives.	Road safety information system workflow	Semantic description of road safety information (accident).
Road Safety System (Logical Design)	Functional requirements of Road safety data collection and analysis system	Road safety information System human-system interface design.	Logical data model for road safety information

Table 30-(Continued) Road Safety Information Architecture (RSIA) based on Zachman Approach

	Where (Network)	**How (Function)**	**When (Time)**
Scope (Contextual)	Locations of road accident data collection and analysis system	Important road accident data collection, analysis and dissemination Process.	Road accident information management significant events
Enterprise and Environment (Conceptual)	Functional decomposition of the road accident information management process	Conceptual activity model of road safety data collection and analysis	Sequencing and Time period assigned to the significant events
Road Safety System (Logical Design)	Physical Deployment Architecture of the road safety data system	Architectural design of road safety data collection and analysis system with function and user Views.	Time representation of significant event with period stereotype

Given the high level architectural representation, the detail description follows. Accordingly, based on the empirical results, international practice and a method to define an enterprise architecture using Zachman framework proposed in (Pereira and Sousa, 2004) a description of an architectural model for road safety data collection and analysis is presented by views/perspectives.

Contextual View- Scope/Planners Perspective

This perspective is concerned with issues important to planning of the road safety data collection and analysis system. This can be seen as a means of defining the scope of the envisaged system. According to (Pereira and Sousa, 2004), there is no dependency among cells' concepts. So, the order of fulfilling for this row is totally free and thus, the process of defining cells can be executed in parallel. Interview results as presented in section 5.3 guided by the 6 motivational questions of Zachman framework, were used to populate cells of this row. In addition, analysis of a road safety strategic plan was also used to define the content of cells at this viewpoint.

Accordingly, the first cell (Context-Motivation) lists motivation for this system in terms of social, health and economical impact case. The content is presented in Table 31. Similarly, the next cell (Context-People) lists essential road safety entities or organizations identified, while the third cell (Context-Content) lists business subjects or entities that the system needs to keep information about. Data hierarchy can be extracted based on data analysis of enterprise strategic plan, enterprise goals and enterprise missions (Rezaei and Shams, 2008). As discussed in section 5.3.3, nature of an accident, road users, vehicles, road and environment are identified as major data elements. It is also found that two entities namely pre-accident and post-accident scenarios are vital for a complete picture of the case at hand.

The fourth cell (Context-Network) addresses the location of an enterprise (Ertaul and Sudarsanam, 2005), which includes regional and national offices along with all stakeholders involved in accessing and working on the data. This is populated from the findings of the analysis of the strategic plan and qualitative interview results as presented in section 5.3.5. The representation is just a list of locations, which the system, road safety data collection and analysis, operates (Radwan and Aarabi, 2011).

The fifth cell (Context-Functions) lists the major processes in which the business operates. The process includes major tasks right from the occurrence of an accident to the dissemination of the information. Another critical aspect of such type of system is the timing issue. Accordingly, the sixth cell (Context-time) addresses the list of significant events for accident information management (Fatolahi and Shams, 2006); (Radwan and Aarabi, 2011). As indicated above and recommended in the literature strategic documents (NRSC, 2011), empirical (section 5.3) and data mining results (section 5.2), the researcher's own observation and lessons from international best practices are analyzed in the process of populating the first row of the model as presented in Table 31 and 32. Motivations, participating entities and accident data content are presented in Table 31 while the place, process and timing aspect of accident data is exhibited in Table 32.

Table 31-Contextual View- Scope/Planners Perspective

	Motivation	People	Content
Scope (Contextual) **Road Safety**	• Develop evidence based national road safety measures/plan by coordinating concerned organizations • Promote awareness and participation of the public for sustained and drastic improvements of road safety	• National Road safety Coordination Office • Federal Police • Federal Road safety Directorate • Regional Road Safety Department • Regional Traffic office • Woreda Traffic office • Health care Institutions • Citizens	• Accident (general) • Road Users (drivers, victims, pedestrians) • Vehicles • Road • Environment • Pre accident scenarios (trajectories) • After accident situations
	Why	**Who**	**What**

Table 32-(continued): Contextual View- Scope/Planners Perspective

	Network	Function	Time
Scope (Contextual) Road Safety	• Accident sites • Traffic offices (Woreda, Regional , Federal) • Road safety departments (Woreda, Regional, Federal) • National Road safety Coordinating Office	• Accident data collection and reporting (Accident location referencing, Accident reporting, Quality and completeness checking, Updating accident data) • Regular in-depth analyses of patterns of accidents (Descriptive analysis, Exploratory analysis, Predictive analysis, Trend analysis) • Information exchange • Analysis result Interpretation and Dissemination	• Pre-accident events • Accident investigation • Accident recording • Data Quality check • Accident analysis • Accident Information Dissemination
	Where	How	When

Conceptual View-Enterprise/Environment Perspective

The second important users view considered is conceptual view which is from the enterprise and environment aspect. Based on the recommendation made in (Rezaei and Shams, 2008) cell contents at this level are derived from the above contextual perspectives. This perspective models the motivation, people, content, network, time and functions of a road safety data collection and analysis system from the viewpoint of business owners. Accordingly, in the first cell (Conceptual-motivation) the major reason for this specific system from enterprise point of view is presented in terms of objectives. The second cell (Conceptual-People) as shown in Table 33 allows exhibiting the general work flow in between and within road safety organizations. As explained in (Pereira and Sousa, 2004), Organization chart or Processes vs. Organization Matrix can be used to model the content of this cell.

Table 33-Conceptual-People using Process Vs Organization matrix

	ADC	QCU	DA	PA	TA	ISD
Woreda Traffic Officers	X	X	X		X	X
Citizens	X					X
Regional Traffic Offices			X		X	X
Federal Police			X		X	X
Regional Road safety Departments				X	X	X
Road Safety Directorate Office				X	X	X
National Road safety Coordinating Office						X
Health institutions		X				X

Key: ADC: Accident data collection (reporting and referencing), QCU: Quality Checking and Update, DA: Descriptive analysis (interpretation), PA: Predictive & exploratory analysis (interpretation), TA: Trend Analysis, ISD: Information Sharing and dissemination

The third Cell (Conceptual-Content) presents semantic description of the road safety data, the business content, including business entities, their attributes and relationship. Entity dictionary can be used to represent this cell (Radwan and Aarabi, 2011). In this perspective, important data entities and their relations can be extracted based on scope/planner's perspective outputs (Rezaei and Shams, 2008).

The fourth cell (Conceptual-Network) defines the location of business nodes, where the system is used (Ertaul and Sudarsanam, 2005). It is also mentioned in (Radwan and Aarabi, 2011) that, the focus of this cell is to represent the conceptual model of "Where", which includes the location of and place, where stakeholders, use from the system and also the operations that they can do related to this. As described in (Radwan and Aarabi, 2011, Fatolahi and Shams, 2006), organizational units within location stereotypes of UML packages associated using dependency relationship is a preferred modelling to represent this cell. The content of this cell is represented accordingly as shown in Figure 20. As can be seen from the figure, four important road accident information management locations are represented with their dependency relationship. It also agrees with the proposed approach in (Pereira and Sousa, 2004), that considers the content of conceptual-function and contextual-network cells to populate this cell.

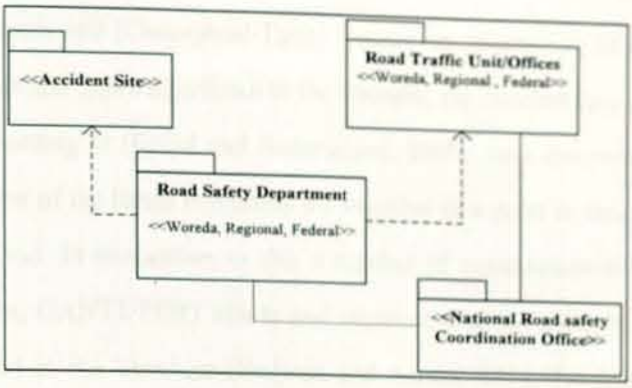


Figure 20 : UML package representing conceptual-network cell

The fifth cell within this perspective, (Conceptual- Function) models the business workflow of the stakeholders interacting with the business (see Figure 21). Flowchart, activity diagram, UML use case diagram and sequence diagrams are common tools for process modelling at this layer (Radwan and Aarabi, 2011). Accordingly, activity diagram is used to represent the content of this cell. It clearly explicates the work flow in between the locations identified earlier. Its content is derived from the cell in the above row, contextual-function.

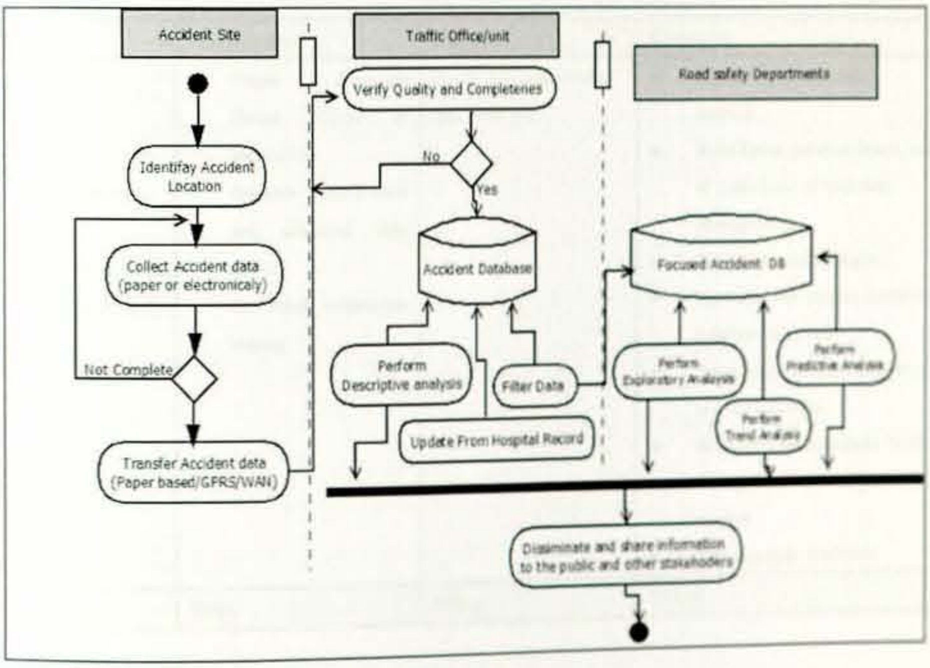


Figure 21: Activity Diagram representing conceptual-function cell

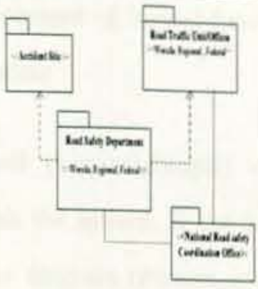
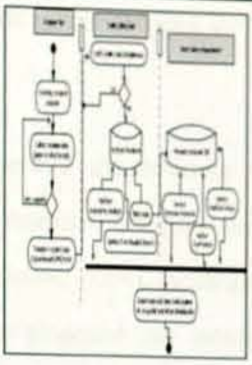
Finally, the sixth cell (Conceptual-Time) focuses on sequencing of the timing of process, events and flows significant to the business, the accident data collection and analysis. According to (Ertaul and Sudarsanam, 2005), time dimension may be of two forms. One of the forms represents the snapshot of a point in time and the other defines a period. In connection to this a number of representations like Business execution plan, GANTT/PERT charts and sequence diagram using business actors, were proposed in the literature (Radwan and Aarabi, 2011, Fatolahi and Shams, 2006). Hence, a list of events with their suggested period is a preferred approach for its understandablity to represent this cell.

Based on the above discussion, the populated architectural description of the second row of the road safety information management system is presented in Table 34 and 35. The conceptual level representations of motivation, people and content are shown in Table 34 while Table 35 contains models and description for the network, function and time dimensions.

Table 34-Conceptual View-Enterprise/Environment perspective

	Motivation	**People**	**Content**
Enterprise and Environment (Conceptual)	• Single Complete Master Source of Information. • Dynamic , multi-view and advanced data analysis • Facilitated information sharing	People vs Process matrix (See Table 33)	• **Accident**; severity, type, location... • **Road Users**; personal details, use of seatbelt, use of stimulant, photograph..... • **Road**; type, nature , name ..... • **Environment**; weather condition, light condition, time... • **Pre accident scenarios**: evasive action, slight miss, • **After accident situations**; health status of victims, road users reaction • **Vehicles**; type, conditions...
	Why	**Who**	**What**

Table 35-(continued): Conceptual View-Enterprise/Environment perspective

	Network	Function	Time
Enterprise and Environment (Conceptual)	 See Figure 20	 See Figure 21	• Pre-accident events • Slight misses • Accident investigation (on spot, immediately) • Accident recording (on spot, immediately) • Data Quality check (on the spot and until one month) • Accident Analysis (Descriptive analysis (day, week, month and yearly base) (Predictive and exploratory analysis, yearly)) • Accident info. Dissemination (based on analysis time)
	Where	**How**	**When**

Logical View- Road Safety Data Collection and Analysis System Perspective

The last important aspect of the accident data collection and analysis process is the systems view. This perspective models the requirements, participation, business content, and process of the accident data collection and analysis system from the viewpoint of a system. It helps to portray the specific functional requirements of the system, human-system interface issue, logical data model, geographical location, timing and layered architectural design of the data collection and analysis system with functions and users views. Architectural artefacts of this perspective are presented in Table 36 and 37.

Accordingly, the first cell (Logical-Motivation) presents the reason of the system in terms of functional requirements. Using data from the interviews and review of existing road safety initiatives, it was possible to develop functional requirement of the road safety data collection and analysis system expressed as behavioural objectives to populate the Motivation (Why) component of this perspective. In (Radwan and Aarabi, 2011), it is recommended to consider the analysis of the cells above and the content of logical-function cell in defining the content of this cell i.e Logical-Motivation

The second cell (Logical-People) describes the structure and contents of user interactions with the system. It can be modelled using Systems vs. Roles Matrix or UML Use Case diagram (Pereira and Sousa, 2004). Accordingly, Use Case model (see Figure 22) is used to represent the proposed user interaction in a road safety information management scenario. The inclusion of citizens and health institutions in the process of accident information management is worth mentioning. Its content is derived from the above cell and analysis of the function perspective of the systems view.

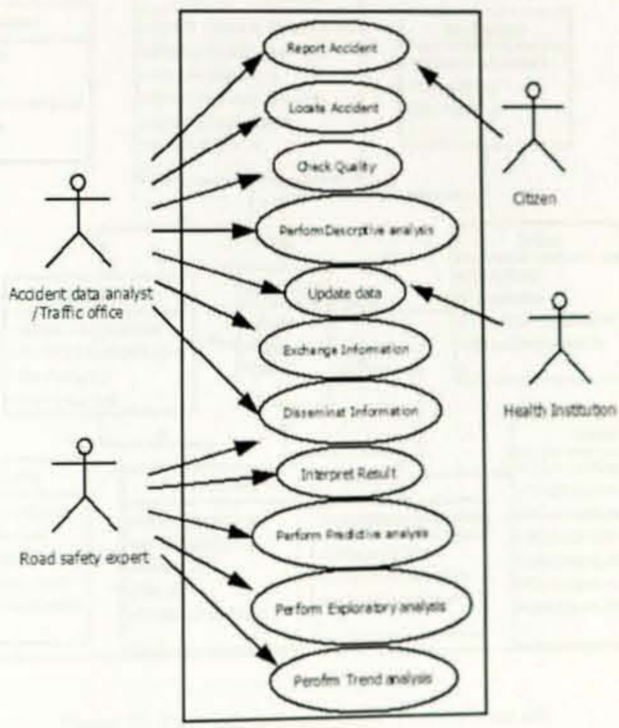


Figure 22: Use case Model representing Logical-People cell

The third cell (Logical-Content) describes the business structure including business entities and their relationship. Example model of this cell could be a business entity-relationship diagram (ERD) that models the business concepts, entities and attributes. The content of this cell could be extracted from owner's perspective output (Rezaei and Shams, 2008, Pereira and Sousa, 2004). In addition to extracting from the previous perspective, the content is also a result of empirical data collected. Thus the detail data elements are upshots of information requirements identified through data mining experiments as shown in Table 25 and qualitative data collection and analysis presented in section 5.3.3 of this dissertation. It is depicted in Figure 23. As can be seen from the figure, Road, Driver, Victim, Pedestrian, Vehicle, Environment, Accident, Pre-accident and After-accident scenarios are important data elements, that the road accident data system should contain and are modelled.

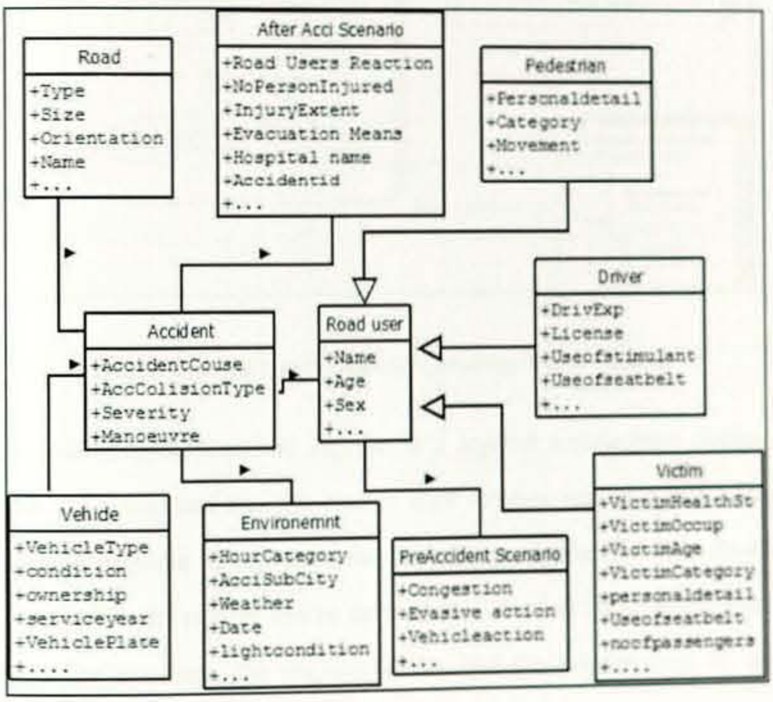


Figure 23: ER Model representing logical-Content cell

The forth cell in this perspective (Logical-Network) represents the available nodes of a whole system/enterprise and logical links in between them. In the road safety domain the accident sites, Road Safety Departments, Road Traffic Units/offices and National Road Safety Coordination Office are modelled with their respective modules are modelled. This cell has the conceptual-network and logical-function artefacts as its base (Pereira and Sousa, 2004). Though, system diagram and UML component are also proposed, Deployment diagram, as shown in Figure 24, using location stereotype of packages is a preferred modelling techniques to represent this cell (Fatolahi and Shams, 2006).

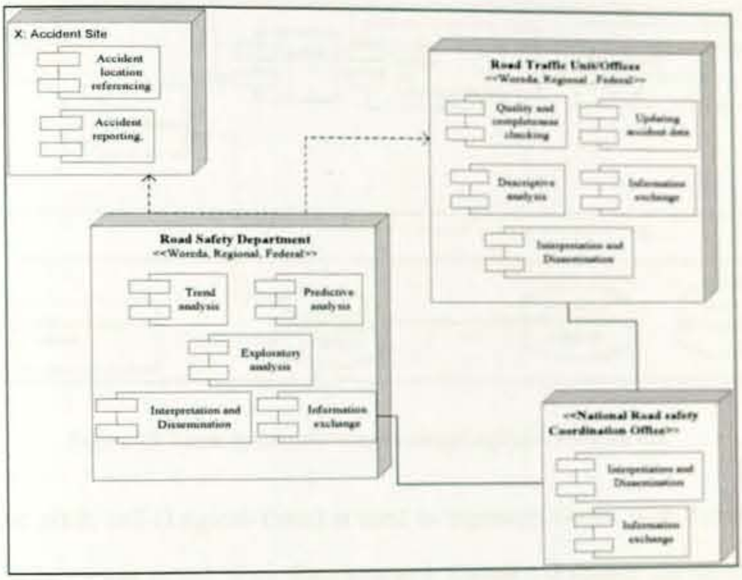


Figure 24 : Deployment diagram representing logical -network

The fifth cell (Logical-function) represents a layered architectural design of road safety data collection and analysis system with function and user views (see Figure 25). It is also possible to represent this cell using an essential data flow diagram (Källgren, 2008). Its content can be derived from the cell immediately above. Thus, it specifies the structure, the responsibilities and the relationships of the design elements of a road safety data collection and analysis system.

As can be seen from the diagram (Figure 25) the system is represented in three layers where the bottom layer represents the data storage while the upper two layers illustrate the business logic and the presentation or end users view respectively.

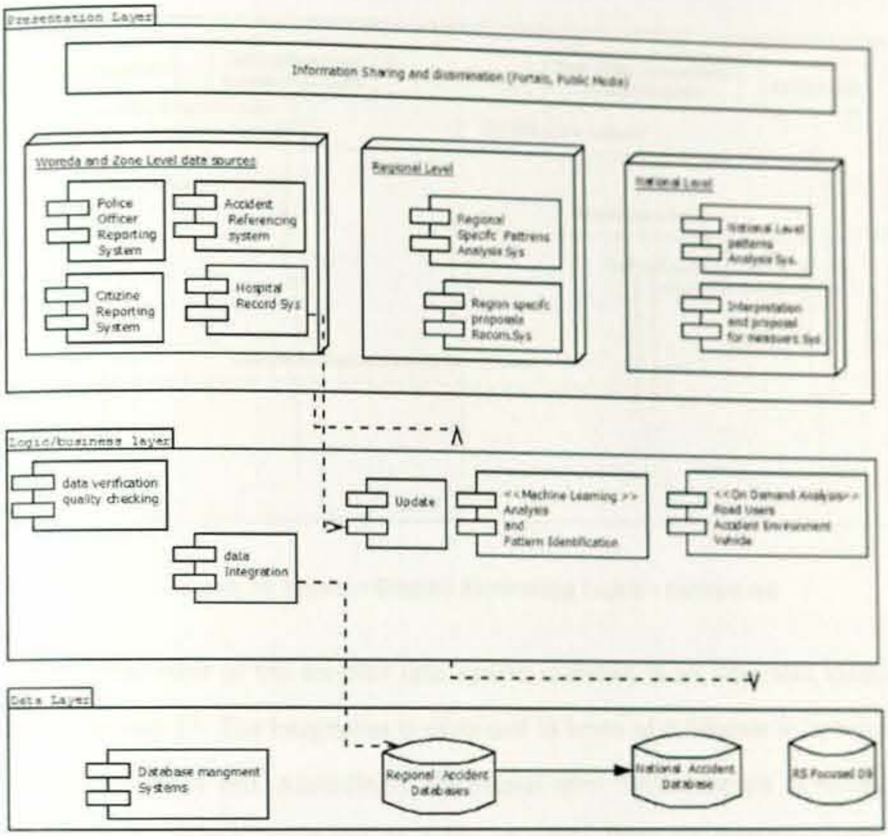


Figure 25: Layer Architecture representing Logical – Function cell

Finally, the sixth cell (Logical-Time) is used to represent events with their sequence and period in a more detail. State diagrams and sequence diagrams are recommended in the literature to define the content of this cell (Pereira and Sousa, 2004). Sequence diagram (see Figure 26) using actors and UML notes for periods is a suggested diagram to represent timing in accident information management (Radwan and Aarabi, 2011, Fatolahi and Shams, 2006). The diagram explicates the process of accident data collection and analysis practices emphasizing on time sequencing of the events.

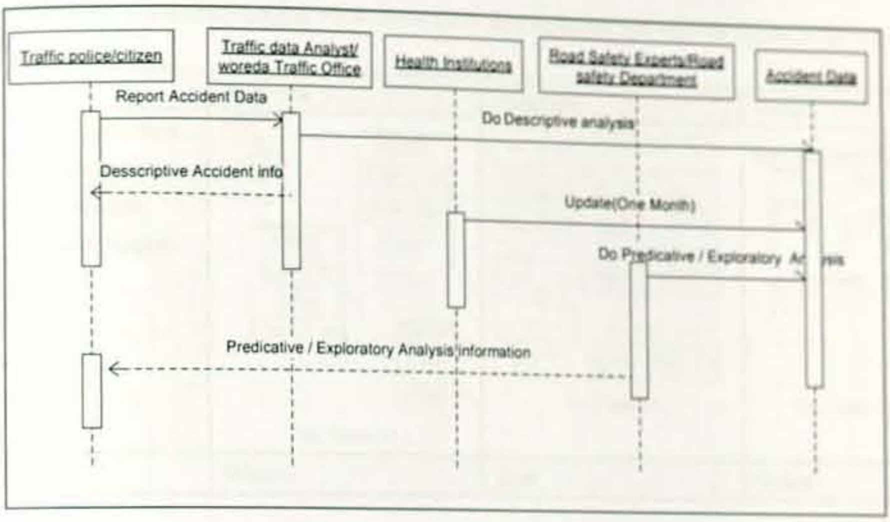


Figure 26: Sequence Diagram Representing Logical– Function cell

The logical view of the accident data system is shown in an integrated manner in Table 36 and 37. The integration is expressed in terms of coherence in between the contents of each cell. Accordingly, the logical level representations of motivation, people and content are shown in Table 36 while Table 37 contains models and description for the network, function and time dimensions.

Table 36- Logical View- Road safety data collection and analysis system perspective

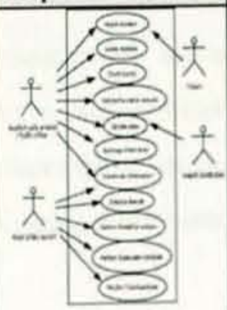
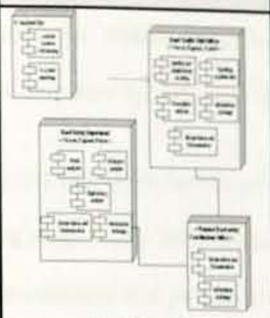
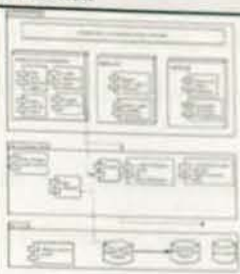
	Motivation	**People**	**Content**
Road safety Data Collection and Analysis System (Logical Design)	• Enable capturing complete, quality and accurate information • Provide a platform for information sharing • Enable periodic descriptive, exploratory and predictive analysis	 See Figure 22	 See Figure 23
	Why	**Who**	**What**

Table 37- (continued) Logical View- Road safety data collection and analysis system perspective

	Network	Function	Time
Road Safety Data Collection and Analysis System (Logical Design)	 See Figure 24	 See Figure 25	 See Figure 26
	Where	**How**	**When**

6.3 Discussion

This section discusses key attributes of the proposed Road Safety Information Architecture (RSIA). The architecture serves as a mechanism to facilitate communication regarding accident data by which it satisfies one of the major purposes of architecture.

Based on the generic Zachman Framework, the Road Safety Information Architecture is represented by the integration of rows and columns. The framework reduces complexity by decomposing enterprise architectures into level of cells that are allocated by column and row. While a column represents a specific dimension each column at each row point has a unique model. Similarly each row at each column point presents a unique perspective. Representation of the architectural descriptions is based on the optimal attributes (more specifically for the data and process dimensions) identified in the data mining experiments and a qualitative data collection and analysis process.

What is to be noted here is that each perspective could be applied to both development of new systems and maintenance of existing systems. Effective enterprise information architecture is more than just populating Zachman cells or modeling information, technology and application aspects independently. In line with the argument forwarded in (Abdullah and Zainab, 2008), without an integrated perspective, all architectural models and strategies will be misaligned, inflexible and will remain unused. As stated in (Abdullah and Zainab, 2008), such type of architectural guidelines portray the "real life" situation and thus considered robust.

Thus, by applying the top three levels of the Zachman hierarchy, it was possible to develop descriptive road safety information architecture. Major procedures includes; determining the dimensions and perspectives required and applicable for the purpose under study, determining the information requirements of the domain through qualitative data collection and data mining experiments, classifying and structuring the information based on the dimensions identified keeping the integrity, representing the information in a required level of detail for the perspective under consideration, evaluating both the process and the content using recommended metrics and methods. Therefore, though key purpose of the research is to design the road safety information architecture of road safety organizations based on a framework, the process and techniques employed as a methodology is another major contribution of this work.

Analyzing the framework's perspectives, (Pereira and Sousa, 2004) verified the existence of a concept related to the Zachman Framework that the authors designate as an "anchor cell". According to the authors, an anchor cell is a cell that on any framework's perspective has an aggregate function relative to the other cells. In line with this, it is easy to learn that two cells representing the intersection of the conceptual and logical perspectives under the function or process dimension are more of aggregate and core, which entails they deserve more discussions and attention. In the context of this specific research, the description of content, process and people at a conceptual and logical level satisfy the 'anchor cell' concept. This is mainly because these cells represent the major issues in an accident information management practice under study. The concept of 'anchor cell' agrees also with the (Barker, 2005) Information Architecture core representations; content, context and users.

Road safety information architecture (RSIA) is defined specifically for low resource settings. This is to mean that the architecture is developed based on empirical data where infrastructure related to road and traffic management is limited which will influence mainly the content and process aspects of accident information management. Countries will use the general architecture to create individual country-specific architectures. In addition to empirical data, RSIA is a result of IA concepts and approach guided by enterprise view using Zachman Framework as applied to road safety data collection and analysis. Although the basic principles of road safety in low and high resource settings are similar, there are many practical differences that drive the need for a dedicated architectural guidance for low resource settings, including dedicated views and viewpoints and unique artefacts. We believe, the lessons learned will, in turn, inform designs and practices in the developed world.

A key distinguishing feature of RSIA is simplicity. Partly, the simplicity emanates from the nature of the guiding framework selected. RSIA attempts to distil out the best procedures, practices and artefacts that have worked or become ingrained in several countries, as learned from literature supported by empirical investigations in the Ethiopian case and generalize these into a general architecture that can be applied in other countries. The intended result is a user-oriented practical architectural model that balances ease of adoption and use with completeness and theoretical rigor. The overarching purpose of the RSIA is to assist in evaluating and/or developing road safety data collection and analysis systems and promoting informed decision-making by road safety organizations.

6.4 Summary of the Chapter

Road Safety Information Architecture, which is the major deliverable of this research work, has been presented in detail in this chapter of the thesis. The discussion includes both the process and the end result. It starts with an overview explaining the link that this chapter has with the previous sections. Then the architecture is presented guided by an enterprise architectural framework. The discussion followed providing details of its applicability and key features of the architecture.

Thus, the development of the road safety information architecture in an enterprise view is an original contribution, which improves and expands the conceptual framework of the research in the field. It is believed that communication and understanding in a road safety information management will be improved.

Chapter Seven

EVALUATION OF THE RESULTS

7.1 Overview

There are various factors that affect the quality of a research. This part of the thesis explicates issues that could have caused a weakened or strengthened quality of the research at hand. Accordingly, a multi-method evaluation technique as described in the methodology chapter was employed to evaluate and validate the work. Generally, it is found that both the process and result of this research are valid and reliable. That means given the stated criteria, the research result adequately describes the phenomena of interest at hand.

As this research follows a design science approach, it also adheres to its evaluation guideline forwarded in literature. Regarding evaluation, Hevner, et al (2004), stated that the utility, quality and efficacy of design artefacts must be rigorously established through well executed evaluation methods. It is also stated that the selection of metrics (like completeness, consistency, ease of use, reliability, usability, fit with the organization and purpose, and other relevant attributes) and evaluation methods (Observational Experimental, Analytical, Descriptive, Testing, Surveying) is crucial (Hevner et. al. 2004; Cleven et al., 2009). Hence, in line with the discussion in design science literature, the evaluation process in this research agrees with the metrics and methods recommended.

Thus, issues of validity, reliability and generalizability of findings, accuracy and interestingness of machine learning experiments, completeness, practical utility, and robustness of the architectural models are examined and explained. In discussing the evaluation process of this research we follow the assertion made in (Cleven et al., 2009) as *"we emphasize that evaluation should not be seen as an isolated process activity but needs to be considered from the beginning of a design process."*

7.2 Validity and Reliability of the Process and Results

Validity and reliability do not carry the same connotation in qualitative researches as they do in qualitative researches (Creswell, 2009). In qualitative researches, validity means that the researcher checks for accuracy of findings by employing certain procedures, while reliability refers to the consistency of the approach used (Gibbs, 2008, Creswell, 2009). The discussion here in this section is basically the last follow-up of the validity and reliability issues that was presented in the methodology chapter. Accordingly, the theoretical support along with the quality of the empirics and data collection and analysis is described in this section.

The Theoretical Support in the Research (Descriptive Evaluation)

An important way of validating a research work is exhibiting the theoretical support. According to design science literature (Xie and Helfert, 2011, Hevner et al., 2004), a descriptive evaluation method involves informed argument using information from relevant researches to build a convincing argument for the process of building artifacts and their utility. Accordingly, attempt has been made to refer relevant literature in the area throughout the whole research process. In line with this, the literature and articles that have been used in this dissertation have been judged by their credibility. By doing this, we believe to improve the validity and reliability of both the process and the end result. Major consideration in assessing the credibility include, slight consideration of publisher, the age of the source and the source matching to the topic. It is obvious that the age is crucial to check, as a result of the dynamic nature and developments of the topic under study. Moreover, referring has frequently been used, to validate statements. As much as possible, multiple sources are used in the process of validating statements and assertions.

Regarding the framework employed, Zachman Framework, as discussed in (Radwan and Aarabi, 2011, Dominguez, n.d) that it has many important features like its capability and appropriateness by being expressive, simple and a problem solving tool. It is also mentioned that it agglutinates, communicates, helps planning and it is "product" neutral, which proves its wide applicability. *"Its usefulness is due to the fact that it is based on well established organizational and practical views as well as on 5WH scheme which in turn helps describing precisely any abstract or concrete entity"* (Dominguez, n.d).

Its applicability in security planning and architecture is also demonstrated in (Ertaul and Sudarsanam, 2005, Mohajerani and Moeini, 2004, Ertaul et al., 2010) as, *"It is shown that Zachman Framework best fits to plan security architecture for an enterprise as any evolving changes in technology can be implemented onto the Zachman Framework without affecting the direction of the enterprise."*

Although, it appears that Zachman Framework has not been used yet to define road safety information architecture, it has been proved to be useful in areas related to information technology like network security, digital library content, enterprise aggregate planning. Thus, as it is shown herein, using a Zachman framework based architectural descriptions improves communication by reducing misunderstandings and hindrance in defining proper information management practices. The use of Cross- Industry Standard Process for Data Mining (CRISP-DM) also confirms the valid theoretical base for data mining experiments conducted.

The Empirics in the Research

As the research is a qualitative case study, mainly, interview was used as a major data collection technique, augmented by observation and document analysis. In line with the concept of validity in a qualitative approach, attempt has been made to present a rich and thick description to convey findings. As indicated in (Creswell, 2009), detailed descriptions of the settings and findings make a research output more realistic and richer. Where applicable data sources triangulation, where themes are described from different data sources namely; interview and group discussion, observation and document analysis, was also used. The process of supporting themes based on converging sources of data adds on the validity of a research findings (Creswell and Miller, 2000, Creswell, 2009).

Another important undertaking in the process of building up the validity of this work is member checking technique. In connection to this, efforts have been made to determine the accuracy of the findings through taking the lessons, understandings and the resulting models back to the research participants for comment and confirmation.

It is also important to illuminate the capability of the participants in understanding and commenting the research results. Accordingly, the fact that they are experts in the area and working for fairly longer time enables the participants to reasonably understand the concepts and presentations regarding the research findings.

Regarding the reliability of the work and findings, the data collection and analysis procedures described in the methodology and result section provide the necessary details to justify the consistency of the research approach. Another worth mentioning issue in a research is the generalizability of findings and results. Though, it is stated in (Yin, 2003) that qualitative case study results can be generalized to some broader theory, other literature (Creswell, 2009, Gibbs, 2008) argue on the intent of a qualitative research as not to generalize findings outside of those under study. Rather, it is stated that the value of qualitative research lies in the particular description and themes developed in the context of specific site. According to (Yin, 2003), the generalization occurs when researchers study additional cases and generalize findings to new cases. Following this argument, it is easy to understand that the results of this research can be further expanded and generalized by studying different cases in different sites as illustrated in (Onwuegbuzie and Leech, 2007).

7.3 Accuracy and Interestingness of the Experimental Results (Experiments)

With respect to the validation of the data mining experiments, experimental techniques recommended in literature were employed. Accordingly, objective like accuracy (prediction success), error rate (misclassification), ROC and subjective validation like domain experts comments on the soundness of the patterns were used. The subjective evaluation helps to exhibit the interestingness of the patterns extracted, while the objective measures explain the correctness or logicality.

Accordingly, the best models selected exhibited acceptable & better accuracy, >90% and ROC score with lower error rate as presented in previous sections. However, it is also identified that all the models perform relatively less in the case of fatal and injury classes, while their performance is very good in determining non injury risk of an accident. Moreover, though, all the techniques are found to be promising in identification of patterns in a road safety domain, TreeNet is shown to be the best method to be used in the domain under study, if the decision is to use a single method. However, ensemble result proves to be the best of all individual models.

Thus, employing well known and appropriate evaluation techniques in a data mining community (Venkatadri and Lokanatha, 2011) in the process of pattern identification for explaining aspects of road safety adds to the validity of the work under consideration.

7.4 Completeness, Robustness and Practical Utility of RSIA (Survey)

This sub section discusses the procedure and actual evaluation survey results from the respondents.

7.4.1 Evaluation Procedure

A formative evaluation was conducted by the principal researcher and employing the help of two experts in the information system field. The purpose of the evaluation is to test that the study and design of road safety information architecture is a working

7.4.2 Evaluation Survey Results

The evaluation method involved supplying a copy of the architectural descriptions and explanation of each concept in the information architecture to the participants, who are actually experts in road safety data collection and analysis. Twenty five road safety experts whose work is directly related to accident data management have participated in the evaluation process. From a total of 25 participants, 9 were new to the study, while the rest 16 were already part of the research process.

As discussed in the previous section, to confirm the validity of the road safety information architecture, 33 questionnaires containing 24 Likert type items and 5 open ended questions were distributed among road safety experts in road safety departments and traffic offices of the research sites. Finally, 25 completed questionnaires were returned and used for analysis. The questionnaire's reliability was analyzed using one of the reliability analysis features of SPSS 15.0 tool. Accordingly, the Cronbach's Alpha ($\alpha = 0.940$) was calculated (see Annex A) which confirms its reliability.

Regarding the analysis of experts' response for Likert type items, standard deviation and mean of descriptive statistics were used. In view of that the standard deviation was calculated for each item of the survey based on the gathered data. As evident from the table in Annex A, it is easy to deduce that the road safety information architecture is acceptable as the survey exhibits standard deviation less than 1 (0.506 – 0.971) for each item.

Another parameter used in the analysis of the experts' response on the acceptability of the information architecture is comparing the mean score results of completed questionnaire with the questionnaire's average. Hence, the mean score of the 25 completed questionnaires was 99.48, which is by far more than the questionnaire's average score, 75. This expresses that most of the questions in the survey are rated above average by respondents, which indicates that the information architecture defined in this research in the opinion of road safety experts is desirable and sought-after.

The next part of the evaluation survey was composed of 5 open ended questions, for which experts responded accordingly. The first two questions were more of general on information architecture and the next three were more specifically on the road safety information architecture presented in this research. Though, the questions are available for all respondents, based on their high interest to participate in an interview session, 8 road safety experts were approached for further and focused structured interview. The first question was about their understanding of the concept "information architecture" and if their organizations do have any to date.

Accordingly, all the respondents communicated their understanding of the concept information architecture from the discussion made and they expressed that they were not aware of it as it appears in the literatures but assumed that it is a way of organizing information using information and communication technology. Therefore, none of the respondent can provide any further explanation about the concept information architecture, which actually is not a surprise as there is various connotations and understandings in the literature too.

In connection to this, it is also learned that no organized information architecture existed in their respective organization. However, two of the respondents from one of the research sites claim that they consider their procedure in collecting and analyzing accident information as a form of information architecture though not well guided and systematized which really make sense.

When asked about any advantage or disadvantage of using such type of information architecture, respondents list a number of advantages and a couple of disadvantages too. This implies, their significant understanding of the material presented. Major advantages of the use of information architecture includes; facilitate access to and update the accident data, create a platform for further investigation, facilitate accident data quality and integration, promote proper awareness creation, motivate further investigation.

It was also interesting to learn some possible limitations of the use of such type of information architecture. Accordingly, three of the respondents mentioned that use of such type of information architecture requires trained manpower, responsible parties' commitment and cost. However, given the level of priority that the road safety sector has got, the seriousness of the problem worldwide and the increasing awareness of the community we argue that the advantages outweigh the disadvantages.

The third question posed in the evaluation interview is centered on the respondents understanding of the road safety information architecture. Accordingly, all respondent confirmed that they really understand the presented road safety information architecture with a qualifying phrase 'to some extent' and 'somehow' in case of the three respondents. The fourth question involved if the respondents felt any missing elements in the road safety information architecture. Only one respondent mentioned some missing elements. The respondent brings up the inclusion of the role of justice and educational institutions in the road safety information architecture. However, their role is not well visible from the accident information management practice point of view.

The last question was about the envisaged potential use of the information architecture within their respective organizations. While, two of the respondents have a reservation on the implementation of the defined road safety information architecture due to its resource and commitment requirement, the other six envisage the potential use of it. Some of the reasons mentioned for its potential use include, the government attention to the road safety issues and the availability of volunteers in the process of road safety problems mitigation. It is also stated that the architecture will help in reviewing the road safety information management practice, which agrees with the potential contribution of this research.

Last but not least, it is worth mentioning that the proposed architecture is in agreement with the set (7) of integrity rules. This is to mean that, all six columns in RSIA (content, motivation, function, people, place and time) are interchangeable in order, each column has a simple generic model representing one aspect, the basic model of each column is unique, each row describes a distinct perspective, each cell is unique and integration of all cell models in one row constitutes a complete model from the perspective of that row.

7.5 Summary of the Evaluation Process

Understanding the importance of evaluation of the process and the end results of a given research, a multi-method evaluation technique was employed to evaluate and validate the work. Generally, it is found that both the process and result of this research are valid, reliable and acceptable. This is exhibited through the theoretical literature support in the process of the research and the experimental and survey evaluation employed to assess the validity and acceptability of the end result.

The reliability of the evaluation survey items itself was also checked using available reliability analysis techniques in addition to the fact that they are adopted from previous researches. In the case of the structured interview the experience and insight that the road safety experts provide a confidence on their better position to comment the research results.

Peer-reviewed conference and journal publications as indicated in own works (Beshah et al., 2011b, Beshah et al., 2012b, Beshah et al., 2012a) also support the validity of the process and result of this research undertaking. However, from the very nature of such type of works, it is believed that the ultimate benefit of the road safety information architecture will be seen in the long run as it requires time for its implementation and its impact.

Summary of the evaluation process is presented in Table 38. The table summarizes the major aspects of the research evaluated, purpose of the specific evaluation, methods and outcome of each evaluation.

Table 38- Summary of evaluation process

S.n	Aspects Evaluated (What)	Purpose (Why)	Methods of evaluation (How)	Evaluation Outcome
1	Road Safety Information Architecture (RSIA)	To confirm its completeness, clarity and Relevance To test that the study and design of RSIA is a working idea	Surveying Descriptive	RSIA is acceptable and sought after
2	Accident Data Analysis Models	To measure accuracy and interestingness of the models and resulting patterns	Experimental (Accuracy, Error Rate, ROC)	Models perform well and patterns are acceptable
3	The research process	To confirm the validity and reliability of the research	Literature support Expert participation	

Chapter Eight

CONCLUSION AND RECOMMENDATIONS

8.1 Overview

This final chapter presents conclusions from the whole research process and summary of the results. As can be seen from chapter one, the purpose of this specific research was to address four research questions, which were mapped in to 7 specific objectives. This section presents summary, contributions, limitations of the research and conclusions as a result of meeting these objectives. Finally, identified future research lines are provided so as to let others continue in exploring the area and improve the results.

8.2 Summary

Through this research, attempt has been made to extend data mining experiments in explaining road safety situations and investigate application of enterprise information architecture concepts. In doing so CART, TreeNet , RandomForest, J48, PART and Ensemble techniques were explored for road accident data understanding and analysis. Survey of the literature enabled to create understanding of state of the art techniques and attempts in a road safety data quality and data analysis domain. One of the main goals of the research was to empirically explore data quality issues, perform trend analysis and explaining road safety situations. Two of the major issues in explaining road safety situation are the role of road users factors, which is identified to be the major factor, on the risk of injury for a road traffic accident and collision related factors.

As to the road user's behaviour; *road user's category, occupation and age* are found to be more important in determining accident severity while road and environment factors are found to be the least important. With respect to collision related factors it is found that factors like *accident type, accident cause* and *sub-city* are major variables determining accident severity. These optimal attributes are used as an input in the design of content demission of the architecture. In addition to revealing patterns, major contribution of this research work includes comparison of analytical predictive models for the domain, highlighting data quality issues, proposing ensemble technique to improve accuracy and trend analysis regarding factors for accident severity, which also informs the function dimension of the architecture. Moreover, use and comparison of different soft computing techniques on the test bed reveal best approach and accuracy in understanding and predicting road safety patterns. In relation to this, ensemble technique proved to be more effective than separate models. However, TreeNet is better over the others in all evaluation metrics.

Detection of accidents risks due to road users related factors assist in designing appropriate counter measures in the effort of reducing the socio-economic impact of road accidents, which ultimately improve road safety. Another advantage of this approach to road traffic accident data understanding and analysis through machine learning is that the hypothesis can easily be formulated for future trends.

Another major aspect of this research is the definition of road safety information architecture. Extensive survey of literature in the area has been made as discussed in literature review chapter to provide evidence for the applicability of enterprise information architecture in other areas. In line with this, a review of literature also enabled us to create a good understanding of international practices and attempts in improving accident data collection and analysis in a road safety domain.

Accordingly, this research presents information architecture for road safety data collection and analysis systems. The defined architecture is based on Zachman's framework separation of concerns. By applying the top three layers of Zachman hierarchy, it was possible to develop descriptive RSIA that can facilitate communication. The development of RSIA in the study and design of a road safety information management is an original contribution which improves and expands the conceptual framework of the research in both road safety domain and Information architecture field.

RSIA can serve as a strategic guide to the review and development of the road safety data collection and analysis systems. It can also be used as a tool for analysis and re-engineering of existing accident data systems. The result of the research helps road safety organizations to revisit their focus of attention in crafting and implementing measures to reduce road safety problems. More specifically, the research indicated that in addition to drivers, education and enforcement measures should address well other road users like pedestrians. It is also worth mentioning that systematic data collection and quality check along with periodic analysis should get due attention so that other measures will be knowledge driven. The research result can also be used as a hypothesis and/or replicated to other developing countries with similar context in the area of road accident data collection and analysis.

Regarding evaluation of the research results, a multi-method evaluation including surveying, experiment and descriptive techniques were employed to evaluate and validate the work. Generally, it is found that both the process and result of this research are valid and acceptable. This is shown through the theoretical literature support in the process of the research and the experimental and survey evaluation employed to assess the validity and acceptability of the end result.

8.3 Contributions of the Research

This research has investigated existing theory and conceptual frameworks on information architecture and accident data analysis, and established the use in the case of road safety information management. It further aimed to broaden these ideas by explaining road safety situations and bringing in components identified in related research areas, with the ultimate goal being the development of integrated road safety information architecture for road safety domain. In meeting the research objectives, the work has generated a number of outcomes that are contributions to the knowledge and practices.

8.3.1 Contributions to the knowledge

The major contributions of this research are the design artefacts (architectural descriptions) collectively named as the Road Safety Information Architecture (RSIA) and knowledge embodied in it. Because RSIA is the first artefact to address road safety information management from an enterprise perspective, its design by it itself is a contribution to design science. Another interesting contribution includes the design process, as it was possible to clearly indicate the process as a base for the design of the architecture. These contributions advance our understanding of how best to structure information assets. The details of these contributions are presented below.

Architectural description of road safety information and process, which will be used to evaluate existing systems and/or design a new one, is a major contribution to both the design science research and the information architecture. Detail description of the content, process, motivation, network, people and time provides a comprehensive view in addressing road safety issues from information management point of view.

Analyzing the nature of the problem area and introducing enterprise view in a road safety domain is another contribution. Sensing the absence of overarching architectural guide and the disintegration in the effort of addressing road safety problem, this work provides an enterprise view so that stakeholder can view the problem domain from different perspectives. As there was no any road safety information architecture from enterprise perspective, this is an important contribution to the Information Architecture knowledge area. The advantage of having an IA from an enterprise view, specifically for road safety organizations includes providing a platform for standardizing the content and process of road safety information management. With this regard issues related to accident data sources and contents were explored. Some key points comprise a recommendation for inclusion of more attributes and recording traffic events with or without collision including slight misses as it improves road safety understanding by allowing better understanding regarding collision process. It also provides a research framework for future efforts of improving road safety.

It is also worth mentioning the methodological contribution to the IA area on how to develop architectural description under Zachman Framework in a specific domain. As discussed in Chapter 6 the design process in pulling data mining results and empirical qualitative data guided by the theoretical support to define information architecture is a contribution. This is because there is no as such prior research attempting to integrate data mining results and qualitative data using varied modeling tools in developing architectural descriptions. This is imperative as the result and processes of this research are found to be acceptable through the evaluation process.

8.3.2 Contribution to the Practice

This subsection presents contributions of the research to the practice and recommendation resulted from the process of the research. Accordingly, based on the architectural model artefacts, which are results of the review of international practices, successive data mining experiments, empirical data from road safety and traffic departments and study of relevant literature, the following basic implications and recommendations to the practice are identified.

Explanation of the road safety situation is one of the major aspects of the research at hand. This was possible through experimenting and suggesting analytical machine learning models in describing the nature and magnitude of the road safety problem. Predictive models in determining factors contributing to accident occurrence and severity add on the effort being made in understanding road safety situations. Hence, selected accident analysis models can be integrated to make periodic analysis of accident data, which is believed to improve road safety measures which are an important contribution to the domain. In this regard the use of multi-classifier systems is one of the main aspects in this research contributing to both the road safety and the analysis of accident data management. Shading light on the data quality issues and accident data analysis trend conducted through this research were also worth mentioning contributions.

Investigation of Zachman Framework in a road safety domain is also another aspect. Zachman Framework has been used in areas like network security planning, education services delivery, determining the content of digital libraries. This work extends the use of the framework in road safety domain for a road safety information management, which will improve understanding and facilitate communication among road safety organizations.

Regarding accident data collection and reporting, much attention should be given in addressing under reporting and data quality issues as recommended in own works (Beshah et al., 2011b, Beshah et al., 2012b, Beshah et al., 2012a) . This is possible by incorporating a data quality and completeness check process at traffic departments (see the function perspectives of the architectural model). It is also worth mentioning that citizens at the spot of the accident could participate in reporting accidents. Moreover, collecting data related to Pre-accident scenarios like evasive action, vehicle action before accident and after accident situations such as health status of victims, road users reaction etc. will add on the quality and completeness of the data for multi-analysis.

With respect to data handling and information exchange, as can be seen from international practices and due to structural issues in the study area, the master data source is recommended to be at traffic departments. However, road safety departments should also have a copy of this accident data excluding privacy related attributes (see Figure 21). This will allow the road safety departments at various levels to conduct predictive and exploratory analysis as shown in previous experiments (Beshah et al., 2011b, Beshah et al., 2012b, Beshah et al., 2012a) which will be used to continually revise and update countermeasures. Hospital records with a unique ID of an accident should also be integrated with police data through updating process.

Accident analysis and dissemination of its result is another very important function of a road safety system in accident information management. The analysis should focus on well identified aspects based on the priority set. Moreover, in addition to answering the "how many questions", it is also recommended to conduct predictive and exploratory experiments by integrating models into a system. The architecture therefore will allow the integration of open-source and/or proprietary analysis tools in to the overall system. It should be borne in mind that very little study of accidents has been possible in Ethiopia due to the problems related to data availability, completeness and quality. Therefore, it is anticipated that with improvements in the data content, collection and reporting, accident analysis can better be enhanced.

Though the research work has resulted in a number of contributions as mentioned above, the completion of this work does not denote the end of the researcher's study of information architectures. It rather marks the beginning by creating an empirical, theoretical and wide-ranging basis for establishing best way to structure information and successful road safety information management. It is believed that this research can have a significant impact on the state of accident information management as road safety organizations are convinced to make use of the information architecture.

8.4 Limitations of the Research

As can be seen from the methodology section, this research has sought to use a versatile approach: a qualitative empirical data using interview, observation and document analysis; experiments using data mining techniques; critical literature review; own and supervisor's expertise. Despite the various efforts to overcome risks on the quality of the research findings and the process employed, some limitations must be acknowledged.

One such possible limitation of this research undertaking is on the validation of the architectural description artefacts. In relation to this, attempt has been made to provide thick description on the empirical data collection and analysis as well as base on multiple references in the development of artefacts in the research process.

As to the evaluation process, use of questioning approach through survey items and employing descriptive evaluation through theoretical support were the major ones.

Thus, particularly the questioning approach through survey items might not be sufficient to practically show the usefulness of the architectural description as it only investigates the road safety (domain) expert's opinion on the road safety information architecture. The inclusion of IT experts would have made the evaluation stronger. In addition, in order to thoroughly evaluate the architecture in a real setting, it would be necessary to demonstrate through instantiations (prototype) and have road safety organizations adopt this enterprise view based road safety information architecture and then evaluate if the process of accident information management is improved and justify the impact on the improvement of road safety situations in general. However, this would require a longitudinal research which needs a bit longer time and is beyond the scope of this research.

The performance of the machine learning experiments is better compared to the previous experiments done in the research area. Yet the current performance could have been even better if quality data and more powerful computer systems were used.

The fact that the data collection has been confined to only three sites should be acknowledged as another limitation of the research. Moreover, out of the three sites Addis Ababa has been the main source of empirical data as it is a region where the magnitude of road safety problem is high and consequently it is a region with relatively large number of road safety experts.

8.5 Future Research Direction

In this research Road Safety Information Architecture is developed and through the process it is established that future research could examine issues in road safety information management in an integrated manner. The result of this study can be used to support future research related to application of information architecture concepts, especially in the context of road safety. Thus, this work can be viewed as a pioneer in establishing an Information Architecture practice in enterprises in general and in a road safety domain in particular in the study area. Hence, using the same integrated information architecture approach as defined for the road safety information management, other information requirements can also be addressed and provide organizations with a more effective way of managing information assets.

Particularly this research undertaking will be extended to the development of a well established accident data and information management system. Accordingly, the research project will carry on, in collaboration with the relevant road safety organs, by using the deliverables of this research as a guide to evaluate the existing accident information management and documenting the status in all the regional states. The aim of this additional research project could be either to evaluate the RSIA as analyzing instrument or to evaluate it as an information system development tool. This will be followed by the implementations of instantiations to the identified problems. The major output of this research, RSIA, assumes that the target be the accident data collection and analysis at this stage. In the future this can be expanded to include congestion and pollution as well. Future research direction could also be on establishing an architecture framework for integration with other transportation systems.

Moreover, based on the major results of this research, it is logical to recommend that further investigation is required to expand and enrich the defined architecture. One aspect for further research could be including more research sites both locally and in other developing countries with similar context. A more detail and diversified predictive analysis of road accidents using a range of as well as a combination of machine learning techniques is also another potential area of future work. In connection to this, the result of this study can also be used to support future research related to machine learning approaches such as ensemble technique especially in the context of road safety.

8.6 Conclusion

In order for road safety organizations achieve their objective of addressing road safety problems, they must excel in road safety information management. The current approach to accident information management is not providing the various stakeholders with the information and report they require in order to make effective decisions. This is due to the fact that road safety organizations namely Traffic Offices and Road Safety Departments at various levels are working in a fragmented manner. The solution to the problem is in the design and usage of information architecture. By using information architecture approach, road safety organizations develop a better understanding of the content, motivation, process, place, time and people of accident information management.

The primary objective of this research is designing a new model to architect road accident information management. The process is guided by Zachman Framework. To achieve this goal, firstly various concepts related to road safety, data mining approaches, information architecture and the guiding framework are described; Second, various data mining experiments and qualitative data collection and analysis were done to determine information requirements and finally the proposed Road Safety Information Architecture (RSIA) is presented and delineated. Thus, this dissertation resulted in a suitable enterprise information architecture model, which has been related to relevant information architecture goals and information sources. It is believed that as the information architecture practice evolves through time and gets more advancement through subsequent researches, it will be empirically proven in real world use cases which will lead to a development of a strategic document and detail guidelines.

References

AACG 2009. Addis Negari Gazeta: Road traffic safety regulation of the Addis Ababa City Governemnt Addis Ababa: Addis Ababa City Government, AA, Ethiopia

ABD RAZAK, R. 2008. An exploratory study of enterprise architecture practices in Malaysia. *Communications of the IBIMA*, 3, 133-137. ISSN 1943-7765. Available From: <http://www.ibimapublishing.com/journals/CIBIMA/>

ABDULLAH, A. & ZAINAB, A. 2008. The Application of zachman framework in architecting a collaborative digital library. Proceeding of the European Conference on Information management & Evaluation (ECIME 2008). Royal Holloway, University of London, United Kingdom 11-12 September 2008, 11p. Available From: <http://dspace.fsktm.um.edu.my/handle/1812/219>

ABOWD, G., BASS, L., CLEMENTS, P., KAZMAN, R. & NORTHROP, L. 1997. Recommended Best Industrial Practice for Software Architecture Evaluation. No. CMU/SEI-96-TR-025, Carnegie-Mellon University, Pittsburgh, Software Engineering Institute. DTIC Document.

ABUGESSAISA, I. 2008a. *Analytical tools and information-sharing methods supporting road safety organizations.* PhD Thesis. Linköping University, Linköping.

ABUGESSAISA, I. 2008b. Knowledge discovery in road accidents database. *International Journal of Public Information Systems,* 4, No 1, 59-85. Available From: <http://www.ijpis.net/ojs/index.php/IJPIS/article/view/54>

ACUNA, E. & RODRIGUEZ, C. 2004. The treatment of missing values and its effect on classifier accuracy. *Classification, Clustering, and Data Mining Applications*, Springer Berlin Heidelberg, pp. 639-647.

ADB 1996. Road safety research in the Asian pacific region: Technical note No.1 Review of recent projects and researches.: Asian Development Bank.

AFUKAAR, F. 2007. Road Traffic Injury Data Systems in Ghana:The key to safety improvement and control. Paper presented at *African Road Safety Conference.* Kwame Nkrumah Conference Centre, Accra, Ghana

AFUKAAR, F. K., ANTWI, P. & OFOSU-AMAAH, S. 2003. Pattern of road traffic injuries in Ghana: implications for control. *Injury Control and safety promotion,* 10, issue 1-2, 69-76. Available From: <http://www.tandfonline.com/doi/abs/10.1076/icsp.10.1.69.14107#.Uj6r0H-F2Bo>

ANASTASOPOULOS, P. C. & MANNERING, F. L. 2011. An empirical assessment of fixed and random parameter logit models using crash-and non-crash-specific injury data. *Accident Analysis & Prevention,* 43, 1140-1147.

ANTENEH, F. 2011. Mining Road Traffic Accident data For Predicting Accident Severity to Improve Public Health- Role of Driver and Road Factors in the case of Addis Ababa. Masters Thesis, Addis Ababa University.

AVISON, D. & MYERS, M. D. 2002. *Qualitative research in information systems: a reader,* SAGE Publications. London.

AXELROD, S. 2010. Quality Data Through Enterprise Information Architecture. Available From: <http://msdn.microsoft.com/en-us/library/bb266338.aspx>

BAGULEY, C. 2001. The importance of a road accident data system and its utilisation. *Presented at International Symposium on Traffic Safety Strengthening and Accident Prevention.* Nanjing, China, 28 -30 November 2011. Available From: <http://www.transport-links.org/

transport_links/filearea/publications/1_795_pa3807-02.pdf>.

BAHILL, A. T., BOTTA, R. & DANIELS, J. 2006. The Zachman framework populated with baseball models. *Journal of enterprise architecture,* 2, No 4, 50-68.

BARKER, I. 2005. What is information architecture? KM Column. Available From: <http://www.steptwo.com.au/papers/kmc_whatisinfoarch/index.html>

BENER, A., ABU-ZIDAN, F. M., BENSIALI, A. K., AL-MULLA, A. A. & JADAAN, K. S. 2003. Strategy to improve road safety in developing countries. *Saudi Medical Journal,* 24, 603-608.

BERRY, M. & LINOFF, G. 2000. Data mining techniques and algorithms. *John Wiley & Sons, Inc.: USA.*

BESHAH, T., ABRAHAM, A. & EJIGU, D. 2011a. The use of Information Architecture Towards effective Road Safety Data Managment. 4th *ICT - a New Paradigm for Citizen Services conference.* 29, June - 03, July 2011, Addis Ababa, Ethiopia.

BESHAH, T., EJIGU, D. & ABRAHAM, A. 2012a. A novel road safety information architecture (RSIA): An enterprise view. *Information and Communication Technologies (WICT), 2012 World Congress.* Oct. 30 -Nov. 2, 2012 pp.1127,1134, doi: 10.1109/WICT.2012.6409244 IEEE.

BESHAH, T., EJIGU, D., ABRAHAM, A., SN ŠEL, V. & KR MER, P. 2012b. Knowledge discovery from road traffic accident data in Ethiopia: Data quality, ensembling and trend analysis for improving road safety. *Neural Network World,* 22, 215.

BESHAH, T., EJIGU, D., ABRAHAM, A., SNASEL, V. & KROMER, P. 2011b. Pattern recognition and knowledge discovery from road traffic accident data in ethiopia: Implications for improving road safety. *Information and Communication Technologies (WICT), 2011 World Congress.*. pp 1241 - 1246, doi: 10.1109/WICT.2011.6141426, IEEE.

BESHAH, T. & HILL, S. 2010. Mining road traffic accident data to improve safety: role of road-related factors on accident severity in Ethiopia. *Proceedings of AAAI Artificial Intelligence for Development (AI-D'10),* pp 22-24. Stanford, CA, USA, March 2010. Avaulable From:< http://www.aaai.org/ocs/index.php/SSS/SSS10/paper/viewPDFInterstitial/1173/1343>.

BHALLA, K., NAGHAVI, M., SHAHRAZ, S., BARTELS, D. & MURRAY, C. 2009. Building national estimates of the burden of road traffic injuries in developing countries from all available data sources: Iran. *Injury prevention,* 15, 150-156.

BOBEVA, M. 2005. *A Framework for Information Architecure for Business Networks.* PhD Thesis, Bournemouth University.

BORZACCHIELLO, M. T., TORRIERI, V. & NIJKAMP, P. 2009. An operational information systems architecture for assessing sustainable transportation planning: principles and design. *Evaluation and Program Planning,* 32, 381-389.

BRAA, J., HANSETH, O., HEYWOOD, A., MOHAMMED, W. & SHAW, V. 2007. Developing Health Information Systems in Developing Countries. *The Flexible Standards Strategy. MISQ,* Vol. 31, No. 2 (Jun., 2007), pp. 381-402. Available :< http://www.jstor.org/stable/25148796>

BURES, P. 2009. The architecture of traffic and travel information system based on protocol TPEG. *Proceedings of the 2009 Euro American Conference on Telematics and Information Systems: New Opportunities to increase Digital Citizenship.* ACM. Available From : <http://dl.acm.org/citation.cfm?id=1551743>.

BURNARD, P. 2004. Writing a qualitative research report. *Nurse education today,* 24, 174-179.

CARCARY, M. 2011. Design Science Research: The Case of the IT Capability Maturity Framework (IT CMF). *Electronic Journal of Business Research Methods,* 9 (2), pp 109-118. Available From:< http://www.ejbrm.com/issue/download.html?idArticle=260>

CARVALHO, J. Á. 2012. Validation criteria for the outcomes of design research. *IT Artefact Design & Workpractice Intervention, a Pre-ECIS and AIS SIG Prag Workshop.* June 10, 2012, Barcelona. Available From: <http://hdl.handle.net/1822/21713>

CHANG, L.-Y., CHU, H.-C., LIN, D.-J. & LUI, P. 2012. Analysis of Freeway Accident Frequency using Multivariate Adaptive Regression Splines. *Procedia Engineering,* 45, 824-829.

CHANG, L.-Y. & WANG, H.-W. 2006. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention,* 38, 1019-1027.

CHEN, B., WENG, X., WANG, B. & HU, X. 2009. Analysis and solution of data quality in data warehouse of Chinese materia medica. *Computer Science & Education, 2009. ICCSE'09 4th International Conference pp* 823 - 827. IEEE. 25-28 July 2009. Available From: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5228165&isnumber=5228118>

CHONG, M., ABRAHAM, A. & PAPRZYCKI, M. 2004. Traffic accident data mining using machine learning paradigms. *Fourth International Conference on Intelligent Systems Design and Applications (ISDA'04), Hungary,* ISBN (Vol. 1047219710, pp. 415-420). Available From : <http://isda03.softcomputing.net/isda-mam.pdf>.

CLEVEN, A., GUBLER, P. & H NER, K. M. 2009. Design alternatives for the evaluation of design science research artifacts. *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology.* ACM. Philadelphia, PA, USA, May 06 - 08, 2009. Available From: <http://dl.acm.org/citation.cfm?id=1555645>

CRESWELL, J. 1998. Qualitative inquiry and research design. Choosing among five traditions. *Lincoln: Sage.*

CRESWELL, J. 2002. *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research,* Saddle River, NJ, Prentice Hall.

CRESWELL, J. W. 2009. *Research design: Qualitative, quantitative, and mixed methods approaches,* Sage Publications, Incorporated.

CRESWELL, J. W. & CLARK, V. L. P. 2007. *Designing and conducting mixed methods research,* Wiley Online Library.

CRESWELL, J. W. & MILLER, D. L. 2000. Determining validity in qualitative inquiry. *Theory into practice,* 39, 124-130.

DAIGAVANE, P. & BAJAJ, P. 2009. Analysis of selective parameters contributing to road accidents on highways for establishing suggestive precautionary strategies. *Emerging Trends in Engineering and Technology (ICETET), 2009 2nd International Conference on.* IEEE. pp. 576, 580, 16-18 Dec. 2009. Available From: < http://ieeexplore.ieee.org /stamp/stamp.jsp?tp=&arnumber=5395401&isnumber=5394994>

DARKE, P., SHANKS, G. & BROADBENT, M. 1998. Successfully completing case study research: combining rigour, relevance and pragmatism. *Information systems journal,* 8, 273-289.

DAUD M.N.R., A. C., D.W., 2007. Human Readable Rule Induction In Medical Data Mining: A Survey Of Existing Algorithms. *WSEAS European Computing Conference.* Sepetmber 25 - 27, 2007, Athens, Greece.

DAVIS, A., QUIMBY, A., ODERO, W., GURURAJ, G. & HIJAR, M. 2003. Improving road safety by reducing impaired driving in developing countries: A scoping study. Unpublished Report. *Transport Research Laboratory*. Availabe From: http://grsp.drupalgardens.com/sites

/grsp. drupalgardens.com/files/Impaired%20driving%20full%20report.pdf
DE O A, J., MUJALLI, R. O. & CALVO, F. J. 2011. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. *Accident Analysis & Prevention*, 43, 402-411.

DELEN, D., SHARDA, R. & BESSONOV, M. 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. *Accident Analysis & Prevention*, 38, 434-444.

DOMINGUEZ, A. n.d. Zachman Architecure- Based Education Service Delivery. Available From: <http://www.poms.org/conferences/cso2007/talks/10.pdf>

DONG, A. & AGOGINO, A. M. 2001. Design principles for the information architecture of a SMET education digital library. *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries.* September 8 - 12 2001, London, UK. ACM. Available From: <http://dl.acm.org/ citation.cfm?id=379699>

DORAN, P. 2007. *Information architecture for Irish grocery retailers using business intelligence tools.* Masters Thesis, Dublin Institute of Technology.

DOWNEY, L. & BANERJEE, S. 2011. Building an information architecture checklist. *Journal of Information Architecture,* Vol 2 Issue 2. Availabe From: <http://journalofia.org /volume2/issue2/03-downey/jofia-0202-03-downey.pdf>

DUCH, W., JANKOWSKI, N. & GRĄBCZEWSKI, K. 2005. Computational intelligence methods for information understanding and information management. *The 4th International Conference on Information and Management Sciences (IMS2005).* July 1-10, 2005, Kunming & Dali & Shangrila, China. Available From: <http://www.fizyka.umk.pl/~norbert/ publications /05-CI-info.pdf>

EKSLER, V. 2007. The role of structural factors in road safety. Young Researchers Seminar 2007. European Conference of Transport. 03 September 2007, Brno, Czech Republic. Available From: <http://www.ectri.org/YRS07/Papiers/Session-7/Eksler.pdf>

ELISH, M. O. & ELISH, K. O. 2009. Application of treenet in predicting object-oriented software maintainability: A comparative study. *Software Maintenance and Reengineering, 2009. CSMR'09. 13th European Conference on.* pp.69,78, 24-27 March 2009, IEEE. Availabe From: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4812740&isnumber=481 2721>

ELVIK, R. 2010. Why some road safety problems are more difficult to solve than others. *Accident Analysis & Prevention,* 42, 1089-1096.

ELVIK, R., VAA, T. & ERKE, A. 2006. *The handbook of road safety measures*, Emerald Group Publishing.

ERTAUL, L., PASHAM, A. R. & PATEL, H. 2010. Enterprise Security Planning using Zachman Framework: Designer's Perspective. Available: <http://weblidi.info.unlp.edu.ar /worldcomp2011-mirror/SAM4357.pdf>

ERTAUL, L. & SUDARSANAM, R. 2005. Security Planning Using Zachman Framework for Enterprises. *Proceedings of EURO mGOV. July 10-12, 2005, Brighton, UK. Availabe From:* <http://www.m4life.org/ proceedings/2005/PDF/16_S039EL-S13.pdf.>

EVERNDEN, R. 1996. The information framework. *IBM Systems Journal,* 35, 37-68.

EVERNDEN, R. & EVERNDEN, E. 2003. Third-generation information architecture. *Communications of the ACM,* 46, 95-98.

EVGENIOU, T. & CARTWRIGHT, P. 2005. Barriers to information management. *European*

Management Journal, 23, 293-299.

FARKAS, B. & WEISTROFFER, H. R. 2012. Some Information Systems Research Trends over the Past Decade. Available From :< http://sais.aisnet.org/2012/FarkasWeistroffer.pdf>

FARZI, S. & DASTJERDI, A. B. 2010. Data quality measurement using data mining. *International Journal of Computer Theory and Engineering,* 2, 1793-8201.

FATOLAHI, A. & SHAMS, F. 2006. An investigation into applying UML to the Zachman framework. *Information Systems Frontiers,* 8, 133-143.

FAZIL, Q.-A., ABDULLAH, Z. & NOAH, S. 2010. Applying Zachman Framework to determine the content of semantic theses digital library. *Information Technology (ITSim), International Symposium.* Vol. 3, pp.1596, 1600, 15-17 June 2010, IEEE. Available From: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5561461&isnumber=5561452>

FRIDSTROM, L. 1999. Economoetric models of road use, accidents and road investement decisions Available From: <www.e-ajd.net/source-pdf/R457_1999.pdf>

FRIEDMAN, J. H. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis,* 38, 367-378.

FRIEDMAN, J. H. & MEULMAN, J. J. 2003. Multiple additive regression trees with application in epidemiology. *Statistics in medicine,* 22, 1365-1381.

GEY, S. & NEDELEC, E. 2005. Model selection for CART regression trees. *Information Theory, IEEE Transactions on,* 51, 658-670.

GIBBS, G. R. 2008. *Analysing qualitative data*, Sage Publications Limited.

GIDEY, A. 2010. Pedestrian vulnerability to road traffic accident in Addis Ababa. Ethiopian Transport Authority. Addis Ababa, Ethiopia.

GOETHALS-SAP-LEERSTOEL, F. 2004. An Overview of Enterprise Architecture Framework Deliverables. Available From:< http://econweb.econ.kuleuven.ac.be/fetew/pdf_publicaties/0570.pdf>

HAI, N. H. 2009. Traffic accidents in Hanoi: data collection and analysis. *4th IRTAD CONFERENCE.* 16-17 September 2009, Seoul, Korea. Available From: <http://internationaltransportforum.org/irtadpublic/pdf /seoul/7- BinHashim.pdf>

HAIXIA, Y. & ZHIHONG, N. 2010. An application of factor analysis on road traffic accident. *Computer Science and Education (ICCSE), 5th International Conference on.* 24-27 Aug. 2010, pp.1355, 1358, IEEE. Available From: < http://ieeexplore.ieee.org/stamp/stamp.jsp?tp= &arnumber=5593713&isnumber=5593391>

HALL, L., BOWYER, K., KEGELMEYER, W., MOORE, T. & CHAO, C. 2000. Distributed learning on very large data sets. Proceedings of the Sixth ACM SIGKDD. International Conference on Knowledge Discovery and Data Mining. Availabe From: <http://citeseerx.ist.psu.edu/ viewdoc/download?doi=10.1.1.37.2971&rep=rep1&type=pdf>

HALL, R. W. 1995. The architecture of transportation systems. *Transportation Research Part C: Emerging Technologies,* 3, 129-142.

HAN, J. & KAMBER, M. 2006. *Data mining: concepts and techniques*, Morgan Kaufmann.

HAND, D. J., MANNILA, H. & SMYTH, P. 2001. *Principles of data mining*, MIT press.

HEVNER, A. R. 2007. The three cycle view of design science research. *Scandinavian Journal of Information Systems,* 19, 87.

HEVNER, A. R., MARCH, S. T., PARK, J. & RAM, S. 2004. Design science in information systems research. *MIS quarterly,* 28, 75-105.

HIZAL HANIS, H. & SHARIFAH ALLYANA, S. 2009. The construction of road accident analysis and database system in Malaysia. *4th IRTAD Conference.* 16-17 September 2009, Seoul, Korea. Available From: <http://internationaltransportforum.org/irtadpublic/pdf/seoul/7-BinHashim.pdf>

HONG, S. J. & WEISS, S. M. 2001. Advances in predictive models for data mining. *Pattern Recognition Letters,* 22, 55-61.

HONGGUO, X., HUIYONG, Z. & FANG, Z. 2010. Bayesian network-based road traffic accident causality analysis. *Information Engineering (ICIE), 2010 WASE International Conference.* IEEE. vol.3, pp. 413, 417. Beidaihe, Hebei, China, 14-15 August, 2010. Available From: < http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5571612&isnumber=5570821>

HYDER, A. A. & BISHAI, D. 2012. Road Safety in 10 Countries: A Global Opportunity. *Traffic Injury Prevention,* 13, 1-2.

JAFARI, M., AKHAVAN, P. & NOURANIPOUR, E. 2009. Developing an architecture model for enterprise knowledge: an empirical study based on the Zachman framework in Iran. *Management Decision,* 47, 730-759.

JANECKA, K. & HULOVA, H. 2011. Using spatial data mining to discover the hidden rules in the crime data. *GIS Ostrava.* Ostrava, Czhec Republic, 23 – 26, January 2011. Available From: < http://gis.vsb.cz/GIS_Ostrava/GIS_Ova_2011/sbornik/papers/Janecka.pdf. >

JANUZAJ, E. & JANUZAJ, V. 2009. An application of data mining to identify data quality problems. *Advanced Engineering Computing and Applications in Sciences. ADVCOMP'09. Third International Conference.* October 11-16, 2009 - Sliema, Malta. IEEE. Avaialbe From: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5359651&isnumber=5359431>

JARVIS, J. & KAMAL, S. 2009. Crash Data System - A new-generation software product approach and a move to improved national systems. *4th IRTAD CONFERENCE.* 16-17 September 2009, Seoul, Korea. Available From: <http://internationaltransportforum.org/irtadpublic/pdf/seoul/7- BinHashim.pdf>

JINLIN, W., XI, C., KEFA, Z., WEI, W. & DAN, Z. 2008. Application of spatial data mining in accident analysis system. *Education Technology and Training, 2008. and 2008 International Workshop on Geoscience and Remote Sensing. ETT and GRS 2008. International Workshop on.* Vol. 1, pp. 472-475, IEEE.

K LLGREN, A. 2008. Towards a Framework for Enterprise Architecture at Vattenfall: Identifyig sutable Models. Master's Thesis, Stockholm, Sweden.

KAR, K. & DATTA, T. K. 2009. An Overview of Mobility and Safety Issues Related to Highway Transportation in India. *ITE Journal,* Vol. 79, No.8, pp 40 - 45.

KASHANI, A. T. & MOHAYMANY, A. S. 2011. Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. *Safety Science,* 49, 1314-1320.

KHAN, M. A., AL KATHAIRI, A. S. & GRIB, A. 2004. A GIS based traffic accident data collection, referencing and analysis framework for Abu Dhabi. *In Proceeding Codatu XI in.* April, 2004, Bucharest, Romania Available From: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.127.5039&rep=rep1&type=pdf.

KIFLE, H. 2009. *"Application of data mining technology to support the prioritization of dangerous crash location: the case of addis ababa traffic office.* Master's thesis, Addis Ababa

University.

KILPEL INEN, T. & NURMINEN, M. 2007. Applying Genre-Based Ontologies to Enterprise Architecture. *paper presented to18th Australasian Conference on Information Systems–ACIS.* December 4, 2007. University of Southern Queensland. Toowoomba Australia. Avaialbe From: <http://urn.fi/URN:NBN:fi:jyu-201303081304>

KITTLER, J., HATEF, M., DUIN, R. P. W. & MATAS, J. 1998. On combining classifiers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* No.20, pp.226-239.

KONONOV, J. & JANSON, B. N. 2002. Diagnostic Methodology for the detection of safety problems at intersections. *Transportation Research Record: Journal of the Transportation Research Board,* 1784 (1), pp. 51-56.

KRISHNAVENI, S. & HEMALATHA, M. 2011. A perspective analysis of traffic accident using data mining techniques. *International Journal of Computer Applications,* Vol. 23 No.7.

KUMAR, R. 2005. *Research Methodology: A step-by-step guide for biginners,* London, Sage Publications.

LAGARDE, E. 2007. Road traffic injury is an escalating burden in Africa and deserves proportionate research efforts. *PLoS medicine,* 4, 170.

LARSON, K., HENNING, K. & PEDEN, M. 2012. The importance of data for global road safety. *Traffic Injury Prevention,* 13, 3-4.

LARSSON, P., DEKKER, S. W. & TINGVALL, C. 2010. The need for a systems theory approach to road safety. *Safety Science,* 48, 1167-1174.

LEVESON, N. G. 2002. System safety engineering: Back to the future. *Massachusetts Institute of Technology*. Available From : <http://203.253.146.106/images/44b/Book2.pdf>

LI, T., CHEN, Y., QIN, S. & LI, N. 2011. Highway road accident analysis based on clustering ensemble. In *Computer Science for Environmental Engineering and EcoInformatics*, pp. 212-217. Springer Berlin Heidelberg.

LIU, P. 2009. A self-organizing feature maps and data mining based decision support system for liability authentications of traffic crashes. *Neurocomputing,* 72, 2902-2908.

MAHMOOD, Z. 2006. Architectural representations for describing enterprise information and data. *Proceedings 10th WSEAS conference on computers.* Vouliagmeni, Athens, Greece, July 13-15, 2006 pp. 728 - 733 Avaliable from: <http://www.wseas.us/e-library/conferences /2006cscc/papers/534-735.pdf.>

MALLOY, M. A., MASEK, E. V., MILLER, R. W. & WINKOWSKI, D. G. 2010. An Information Architecture Framework for the USAF. Available From:<*www.dtic.mil/dtic/tr/fulltext/u2/a576472.pdf*>

MARSHALL, M. N. 1996. Sampling for qualitative research. *Family practice,* 13, 522-526.

MATHERS, C., FAT, D. M. & BOERMA, J. 2008. *The global burden of disease: 2004 update*, World Health Organization.

MENON, G. R., GURURAJ, G., TAMBE, M. & SHAH, B. 2010. A multi-sectoral approach to capture information on road traffic injuries. *Indian Journal of Community Medicine: Official Publication of Indian Association of Preventive & Social Medicine,* 35, 305.

MOHAJERANI, M. & MOEINI, A. 2004. Using enterprise architecture framework to design network security architecture. *WSEAS Transactions on Communications,* 3, 688-693.

MOHAN, D. & ROBERTS, I. 2001. Global road safety and the contribution of big business: Road safety policies must be based on evidence. *BMJ: British Medical Journal,* 323, 648.

MOON, J. 2003. APRAD and road safety in ASEAN. *The 3rd GRSP ASEAN Seminar.* UNESCAP, Bangkok, Thailand. March, 2003.

MORGAN, A. & MANNERING, F. L. 2011. The effects of road-surface conditions, age, and gender on driver-injury severities. *Accident Analysis & Prevention*, 43, 1852-1863.

MORVILLE, P. & ROSENFELD, L. 2006. *Information Architecture for the world wide web: designing large-scale web sites*, O'Reilly Media, Incorporated.

MOSEDALE, J., PURDY, A. & CLARKSON, E. 2004. Contributory factors to road accidents. *Transport Statistics: Road Safety.* Department for Transport. Available From: < http://trid.trb.org/view.aspx?id=864416>

MOSSIE, G. 2009. *"Applying data mining with decision tree and rule induction techniques to identify determinant factors of drivers and vehicles in support of reducing and controlling road traffic.* Master's thesis, Addis Ababa University.

MYERS, M. D. 2009. *Qualitative research in business & management*, Sage Publications Limited.

NAKAHARA, S. & WAKAI, S. 2001. Underreporting of traffic injuries involving children in Japan. *Injury prevention*, 7, 242-244.

NAYAK, R., EMERSON, D., WELIGAMAGE, J. & PIYATRAPOOMI, N. 2011. Road crash proneness prediction using data mining. *Proceedings of the 14th International Conference on Extending Database Technology.* Uppsala, Sweden, March 21 - 24, 2011, ACM. Avaialbe From: < http://dl.acm.org/citation.cfm?id=1951365>

NIEDERMAN, F., BRANCHEAU, J. C. & WETHERBE, J. C. 1991. Information systems management issues for the 1990s. *MIS quarterly*, 15, 475-500.

NIKOLAIDOU, M. & ALEXOPOULOU, N. 2008. Enterprise Information System Engineering: A Model-Based Approach Based on the Zachman Framework. *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual.* IEEE. pp.399, 399, 7-10 January, 2008. Available From: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp= &arnumber=4439104&isnumber=4438696>

NRSA 2006. Research, Information and Development Policy: Governepont Document. The national Road Safety Authority of Israel : Information, Research and Development Division. Israel.

NRSC 2011. National Road Safety Strategic Plan - 2011 to 2020. Addis Ababa: National Road Safety Council. Addis Ababa, Ethiopia.

OFFERMANN, P., LEVINA, O., SCH NHERR, M. & BUB, U. 2009. Outline of a design science research process. *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology.* ACM. Philadelphia, PA, USA — May 06 - 08, 2009. Available From: <http://dl.acm.org/citation.cfm?id=1555619>

ONWUEGBUZIE, A. J. & COLLINS, K. M. 2007. A typology of mixed methods sampling designs in social science research. *The Qualitative Report*, 12, 281-316.

ONWUEGBUZIE, A. J. & LEECH, N. L. 2007. Sampling designs in qualitative research: Making the sampling process more public. *The Qualitative Report*, 12, 238-254.

PADGETT, D. K. 2008. *Qualitative methods in social work research*, Sage Publications, Incorporated.

PAKGOHAR, A., TABRIZI, R. S., KHALILI, M. & ESMAEILI, A. 2011. The role of human factor in incidence and severity of road crashes based on the cart and lr regression: a data mining approach. *Procedia Computer Science*, 3, 764-769.

PARIS, I. H. M., AFFENDEY, L. S. & MUSTAPHA, N. 2010. Improving academic performance prediction using voting technique in data mining. *World Academy of Science, Engineering and Technology*, Vol. 62.

PEDEN, M., RICHARD S., DAVID S., DINESH M., ADNAN A.H., J., E. & M., C. (eds.) 2004. *World Report on road traffic injury prevention*, Geneva: World Health Organization.

PEREIRA, C. M. & SOUSA, P. 2004. A method to define an Enterprise Architecture using the Zachman Framework. *Proceedings of the 2004 ACM symposium on Applied computing*. ACM. March 14-17, *2004*, Nicosia, Cyprus. Available From: http://dl.acm.org/citation.cfm?id=968175

PERSSON, A. 2008. Road traffic accidents in Ethiopia: magnitude, causes and possible interventions. *Advances in Transportation Studies*, vol. 15, pp 5-16.

PIENIM KI, T. 2005. A Business Application Architecture Framework in Manufacturing Industry. *Tampereen teknillinen yliopisto. Julkaisu-Tampere University of Technology. Publication; 530*. ISBN 952-15-1416-7, Tampere, Finland

PUVANACHANDRA, P., HOE, C., EL-SAYED, H., SAAD, R., AL-GASSEER, N., BAKR, M. & HYDER, A. 2012. Road Traffic Injuries and Data Systems in Egypt: Addressing the Challenges. *Traffic Injury Prevention*, 13, 44-56.

RADWAN, A. & AARABI, M. 2011. Study of Implementing Zachman Framework for Modeling Information Systems for Manufacturing Enterprises Aggregate Planning. *Simulation*, 16, 18.

RANAWANA, R. & PALADE, V. 2006. Multi-Classifier Systems: Review and a roadmap for developers. *International Journal of Hybrid Intelligent Systems*, 3, 35-61.

REGASSA, Z. 2009. *Determining the degree of driver's responsibility for car accident: the case of addis ababa traffic office*. Master's thesis, Addis Ababa University.

RESMINI, A. & ROSATI, L. 2011. A Brief History of Information Architecture. *Journal of Information Architecture*, Vol 3, Issue 2.

REZAEI, R. & SHAMS, F. 2008. A methodology to create data architecture in Zachman framework. *World Applied Sciences Journal*, 3, 43-49.

ROBBINS, D. E., GURUPUR, V. P. & TANIK, J. 2011. Information architecture of a clinical decision support system. *Southeastcon, Proceedings of IEEE*. pp.374, 378, 17-20 March 2011, IEEE. Avaialbe From: < http://ieeexplore.ieee.org/stamp/ stamp.jsp?tp=&arnumber=5752969 &isnumber=5752891>

ROBSON, C. 2002. *Real world research: a resource for social scientists and practitioner-researchers*, Blackwell Oxford.

ROSENFELD, L. 2002. Information architecture: Looking ahead. *Journal of the American society for information science and technology*, 53, 874-876.

ROSENFELD, L. & MORVILLE, P. 2002. *Information architecture for the World Wide Web: Designing large-scale Web sites*, O'Reilly Media, Incorporated.

RUNYAN, C. W. 1998. Using the Haddon matrix: introducing the third dimension. *Injury prevention*, 4, 302-307.

RUNYAN, C. W. 2003. Introduction: back to the future—revisiting Haddon's conceptualization of injury epidemiology and prevention. *Epidemiologic Reviews*, 25, 60-64.

SABEL, C., KINGHAM, S., NICHOLSON, A. & BARTIE, P. 2005. Road Traffic Accident Simulation Modelling-A Kernel Estimation Approach. *The 17th Annual Colloquium of the Spatial Information Research Centre*. 24th - 25th November, *University of Otago, Dunedin, New Zealand*.

SAMSON, F. 2006. *Analysis of Traffic Accident In Addis Ababa: Traffic Simulation*. Master's Thesis, Addis Ababa University.

SAUNIER, N., MOURJI, N. & AGARD, B. 2011. Mining Microscopic Data of Vehicle Conflicts and

Collisions to Investigate Collision Factors. *Transportation Research Record: Journal of the Transportation Research Board*, 2237, 41-50.

SAVOLAINEN, P. T., MANNERING, F. L., LORD, D. & QUDDUS, M. A. 2011. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis & Prevention*, 43, 1666-1676.

SCHAPIRE, R. E. 1999. A brief introduction to boosting. *International Joint Conference on Artificial Intelligence*. LAWRENCE ERLBAUM ASSOCIATES LTD. In *Ijcai*, Vol. 99, pp. 1401-1406. Available From: <http://www.cs.princeton.edu/~schapire/papers/Schapire99c.pdf>

SOHN, S. Y. & SHIN, H. 2001. Pattern recognition for road traffic accident severity in Korea. *Ergonomics*, 44, 107-117.

SOWA, J. F. & ZACHMAN, J. A. 1992. Extending and formalizing the framework for information systems architecture. *IBM Systems Journal*, 31, 590-616.

STATA, R. 1989. Organizational Learning—The Key to Management Innovation. *The training and development sourcebook* 2 (1989)

STEINBERG, D. & GOLOVNYA, M. 2006. CART 6.0 User's Manual. *Salford Systems, San Diego, CA*.

STEINBERG, D., GOLOVNYA, M. & TOLLIVER, D. 2002. TreeNet User Guide. Salford Systems. *Inc, San Diego, CA*.

STOOP, J. A. & THISSEN, W. A. 1997. Transport safety: Trends and challenges from a systems perspective. *Safety Science*, 26, 107-120.

SULAIMAN, S., MARIYAM SHAMSUDDIN, S., ABRAHAM, A. & SULAIMAN, S. 2011. Intelligent web caching using machine learning methods. *Neural Network World*, 21, 429.

TABOR, A. 2009. Bayesian approach for analysis of road traffic accidents: The case of Addis Ababa. Master's thesis, Addis Ababa Univesity.

TAN, P., STEINBACH, M. & KUMAR, V. 2006. Introduction to Data Mining, Ed. Pearson Addison Wesley Educational Publishers, Boston, MA, USA.

TESEMA, T. B., ABRAHAM, A. & GROSAN, C. 2005. Rule mining and classification of road traffic accidents using adaptive regression trees. *International Journal of Simulation*, 6, 80-94.

THEARLING, K. 2003. An introduction to data mining. *Whitepaper. Avaialbel From: < http://www3.shore.net/~ kht/dmwhite/dmwhite. htm>*

THOMPSON, C. L. 2006. *Scaling the Zachman Framework: A Software Development Methodology for Non-enterprise Applications.* Doctoral Dissertation, Regis University.

TIGLACO, N. C. C. 1998. Development of Traffic Accident Information System using Geographic Information System (GIS) *AARS. In Proceddings of 19th Asian Confrornce On Remote Senseing, 1998.*

TORMO, M. T., SANMARTIN, J. & PACE, J.-F. 2009. Update and improvement of the traffic accident data collection procedures in Spain: The METRAS method of sequencing accident events. *4th IRTAD Conference. Seoul, Korea.* 16-17 September 2009. Available From: <http://internationaltransportforum.org/irtadpublic/pdf/seoul/7- BinHashim.pdf>

TWO CROWS CORPORATION. 2005. Introduction to Data Mining and Knowledge Discovery. Available: <http://www.twocrows.com/glossary.htm >

URBACZEWSKI, L. & MRDALJ, S. 2006. A comparison of enterprise architecture frameworks. *Issues in Information Systems*, 7, 18-23.

VAN DER WALT, P. W. & DU TOIT, A. 2007. Developing a scaleable information architecture for an enterprise-wide consolidated information management platform. *Aslib Proceedings*. Vol. 59.

No. 1. Emerald Group Publishing Limited.

VASCONCELOS, E. 2000. Strategies to improve traffic safety in Latin America. *Workshop on Urban Transport Strategy., Santiago de, Chile, November 6 - 9 2000.* Available From: <http://siteresources.worldbank.org/INTURBANTRANSPORT/Resources/2vasconc.pdf>

VENABLE, J. 2006. The role of theory and theorising in design science research. *Proceedings of the 1st International Conference on Design Science in Information Systems and Technology (DESRIST 2006).* February 24-25, 2006, Claremont, California. Available from: < http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.110.2475&rep=rep1&type=pdf>

VENKATADRI, M. & LOKANATHA, C. R. 2011. A Review on Data mining from Past to the Future. *International Journal of Computer Applications,* 15, 19-22.

VORKO-JOVIĆ, A., KERN, J. & BILOGLAV, Z. 2006. Risk factors in urban road traffic accidents. *Journal of Safety Research,* 37, 93-98.

WANAS, N. M. 2003. Multiple Classifier Systems. U. W. Multiple Classifiers Focus Group, Ed., Avaialbel From: <http://watsup.uwaterloo.ca/groups/mcs/MCSFGmeeting_Dec12.pdf>

WHITE, M. 2004. Information Architecture: Viewpoint. *Emerald -The Electronic Library,* 22, 218-219.

WHO 2009 Global status report on road safety: time for action. Geneva: World Health Organization.

WHO 2013. Global status report on road safety 2013: Supporting a decade of action. Luxembourg: World Health Organization.

WILLIAMS, A. F. 1999. The Haddon matrix: its contribution to injury prevention and control. In McClure, Roderick (Ed.) *Third National Conference on Injury Prevention and Control,* 9-12 May 1999, Brisbane, Queensland. Avaialbe From:<http://eprints.qut.edu.au/10081/>

WITTEN, I. H. & FRANK, E. 2005. *Data Mining: Practical machine learning tools and techniques,* Morgan Kaufmann.

WITTEN, I. H., FRANK, E., TRIGG, L. E., HALL, M. A., HOLMES, G. & CUNNINGHAM, S. J. 1999. Weka: Practical machine learning tools and techniques with Java implementations. (Working paper 99/11). Hamilton, New Zealand: University of Waikato, Department of Computer Science. Availabe From:<http://researchcommons.waikato.ac.nz/handle/10289/1040>

XIE, S. & HELFERT, M. 2011. Towards an Information Architecture Oriented Framework for Emergency Response System. *Proceedings of the 8th International ISCRAM Conference,May 2011,Lisbon,Portugal.Avaialbelfrom:*<http://www.iscramlive.org/ISCRAM2011/proceedings/papers/225.pdf.>

XIONG, H., PANDEY, G., STEINBACH, M. & KUMAR, V. 2006. Enhancing data analysis with noise removal. *Knowledge and Data Engineering, IEEE Transactions on,* 18, 304-319.

YANNIS, G., EVANGELIKOS, P. & CHAZIRIS, A. 2009. CADaS–A common road accident data framework in Europe. Presentation to the 4th IRTAD Conference, 16-17 September, 2009 Seoul, Korea. AvailableFrom:<http://www.internationaltransportforum.org/irtadpublic/pdf/seoul/3-Yannis.pdf>

YIN, R. 1994. *Case study research: Design and methods,* Thousand Oaks, CA: Sage Publishing.

YIN, R. 2003. *Case study research: Design and methods,* Thousand Oaks, CA: Sage.

YOUNG, C. 2001. The unexpected case for enterprise IT architectures. *Gartner Group Strategy, Trends & Tactics,* 9.

YUNUS, Y. A. & RAHMAN, A. A. 2008. The Evolution of Information Architecture. *International*

Symposium on Information Technology, ITSim. Vol. 4, pp. 1-6 2008. IEEE.

YUNUS, Y. A., RAHMAN, A. A. & IAHAD, N. A. 2008. Framework for developing Malaysian Natural Product Information Architecture. Procedings of *4th International Conference on Information Technology and Multimedia* UNITEN, 17–19 November 2008, Malaysia. AvailableFrom:<http://web.utm.my/fsksm/content/research/info/doc/papers/Framework_for_developing_Malaysian_Natural_Product_Information_Architecture.doc>

ZACHMAN, J. 2003. Excerpted from The Zachman Framework: A Primer for Enterprise Engineering and Manufacturing (electronic book). Available :< http://www.zachmaninternational.com>

ZACHMAN, J. A. 1987. A framework for information systems architecture. *IBM Systems Journal*, 26, 276-292.

ZACHMAN, J. A. 1999. A framework for information systems architecture. *IBM systems journal*, 38, 454-470.

ZEIN, S. R. & NAVIN, F. P. 2003. Improving traffic safety: A new systems approach. *Transportation Research Record: Journal of the Transportation Research Board*, 1830, 1-9.

ZHENG, Y. 2007. *The impact of localized road accident information on road safety awareness*. Doctoral Dissertation, University of Glasgow.

Appendices

Annex A- Evaluation Result (Descriptive and Reliability)

This annex contains evaluation results and its reliability (descriptive and reliability)

****

DESCRIPTIVES
  VARIABLES=arcmcom arcmofle arcmoapp arcmore arcmocos descomo descost
  descoec descfucl descticl descpcl decomoc decostcl decopecl decofucl
  decoticl decoplcl desymoc desystcl dessypc dessyfcl dessytcl dessypcl
  aparcmoe
  /STATISTICS=MEAN STDDEV .

Descriptives

[DataSet1] C:\Users\Administrator\Desktop\evaluation result.sav

Descriptive Statistics

	N	Mean	Std. Deviation
The architectural model comprehensive	25	4.04	.841
The architectural model is flexible	25	4.08	.702
The architectural model is appropriate	25	4.40	.645
The architectural model is relevant	25	4.64	.638
The architectural model is complete for the scope	25	3.84	.850
The description of the scope/context-motivation is clear	25	4.56	.507
The description of the scope/context-structure/data is clear	25	4.28	.542

The description of the scope/context-people is clear	25	4.20	.866
The description of the scope/context-function/process is clear	25	4.20	.645
The description of the scope/context-time is clear	25	3.80	.816
The description of the scope/context-place is clear	25	4.00	.866
The description of the conceptual motivation is clear	25	4.24	.926
The description of the conceptual-structure/data is clear	25	4.12	.927
The description of the conceptual-people is clear	25	4.04	.935
The description of the conceptual-function/process is clear	25	4.00	.866
The description of the conceptual-time is clear	25	3.72	.936
The description of the conceptual-place is clear	25	3.88	.971
The description of the system-motivation is clear	25	4.48	.510
The description of the system-structure is clear	25	4.40	.707
The description of the system-people is clear	25	4.28	.843
The description of the system-function/process is clear	25	4.28	.458
The description of the system-time is clear	25	4.00	.866
The description of the system-place is clear	25	3.88	.781
The application of this architectural model is easy	25	4.12	.526
Valid N (listwise)	25		

```
RELIABILITY
  /VARIABLES=arcmcom arcmofle arcmoapp arcmore arcmocos descomo descost
  descoec descfucl descticl descpcl decomoc decostcl decopecl decofucl
  decoticl decoplcl desymoc desystcl dessypc dessyfcl dessytcl dessypcl
  aparcmoe
  /SCALE('ALL VARIABLES')  ALL/MODEL=ALPHA.
```

Reliability

[DataSet1] C:\Users\Administrator\Desktop\evaluation result.sav

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	25	100.0
	Excluded	0	.0
	Total	25	100.0

Reliability Statistics

Cronbach's Alpha	N of Items
.940	24

Annex B- Interview questions

Interview questions to collect empirical data on accident information management practices and guide the discussion.

1. How do you explain the road safety situation in Ethiopia?
2. How do you describe the effectiveness of accident data collection and analysis practice?
3. What are the available accident data sources?
4. What aspects of an accident should be recorded?
5. Do you believe that the accident reporting format being used is complete enough in recording all details about an accident? If not what sort of information you think is missing?
6. Can you describe please accident data reporting and analysis process in general?
7. Do you feel that the analysis level is sufficient for the purpose required by the users?
8. How do you deal with accident data? What sort/type of analysis do you make on the accident data at your organization? Are statistical analysis/ machine learning methods involved in your daily work?
9. What aspect of accident information and accident prediction models should get priority in Ethiopian context? (Severity analysis, black spot identification, vulnerable groups' identification, all accident types, injury accidents, crossing accidents, rear-ends accidents, turning accidents, responsibility identification…..)
10. How is accident data updated?
11. What is the prime motivation of managing accident data/information?
12. Who are the primary users of road accident information? What sort of information is required by stakeholders regarding road accident?
13. Who do you think is responsible in accident data reporting?
14. Who do you think should participate in accident data management and analysis?
15. Can you mention a specific place (like organizational units) where data reporting, analysis and dissemination should happen?
16. Would you comment please on the timing for accident data reporting, analysis and dissemination?

Annex C- Road Safety Information Architecture (RSIA) - Evaluation Instruments

This Annex contains a copy of questionnaire items and structured interview questions to evaluate the information architecture.

Part I- Questionnaire items

	Strongly Agree	Agree	No Comment	Disagree	Strongly Disagree
The architectural model is comprehensive					
The architectural model is flexible					
The architectural model is appropriate					
The architectural model is relevant					
The architectural model is complete for the scope					
The description of the scope/context - motivation is clear					
The description of the scope/context – structure/data is clear					
The description of the scope/context – People is clear					
The description of the scope/context – function/process is clear					
The description of the scope/context – time is clear					
The description of the scope/context – place is clear					
The description of the conceptual - motivation is clear					
The description of the conceptual– structure/data is clear					
The description of the conceptual – People is clear					
The description of the conceptual–					

function/process is clear					
The description of the conceptual–time is clear					
The description of the conceptual–place is clear					
The description of the system-motivation is clear					
The description of the system–structure/data is clear					
The description of the system– People is clear					
The description of the system–function/process is clear					
The description of the system– time is clear					
The description of the system– place is clear					
The applicability of this architectural model is easy					

Part II- Structured interview questions

- Q1. What do you understand with the term Information architecture? Does your organization have any?
- Q.2 After viewing the information architecture do you think there are any advantages/ disadvantages to developing and using information architecture to help in the deployment and development of accident data collection and analysis systems?
- Q3 Do you understand the road safety information architecture presented?
- Q4 In your opinion is there any aspects of an information architecture missing?
- Q5 Do you envisage any use or potential use for information architecture with in your organization? If so what are they?

Annex D- Profile of Interview and Focused Group Participants

Participants *(by pseudonym)*	Organization/ Agency	Position	Region	Interview	group discussion
Participant 1	Traffic Office	Accident Data Analyst	Gambela	X	
Participant 2	Transport Authority RS Dep.	Road Safety Expert	Gambela	X	
Participant 3	Traffic Office	Accident Data Analyst	Gambela	X	
Participant 4	National Road Safety Council	Road Safety Expert	Addis Ababa	X	X
Participant 5	Transport Authority RS Dep.	Road Safety Expert	Addis Ababa	X	X
Participant 6	Federal Police	Crime & Accident analyst	Addis Ababa	X	X
Participant 7	National Road Safety Council	Road safety Expert	Addis Ababa	X	X
Participant 8	Traffic Office		Addis Ababa	X	
Participant 9	Transport Authority RS Dep.	Road Safety Director	Addis Ababa	X	
Participant 10	Traffic Office	Accident Data Analyst	Addis Ababa	X	X
Participant 11	Transport Authority RS Dep.	Road Safety Expert	Addis Ababa	X	X
Participant 12	Traffic Office	Traffic Office PR	Addis Ababa	X	
Participant 13	Traffic Office	Accident Data Analyst	Amhara	X	
Participant 14	Traffic Office	Accident Data Analyst	Amhara	X	
Participant 15	Traffic Office	Accident Data Analyst	Amhara	X	
Participant 16	Transport Authority RS Dep.	Road Safety Expert	Amhara	X	
Participant 17	Traffic Office	Accident Data Analyst	Amhara	X	

Annex E- Sample Data Reduction and Analysis Table

Category	Questions	Key words /concepts	Supporting text	Remark
General	How do you explain the road safety situation in Ethiopia?	Problematic Top priority		Observed
	How do describe the effectiveness of accident data collection and analysis practice?	Not effective enacting legal actions on the accused party, disintegrated effort	One respondent commented that: *since traffic office mainly focuses on the responsibility identification, the learning process is somehow disabled*	The observation also confirmed the same Also suggested that road safety departments and traffic offices be organized under same structure as their main agenda is road safety.
Content / What	What are the available accident data sources?	Traffic Police record	'...incompleteness is very serious problem at a national level"	accident report containing only number of accident, accident place, time and date in a memo format in one of the research site is observed
	What aspects of an accident should be recorded?	Nature of an accident, road users, vehicles, road and environment. pre-accident scenarios like 'slight misses' and 'evasive actions' *'Road user's reaction'* after an accident, *'use of seatbelt'*, *'use of stimulant'* like "khat" and *'photograph of drivers'*	*"whenever possible including photograph of the driver and/or involved road users in accident data will improve the completeness of data"*.	
	Do you believe that the accident reporting format being used is complete enough in recording all details about an accident? If not what sort of information you think is missing?	Underreporting Problems on integrity completeness and usefulness of road accident information	*"There are many cases where road users agree not to call a traffic police after an accident, just to avoid procedural steps required in resolving the case"*	-

Annex F- Sample Traffic Accident Statistics Form

ሐ. የትራፊክ አደጋ ስታትስቲክስ

1 በዕለት የደረሱ አደጋዎች

ተ.ቁ	ዕለት	ብዛት
1	ሰኞ	
2	ማክሰኞ	
3	ረቡዕ	
4	ሐሙስ	
5	ዐርብ	
6	ቅዳሜ	
7	እሑድ	
	ድምር	

2 በሰዓት የደረሱ አደጋዎች

ተ.ቁ	ሰዓት	ብዛት	ተ.ቁ	ሰዓት	ብዛት	ተ.ቁ	ሰዓት	ብዛት
1	0100 — 0200		9	0900 — 1000		17	1700 — 1800	
2	0200 — 0300		10	1000 — 1100		18	1800 — 1900	
3	0300 — 0400		11	1100 — 1200		19	1900 — 2000	
4	0400 — 0500		12	1200 — 1300		20	2000 — 2100	
5	0500 — 0600		13	1300 — 1400		21	2100 — 2200	
6	0600 — 0700		14	1400 — 1500		22	2200 — 2300	
7	0700 — 0800		15	1500 — 1600		23	2300 — 2400	
8	0800 — 0900		16	1600 — 1700		24	2400 — 0100	
							ጠ/ድምር	

3 አደጋ የፈጸሙ አሽከርካሪዎች ዕድሜ

ተራ ቁጥር	ዕድሜ	አደጋው ያስከተለው ጉዳት				ድምር
		ሞት	ከባድ የአካል ጉዳት	ቀላል የአካል ጉዳት	ንብረት ጉዳት	
1	ከ18 ዓመት በታች					
2	ከ18—30 ዓመት					
3	ከ31—50 ዓመት					
4	ከ51 ዓመት በላይ					
5	ያልታወቀ					
	ድምር					

4 የአሽከርካሪዎች ፆታ

ተራ ቁ	ፆታ	አደጋው ያስከተለው ጉዳት				ድምር
		ሞት	ከባድ የአካል ጉዳት	ቀላል የአካል ጉዳት	ንብረት ጉዳት	
1	ወንድ					
2	ሴት					
3	ያልታወቀ					
	ድምር					

5 አደጋ የፈጸሙ አሽከርካሪዎች የትምህርት ደረጃ

ተራ ቁ	የትምህርት ደ/ጃ	አደጋው ያስከተለው ጉዳት				ድምር
		ሞት	ከባድ የአካል ጉዳት	ቀላል የአካል ጉዳት	ንብረት ጉዳት	
1	ማይም					
2	መሠረተ ትምህርት					
3	1ኛ — ደ· ት/ቤት					
4	— 2ኛ ደ· ት/ቤት					
5	ከፍ 2ኛ ደ· ት/ቤት					
6	ከከፍ 2ኛ ደ· ት· በላይ					
7	ያልታወቀ					
	ድምር					

6 የአሽከርካሪና የተሽከርካሪ ግንኙነት

ተራ ቁ	ግንኙነት	አደጋው ያስከተለው ጉዳት				ድምር
		ሞት	ከባድ የአካል ጉዳት	ቀላል የአካል ጉዳት	ንብረት ጉዳት	
1	የተሽከ/ባለቤት					
2	ተቀጣሪ					
3	ሌላ					
4	ያልታወቀ					
	ድምር					

7.1 አደጋ የፈጸሙ አሽከርካሪዎች የመንጃ ፈቃድ ደረጃ

	የመንጃ ፈቃድ ደረጃ								
	1ኛ	2ኛ	3ኛ	4ኛ	5ኛ	ሌላ ተንቀሳቃሽ	ፈቃድ የሌለው	ያልታወቀ	ድምር
ብዛት									

7 የአሽከርካሪው የማሽከርከር ልምድ

ተራ ቁ	የማሽከርከር ልምድ	አደጋው ያስከተለው ጉዳት				ድምር
		ሞት	ከባድ የአካል ጉዳት	ቀላል የአካል ጉዳት	ንብረት ጉዳት	
1	መንጃ ፈቃድ የሌለው					
2	ከ1 ዓመት በታች					
3	ከ1—2 ዓመት					
4	ከ2—5 ዓመት					
5	ከ5—10 ዓመት					
6	ከ10 ዓመት በላይ					
7	ያልታወቀ					
	ድምር					

8 የተሽከርካሪው የአገልግሎት ዘመን

ተራ ቁጥር	የአገልግሎት ዘመን	አደጋው ያስከተለው ጉዳት				ድምር
		ሞት	ከባድ የአካል ጉዳት	ቀላል የአካል ጉዳት	ንብረት ጉዳት	
1	እስከ 1 ዓመት					
2	ከ1—2 ዓመት					
3	ከ2—5 ዓመት					
4	ከ5—10 ዓመት					
5	ከ10 ዓመት በላይ					
6	ያልታወቀ					
	ድምር					

Declaration

I hereby declare that the dissertation thesis titled "**Integrated Information Architecture in Support of Road Safety Organizations: The Case of Ethiopia**" is my own work and any additional sources of information have been duly cited. No part of this work has been accepted for an award of any degree or diploma at any institute. Any contribution made to the research by others, with whom I have worked, is explicitly acknowledged in the paper.

Name Tibebe Beshah Signature ______ Date 15/10/2013