HiLCoE Journal of Computer Science and Technology

Volume 2, Number 1, December 2013

HiLCoE

School of Computer Science and Technology

HiLCoE Journal of Computer Science and Technology is published twice a year by HiLCoE. Individual articles can be accessed and downloaded at http://www.hilcoe.com.et.

Address all correspondences to: *HiLCoE Journal of Computer Science and Technology*, P. O. Box 33465, Addis Ababa, Ethiopia; e-mail: hilcoe@ethionet.com.et; Tel. +251 114162121/ +251 114 673585; Fax +251 114162474.

Papers for publication are welcome from researchers in Computer Science and Technology. Papers will be peer-reviewed by at least two reviewers who are active researchers in the respective area of each topic of the paper unlike the first two issues which are neither peer-reviewed nor the responsibility of the affiliated institutions.

Contents

Syllabification Design and TTS System for Afaan Oromo Using Unit Selection

Argaw Korssa
Pact Ethiopia
argaw100@gmail.com

Sebsibe Hailemariam
Department of Computer Science, Addis Ababa University, Ethiopia
sebsibe2004@yahoo.com

Abstract

Though Afaan Oromo is widely spoken in Ethiopia, there is no previous work done that generates quality TTS synthesis using unit selection speech synthesis as well as no syllabification algorithm is developed for it. To address the research problem, review of documents and consulting experts on the linguistic structure of the language were done which are vital to construct new phone set, pronunciation lexica, G2P conversion and syllabification algorithm design. Also recording of speech using praat and C++ code for G2P converter, manual labeling using wavesurf and implementation of syllabification algorithm was used. Furthermore, modules in Festival framework and Festvox where one can build speech synthesis with their own voice are used to develop a TTS prototype.

Finally, the syllabification algorithm is implemented and tested using selected words. The result showed 100% word accuracy rate which depicts the rule-based approach works very well but so far no comparison is made with the other approaches. In TTS synthesizer, new phone set, pronunciation lexica and rules for G2P conversion are constructed. Subjective tests have been carried out for evaluating the synthesized speech quality and the ultimate speech synthesis system got 4.24 MOS.

Keywords: Text-to-speech synthesis; Unit selection speech synthesis; Syllabification; Rule-based techniques; Grapheme-to-phoneme, CV

1. Introduction

TTS synthesis is an automatic conversion of unrestricted natural language sentences into speech [1, 3, 4]. TTS system must be developed for every language to extend the applicability and use of computer technology. It is also arguable that speech synthesis developed for one language is not applicable for others due to differences in language structure, intonation, duration, accent, tone, stress and other contexts. And many of the fully functional TTS products are based on the languages of developed countries. It is with this reality that Morka [6] and Sebsibie [8] tried to develop a prototype TTS for Afaan Oromo using diphone and Amharic language, respectively. However, in the previous work the quality of synthetic speech in Afaan Oromo language was limited and appears now to be insufficient in modern applications.

This study tried to address how to select speech unit and Afaan Oromo phone set for suitable TTS synthesis, how can unit selection be used to represent naturalness and intelligent synthesis and how to design and develop a rule-based syllabification algorithm.

To do so methodologies like data collection that helped to prepare appropriate training and testing dataset (corpus) for the desired system, tools such as speaker for playback record, C++ programming language to develop G2P converter and syllabification rule and praat for manual labeling were used. In addition, a new TTS system with a unit selection compatible voice was built from a new designed and developed large speech corpus within Festival/Festvox framework. Festvox offers general tools for building unit selection voices in new languages.

2. Analysis of Phonology of Afaan Oromo Language

As pointed out by Beekamaa [2] and Thaha [10], the script of Afaan Oromo language is commonly known as phonetic language. In Afaan Oromo there are five short vowels, five long vowels, seven digraph including foreign ones and 28 consonants. With the exception of vowels which stand by their own, all consonants have the combination of both short and long vowels to form a meaning. Afaan Oromo has another glottalized phone that is more unusual, an implosive retroflex stop, "dh" in Oromo orthography, a sound that is like an English "d" produced with the tongue curled back slightly and with the air drawn in so that a glottal stop is heard before the following vowel begins [2, 5].

Afaan Oromo words use both consonantal and vowel roots with vowel variation expressing difference in interpretation. Each syllable pattern comes in different orders, reflecting the 10 vowel sounds with the exclusion of the vowels themselves. There are 28 basic forms, giving 10*28 syllable patterns (syllographs) [2, 5, 10].

There are about 38 phonemes of Oromiffa alphabets (Table 1). These phonemes have been used to contain the acoustics units of the unit selection and speech file which are then used to get the synthesized speech corresponding to the given word. For this study, we have used the eight main templates of syllabic synthesis units (V, VV, VC, CV, CVV, VVC, CVC, and CVVC clusters).

Table 1: Phonetic representation of Afaan Oromo consonants and its IPA Equivalence

Manner		*Bilabial*	*Labiodental*	*Alveolar*	*Palatal*	*Velar*	*Glottal*
Stops	Voiced Voiceless Ejective Implosive	B (p) ph /p□/		D T x /t□/ dh /□/		G K q /k□/	 ' /□/
Fricatives	Voiced Voiceless	(v)	 F	(z) S	(zy) sh /s'/		 H
Affricative	Voiced Voiceless Ejective				J Ch (c') C		
Nasals		M		N	ny /□/		
Lateral				L			
Rhotic				R			
Semi-vowel		W			Y		

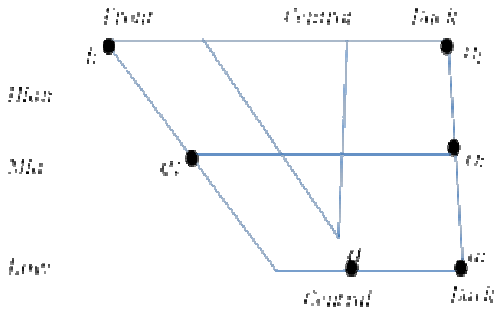

Figure 1: Diagram of Afaan Oromo vowels (Habte and Adugna: 6, 10)

3. System Architecture

3.1 Syllabification Model Design and Architecture

Having germination handling, syllabification and deletion rules and syllable templates of the language, we developed a rule-based syllabification model. Syllabification of any words with the exception of abbreviations and acronyms were considered in our system development. In order to rigger the module, Afaan Oromo text is used. Subsequently

normalization is followed. In the normalization module grammatically wrong characters that appear in the word or text will be deleted using devised rule. Mapping all phonemes and digraphs is carried out, and if digraphs exist re-mapping is carried out. The normalized text then goes to the germination module. Following the output from germination, syllabification is carried out by the syllabifier module. This is done using an algorithm to syllabify any input words in their legal sequence. The final output will be syllable boundary marked Afaan Oromo text as indicated in Figure 2.

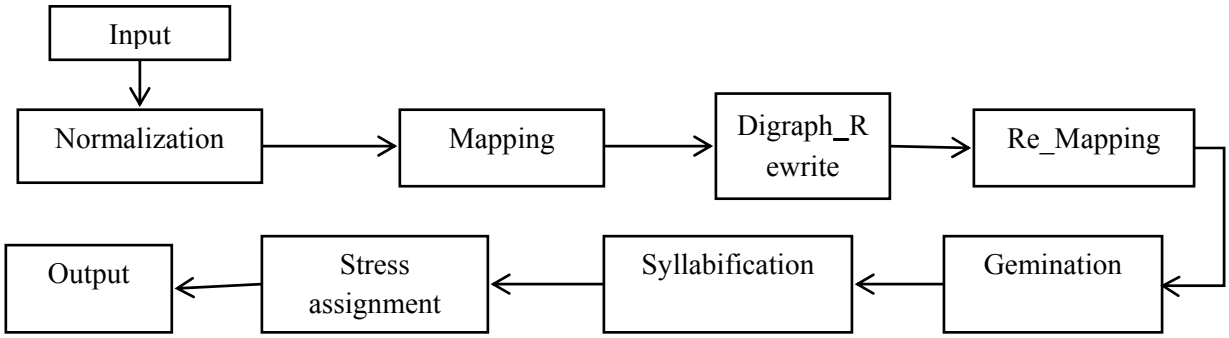

Figure 2: General Automatic Syllabification Model for Afaan Oromo Word

A word may be syllabified in different syllable structure but the algorithm selects the legal syllable structure sequence for the input word. Here, the well-known linguistic syllabification implementation principles namely the maximum-onset principle and sonority hierarchy principle are implemented. The expected output is an identical phonetic representation as input but including modified if incorrect input provided and syllable boundary markers.

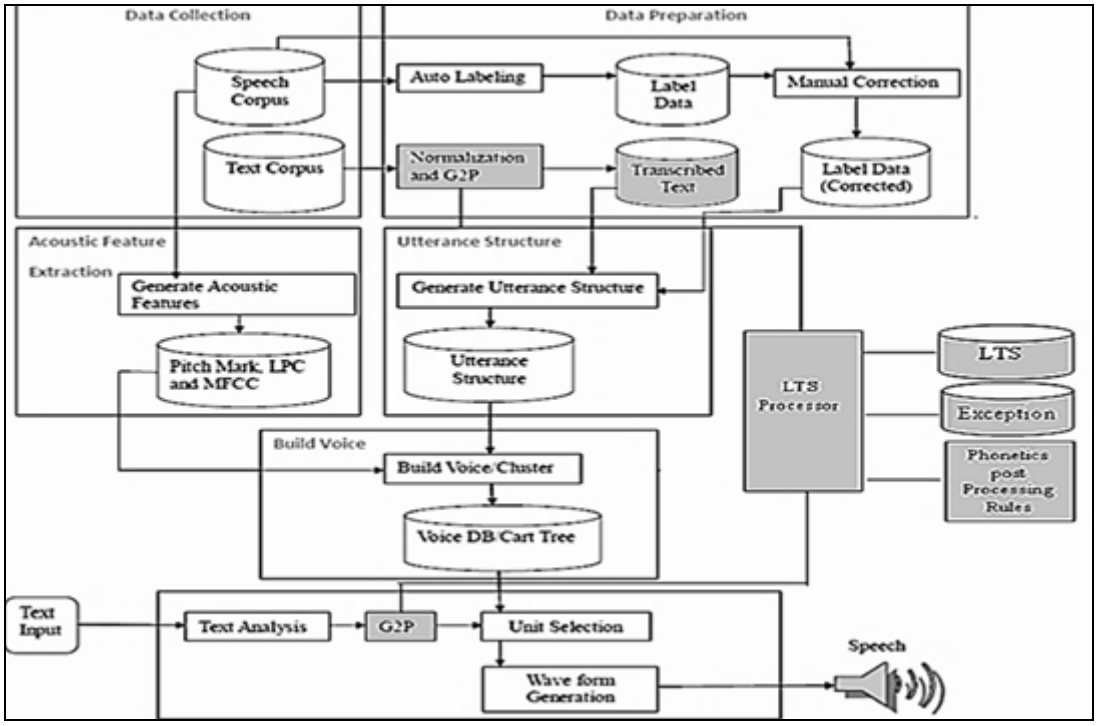

Figure 3: Afaan Oromo TTS System Implementation Model

3.2 Afaan Oromo TTS design and Architecture

Phonemes are the smallest units of speech sound in a language that can serve to distinguish one word from another [9]. Afaan Oromo alphabet has 38 letters classified as 10 vowels and 28 consonants including five adopted foreign phoneme sounds [7, 10]. In our system for phonetic transcriptions we

adopted a phoneme set based on the IPA SAMPA standard.

G2P/LTS conversion, developed using C++ programming language, is best suited for Afaan Oromo words due to close relationship between orthography and phonology of the language. However, some exceptional loanwords might exist. Afaan Oromo pronunciation lexicon is mainly used for determining the pronunciations of these words [10].

4. Implementation

4.1 The Syllable Structure and Syllabification of the Language

All the syllable type of Afaan Oromo can also be defined at the phonological representation. We can classify the eight templates into different classes based on the grammatical structure and/or weight distinction of the syllable templates, for instance, open or closed, long or short, geminated or not, etc. In a simple weight distinction, there are heavy and light syllables in Afaan Oromo. A heavy syllable is a syllable with branching nucleus or a branching rime. A branching nucleus generally means the syllable has a long vowel; this type of syllable is abbreviated as CVV. A syllable with a branching rime is a closed syllable, that is, one with coda; this type of syllable is abbreviated as CVC or CVVC. Light syllable with a short vowel as the nucleus and no coda (CV). A closed syllable is one that ends in a consonant and an open syllable is one that ends in a vowel. Short vowel open syllables are light, all others are heavy. However, heavy syllables attract stress, while light syllables only get stress if they happen to be in the right location in the string of syllables. The following example and Table 2 explain about heavy, light, open and closed syllables.

a. Humnaan hum-naan (By force)

b. Firfirse fir-fir-se (be distributed)

c. Jijijjiiramaa ji-jij-jii-ra-maa (Changeover)

In the words (a, b and c) above, *se, ji, ra, maa* are open syllable whereas *hum, naan, fir and jij* are closed syllable. Syllables such as *naan, jii and maa* are heavy syllable while *fir, se, ji and ra* are light syllables.

Table 2: Summary of Different Kinds of Afaan Oromo Syllable Structure

Kind	*Description*	*Example*
Heavy	Has a branching rhyme. All syllables with a branching nucleus (long vowels) are considered heavy.	CVVC, CVV, VVC CVC
Light	Has a non-branching rhyme (short vowel).	CV
Closed	Ends with a consonant coda.	CVC, VC,
Open	Has no final consonant	CV, V, CVV

4.2 Consonants cluster and their syllable structure

In Afaan Oromo, the maximum number of allowable same or different consonant sequences in a cluster is two. In case if below conditions (a-g) violated, there must be deletion of the phoneme/s. Accordingly, Afaan Oromo word structure allows the following consonant sequence.

a) No two consonant consecutively appear at the beginning or end of the word except ch, dh, ny, ph, ts, sh and zy.
b) No three consonant clusters consecutively appear anywhere in the text.
c) Single quotation (‘) is used to separate between two different vowels and vowels that have different sound and/or length either same or different, e.g., re’ee (goat), bu’aa (result), etc.
d) No three cluster vowels consecutively appear anywhere in a word without separator.
e) No two different vowels come together, if appeared separator is used.
f) No two same short or long sound vowels consecutively appear in a word, if appeared separator is used.
g) Clustering and germination of consonants is possible only word medially

4.3 Epenthesis in Afaan Oromo words

Though the process of epenthesis is not as such significant, we developed a rule for those foreign words. In case there are such words, epenthesis can occur at word-initially and final frequently and word medial sometimes. For instance the English word 'sport' is rewritten as 'ispoortii'. In this letter 'i' is inserted at beginning and end of the word. In the other case, the word 'president' is rewritten as 'piresedaantii' and others. This insertion is to obey the grammatical writing principle of the language unless it has no meaning and difficult to understand.

4.4 Rules and Algorithms

The syllables in Afaan Oromo are formed with the combination of consonants and vowels. Syllables are generally formed from 2-4 characters and contain a vowel and consonants. The most frequently used syllable for words which belong to the Afaan Oromo language are shown in Table 3. To dos so, a minimal set of rules were developed.

Table 3: The structure of Afaan Oromo Syllables-Most Frequently Used Syllables

Syllable structure	*Sample Syllables*
V	Eda (last night) -> e-da. Others start with single a, o, u, i
VV	Aanaa (woreda) -> aa-naa and others
CV	Ba, ci, fi, to… (Bakakkaa-thunder, cimaa-strong, fili-select, etc.)
VC	Ol, ej, ul … (e.g. Ol-up, ejjuu-prostitute, ullaa-, allaattii-birds, etc.)
CVV	Baa, dee, fii… (e.g. baala-leaf, deega-poorness, fiiguu-to run, etc.)
CVC	Bor, gad, bar… (e.g. barruu-inside thumb, barreessuu-write, etc.)
CVVC	Foon, baal, kaan…(e.g. baachuu-carry, foon-meat,)
VVC	Oof, aal… (e.g. aalbee-knife, eengadda-before yesterday, etc.)

In relation with deletion of insupportable phoneme/s, study was not undergone. Thus, from our empirical observation deletion of vowel and consonant has great role in syllabification of Afaan Oromo words. It can occur at word initial, word medial and word final position. Before we design the model of syllabification algorithm we first model the deletion of the inappropriate vowel and consonant in separate module. Both the syllabification and deletion of vowel and consonant algorithm reads input from left-to-right. The language normalization, mapping, germination, deletion and syllabification algorithm identified are presented in the form of a formal algorithm implemented in C++ Programming language.

Deletion of vowel and consonant procedure:

1. Accept input word and scan from left to right.
2. If consonant cluster occurs at word initial or final position, delete the preceding one.
   a) *Exception*: If the first phoneme or consonant is digraph. (Rule #1)
3. If 3 consonants are appeared in sequence word medially, delete the preceding one. (Rule #2)
4. If 3 vowels are appeared in sequence word initial, medially or final, delete the preceding one. (Rule #2)
5. If a cluster of consonants contains the geminate and singleton in sequence, delete after the geminated consonants. (Rule #3)
6. If a cluster of consonants contains the singleton and geminate in sequence, delete one after the geminated consonants. (Rule #3)
7. If a cluster of consonants contains two different geminates in sequence, delete either of the two geminate consonants. (Rule #4)
8. Repeat 2 up to 7 until all the phonemes are parsed in the phonemes list.

The following is the proposed syllabification procedure in Afaan Oromo:

1. Accept the input from normalization algorithm and scan from left to right.
2. At word initial position if one vowel phoneme (V) and (CV) occurs in sequence, mark syllable boundary between first V and C, which results V-CV.

3. At word initial position if two vowels phonemes (VV) occurs in sequence, and the next is (CV), mark syllable boundary between VV-CV but if (VVCCV) mark as VVC-CV.
4. If the initial phoneme is vowel and the next two phonemes are consonant and vowels respectively; mark the syllable boundary just at the second (VC-CV)
5. If (VCCV) pattern occurs at any position, mark syllable boundary between the two consonant clusters.
6. If (VCVC) pattern occurs at word initial position, mark syllable boundary before the second vowel.
7. If (CVV) type sequence occurs at any position, mark syllable boundary at the end of the second vowel.
8. If (CVCCV) phoneme sequence occurs at word initial position mark syllable boundary between the middle consonant clusters (CVC-CV).
9. If (CVVC) pattern occurs at word final position and if there is phoneme before the first consonant mark syllable boundary before the initial consonant in this pattern.
10. If (CVCV) pattern occurs at any position, syllable boundary becomes CV - CV pattern.
11. If (CVCCVC) pattern occurs in a word mark syllable boundary between the geminated/clustered consonants. (CVC- CVC).
12. Repeat 2 up to11 until all phonemes are parsed.

4.5 Voice Building

4.5.1 Creation of Speech Database

To build Afaan Oromo speech database, the text corpus represents different sections of Afaan Oromo texts like political essays, literary works, sports sections, science, newspaper, and business pages in which some exist in hardcopy and others in softcopy. A single male speaker reads carefully the designed script. The selection of the speaker was based on subjective aspects. Signal was recorded using 16 bits and 16 kHz of sampling rate. To this effect, our database can include single or double letter sounds as the smallest phoneme in the current system. The minimum number of phonemes kept in the database is calculated as 598 and the rest of the syllables are formed from the concatenation of these sounds.

4.5.2 Feature Extraction and Clustering

The labeled speech database was processed by applying simple power normalization on each utterance. The maximum and minimum pitch value of the speaker was determined using the Wavesurfer Pitch Marker Tool. The Festvox pitch extraction parameters were adjusted accordingly to obtain pitch features for the utterances. MFCCs were also extracted. The Unit selection was built by applying unit clustering algorithm on the units of the database.

5. Experimental Results and Evaluation

5.1 Performance of Syllabification Algorithm

The test corpus selected and used for syllabification performance measurement and evaluation is 1000 words which satisfies all syllabification rule and regulation of the language. Each syllabified word contains one to seven syllables. The expert compares the input phoneme sequence and the output of the algorithm, and gives their remark on each output. Finally, the expert agreed 100% word accuracy rate.

Table 5: Distribution of syllable patterns over the test set

Syllable Pattern	*Frequencies*			*Total*	*Percentage (%)*
	Word Initial	*Word medial*	*Word final*		
V	70	–	–	70	2.5%
VV	13	–	–	13	0.5%
CV	288	404	392	1084	39.0%
VC	63	–	–	63	2.3%
CVV	167	224	290	681	24.5%
CVC	234	155	48	437	15.7%
CVVC	84	142	193	419	15.1%
VVC	10	–	–	10	0.4%
Total	*929*	*925*	*923*	*2777*	*100%*

5.2 Intelligence and Naturalness of TTS

5.2.1 Evaluation and Methods

To evaluate the quality of the synthetic voice produced by the developed system, we carried out formal listening tests. In such a way, we decided the listeners to rank the voice quality using a Mean Opinion Score (MOS) like scoring as a suitable method for overall evaluation of synthetic speech. The MOSs for speech synthesis are generally given in three categories: intelligibility, naturalness, and pleasantness; and has a five level scale from excellent to bad and it is easy to estimate the listening quality using a 5-point scale: 1-bad, 2-poor, 3-fair, 4-good, and 5-excellent.

In order to validate speech quality, participants that speak different dialects were selected from different parts of Oromia. Accordingly, 12 participants (3 females) were used and they listened the sentences using headphones. The subjects were familiarized with the speech synthesis by listening some example utterances of varying quality.

Of the total recorded and collected corpus, 48 complete sentences in Afaan Oromo had been selected randomly from the text corpus. Each of these sentences contains on average 10 to 32 words. The participants are allowed to listen to the recorded male voice samples before they check the developed TTS. Subsequently, each participant plays the sample voice to check the quality of the voice. However, in order to make the test effort easy and understandable by the participants, a questionnaire was delivered beforehand to familiarize with it. After listening audios, the participants were requested to fill the questionnaire properly.

The means and standard deviations were calculated as per the respondent's opinion. The participants are satisfied with pleasantness and the naturalness of the TTS voices with the average value of the mean being 4.24; and the acceptability of the developed system found to be very encouraging.

6. Conclusion and Recommendation

6.1 Conclusion

In this paper, corpus-based concatenative unit selection speech synthesis system architecture for Afaan Oromoo has been designed and implemented. We have done a literature survey on corpus-based concatenative speech synthesis research. Different techniques have been investigated and applied for optimal speech corpus design.

A unit selection algorithm based on dynamic programming has been implemented. Target and concatenation costs to be used in unit selection have been extracted. A speech representation based on harmonic coding has been implemented. As a result a unit selection based concatenative speech

synthesis system capable of generating highly intelligible and natural synthetic speech for Afaan Oromo has been developed. Subjective tests have been carried out to assess the speech quality generated by the system in terms of intelligence, naturalness and pleasantness. Finally, the system got 4.24 MOS. The variation emerged due to prosody issues, and concatenation of phones which will need further research.

An effort was also made in modeling syllabification algorithm and implementation for Afaan Oromo language. A set of formal rule was defined for the syllabification and epenthesis of its words. The algorithm achieves an overall 100% word accuracy which is well acceptable as rated with the experts. The experts are satisfied with the result. Thus, the rule-based syllabification algorithm developed for Afaan Oromo words is impressive given the linguistic rules and syllabification principles.

6.2 Recommendation

The following are the recommendations to further improve the quality of the Afaan Oromo synthesizer and syllabification:

- Handling prosodic issues (intonation, stress, and duration modeling)
- Developing germination handling algorithm.
- Application of HMM-based speech synthesis method

Similarly, it is also possible to make better and work the syllabification in a different approach:

- Stress and syllabification have their own relationship. Therefore, by studying their relationship thoroughly, we can have stress assignment algorithm.
- A comparison study using data-driven approaches.

References

[1] Samuel Thomas, “Natural Sounding Text-To-Speech Synthesis, based on Syllable-Like Units”, Master of Science, India, May 2007.

[2] Beekamaa Likkaasaa, “Seer Lugaa Afaan Oromoo”, Finfinne, 2004.

[3] Per Olav Heggtveit “An overview of Text-to-Speech Synthesis” Perolavhe Ggtveit, 2003.

[4] Hasim Sak, “A corpus-based concatenative speech synthesis system for Turkish”, 2000.

[5] Habte Bulti, “Analysis of Tone in Oromo”, Addis Ababa University, June 2003.

[6] Morka Mekonen, “TTS Synthesis for Afaan Oromo Language using Diphone Method”, 2001.

[7] Adugna Barkessa, “Concept of Afaan Oromoo Grammar”, Finfine, 2012.

[8] Sebsibe H/Mariam, S P Kishore, Alan W Black, Rohit Kumar, and Rajeev Sangal, “Unit Selection Voice for Amharic using Festvox”, 5th ISCA Speech Synthesis Workshop, Pittsburgh.

[9] Claire-A. Forel and Genoveva Puskás, “Phonetics and Phonology”, Geneva, March 2005.

[10] Thaha M. Roba, “Modern Afaan Oromo Grammer”, USA, 2004.

Architecture Description Language for COTS Integration in Aspect Oriented Software Development

Bereket Tolcha
berekettolcha@gmail.com

Mesfin Kifle
Department of Computer Science, Addis Ababa University, Ethiopia
kiflemestir95@gmail.com

Abstract

Commercial Off-The-Shelf systems or components are being popular in large system developments to save time and money to support reusability. Usually, COTS are assumed thoroughly tested in many situations and environments. They are relatively safe to integrate them to other systems. But the problem is that their integration is not easy since their internal logic is not exposed, they exist as a black box.

Hence, integrating COTS into large systems needs careful design at the architecture stage. Architecture Description Languages come into picture for this purpose. Nevertheless, the existing ADL do not support integration of COTS into Aspect Oriented Software Development approach.

This paper proposes a solution to model interaction of a COTS component with another respective software component. The paper also evaluates the proposed solution by taking two systems and integrating COTS components into them. A prototype is developed to demonstrate the proposed solution in two levels: at design time ADL specification and code level implementation.

The proposed solution uses Aspectual ACME as a base ADL by using its basic elements in representing architecture of a system to design the new artifact because it is an already developed ADL for Aspect oriented software development methodology.

Keywords: Architecture Description Language; Aspect Oriented Software Development; COTS System; Integration

1. Introduction

In Aspect Oriented software development approach, cross cutting concerns are considered as *aspects* and not just objects. Cross cutting concerns are those objects that touch many objects of the system.

There are many efforts in Aspect Oriented Software Development to manage the non-functional requirements and the tangling code problems properly [1]. The main advantage of this approach is implementing the known software development concept “separation of concern”, which dictates everything is modularized based on their concern.

Architectural Description Languages (ADL) is a language that provides primitives to specify components and connectors. A standard notation, in ADL, for representing architectures helps to promote mutual communication, the embodiment of early design decisions, and the creation of a transferable abstraction of a system. ADLs result from a linguistic approach to the formal representation of architectures, and as such they address its shortcomings. Also important, sophisticated ADLs allow for early analysis and feasibility testing of architectural design decisions.

There are different ADLs, in which some are for general purpose and some others for specific purpose. Most ADLs are developed for a specific software development process like OO Software Development, Aspect Oriented Software Development or others.

To represent the architecture of software in common and understandable way, many architects have used Architecture Description Languages (ADL) for many years [2].

The building blocks of an ADL are components, connectors, ports, roles and architectural configurations or properties. Components and connectors may have associated interfaces, types, semantics and constraints, but only explicit component interfaces are a required feature for ADLs. A component's interface is a set of interaction points between itself and the external world. It specifies the services (messages, operations and variables) a component provides and the services it requires of other components. Connectors model interactions among components and specify rules that govern those interactions. A connector's interface specifies the interaction points between the connector and the components and other connectors attached to it. It enables proper connectivity of components by exporting as its interface those services it expects of its attached components. Configurations define architectural structure and how components and connectors are connected [3].

The AspectualACME is a more generic ADL designed for aspect oriented software development by enhancing the traditional ADL called ACME which is a general purpose ADL. It also supports the representation of integration of Commercial Off-the-Shelf (COTS) software using Aspect Oriented development method. But it is more general and it doesn't represent some specific integration concerns.

Currently available AspectualACME represents the integration of third party software like COTS by considering them as another component of the system. But the integration of COTS should be handled carefully and thoroughly because of their black box nature for changeability and maintenance.

Basically acquiring COTS components should ease up the development process of a new software system quite a lot, because the development team doesn't have to "reinvent the wheel" for some particular features anymore. However, it is really rare that a COTS component can be just installed straight away to the existing production environment [4].

For the purposes of this paper, the term COTS means "a software product, supplied by a vendor, that has specific functionality as part of a system - a piece of prebuilt software that is integrated into the system and must be delivered with the system to provide operational functionality or to sustain maintenance efforts and it is a black box" [5].

More and more software projects are using COTS components. Using COTS components brings both advantages and risks. To manage some risks in using COTS components, it is necessary to increase the reusability of the glue-code so that the problematic COTS components can easily be replaced by other components. Aspect-oriented programming (AOP) claims to make it easier to reason about, develop, and maintain certain kinds of application code [1].

With AOP, each aspect can be expressed in a separate and natural form, and can then be automatically combined into a final executable form by an aspect weaver. As a result, a single aspect can contribute to the implementation of a number of procedures, modules, or objects. It therefore helps to increase reusability of the source code. AOP brings several new ideas and definitions

The main benefit of using AOP is that, all the changes are centralized in the aspect without scattering everywhere in the whole system. AOP users should ensure that the cross-cutting concerns in the glue-code are homogeneous [5]. If cross-cutting concerns in COTS glue-code can be separated into aspects, it will be easier to understand and change the system [6].

When integrating COTS components, the internal implementation may cause mismatches between these COTS components. Therefore, glue-code may be needed to integrate these COTS components and make them work together.

Hence, the following are the problems addressed in the paper: how the existing ADL for Aspect Oriented Software Development represents aspectual concerns and its limitations, how the existing ADL for Aspect Oriented Software Development represents integration of COTS, and what should be done to fill

the gap of representing the integration of COTS using an ADL.

The methodology used in this research is mostly exploration and investigation by reading the available papers about the working ADLs and after gathering enough information about the existing ones, the new solution will be designed using the basic notations and semantics used by the traditional ADL called ACME. The designed solution is tested by using two medium sized software using AspectualACME syntax and code level implementation.

2. Related Work

There are two different ADLs reviewed for this paper. One is designed as a new ADL based on traditional ones to represent all concerns of Aspect Oriented Software Development and it is called AC2-ADL [2]. The other one is a traditional ADL which is for general purpose and made some enhancements to make it fit for Aspect Oriented and it is called AspectualACME. The modification is to add a new artifact for representation of cross-cutting concerns of the Aspect Oriented Method [7] and their interaction with other components of the system. Because the second option is closest to the proposed solution, its detail is discussed below.

2.1 AspectualACME and how it works

The basic elements of ACME are not enough to properly modularize and compose crosscutting concerns (or features) with other system concerns. Crosscutting concerns are concerns that get scattered and tangled with other concerns realized by the system components. AspectualACME extends ACME in order to modularly represent crosscutting concerns at the architectural level.

It proposes modeling crosscutting concerns as regular components, and enriching composition mechanisms of ACME to support the definition of crosscutting relations. AspectualACME introduces the Aspectual Connector (AC), a special connector that encapsulates the crosscutting component interactions.

2.2 Aspectual Connector

An aspectual component is a component that represents a cross-cutting concern in a crosscutting interaction. The traditional connector is not enough to model the crosscutting interaction because the way that an aspectual component composes with a regular component is slightly different from the composition between regular components only. A crosscutting concern is represented by providing services of an aspectual component and it can affect both provided and required services of other components which can be, in turn, regarded as structural join points at the architectural level.

Aspectual Connector (AC) is an architectural connection element that is based on the connector element but with a new kind of interface. The purpose of such a new interface is twofold: to make a distinction between the elements playing different roles in a crosscutting interaction, i.e., affected base components and aspectual components; and to capture the way both categories of components are interconnected. The AC interface contains: (i) base roles, (ii) crosscutting roles, and (iii) a glue clause.

The base role may be connected to the port of a component (provided or required) and the crosscutting role may be connected to a port of an aspectual component. The distinction between base and crosscutting roles addresses the constraint typically imposed by many ADLs about the valid configurations between provided and required ports. An aspectual connector must have at least one base role and one crosscutting role. The composition between components and aspectual components is expressed by the glue-clause. The aspectual glue specifies the way an aspectual component affects one or more regular components.

3. The Proposed Solution

Based on the detailed study being conducted on articles related to ACME, Architectural structure is described in ACME with components, connectors, ports, connections and systems. The interaction

between those components will be represented by attachments.

The Aspectual Connector in AspectualACME specializes on the conventional connector abstraction to support the description of interactions among components that have a crosscutting impact and other components.

There is no special way of representing the integration of plug-in software mainly COTS type in the traditional ACME or on the modified version of it called AspectualACME. The integration of COTS is represented in the architecture simply by normal components and connections available in the traditional ACME.

The artefact which will be used to facilitate the integration of COTS component into a system that is developed using Aspect Oriented Software Development method is designed using the basic architectural elements of AspectualACME ADL (connector and role).

The new artifact is designed by making the following considerations:

- It uses the wrapper to abstract the black box behaviour of COTS based on the comparison of the three methods in Table 1. This will make the integration changeable and maintainable because it hides the detail complexities of the COTS components and it is better than the other two options because we can detach it including the COTS component from the main system without affecting it.
- The COTS component will be considered as another component of the system using its wrapper to communicate with other components of the Aspect Oriented system.

So based on these considerations, the new artefact is a connector (named COTS_Connector) with some modifications from the basic connector element of ACME to represent the specific connection characteristics related to COTS integration.

3.1 Architecture of the Solution

The newly designed connector called COTS_Connector will contain two roles. The first role is a normal AspectualACME role in the System component side (R) and the other is from the COTS component side (CR) which will be abstracted through the wrapper role WR.

Table 1: Comparison between three techniques of COTS integration

Method	*Wrapper*	*Glue-ware*	*Proxy*
Hides complexity of the COTS component	Yes	Yes	Yes
Makes the COTS component as a system component	Yes	No	No
Easy to change the COTS component	Yes	No	Yes

Because the wrapper abstracts the existence of the COTS component, it should expose the services offered by the COTS component using COTS Role (CR) through the Wrapper Role (WR). The COTS_Connector will make use of the Wrapper Role to access the COTS_Role to connect to the COTS components services which makes the COTS component abstract so that there is no direct interaction between system components and COTS component. This will make the architecture more flexible for maintenance and change of COTS component for many reasons like if new functionality is required or the previous COTS is updated and needs to be replaced by the updated one.

Table 2: The syntax difference between ACME connector and COTS connector

Regular connector in ACME	*COTS_Connector in AspectualACME*
Connector Connector1 = { Role aRole1; Role aRole2; }	COTSConnector Connector1 = { Role R; Wrapper_RoleWR; COTS_Role CR; }

Table 2 compares the COTS connector with a normal component connector without aspectual concerns. But the COTS connector can also connect an aspectual component with COTS component. In this case the COTS definition will change accordingly to reflect the characteristics of the aspectual component.

Table 3: The syntax difference between AspectualACME connector and COTS connector

Aspectual connector in AspectualACME	*COTS_Connector in AspectualACME*
AspectualConnector Connector1 = { Role AspectualRole1; Role Role1; }	COTSConnector Connector1 = { Role AspectualR1; Wrapper_Role WR1; }

Wrapper_Role WR1 = {

systemRole SR1;

COTS_Role CR1;}

- Where systemRole represents the role of the larger system which integrates COTS.

The basic difference between COTS connector and the other two connectors (normal and Aspectual), as shown in table 3, is that in case of COTS connector it has another additional role for the wrapper to refer to the COTS role and the wrapper role will have its own definition to combine its role with the COTS role. This way of integrating COTS components will guarantee the abstraction of the black box nature of COTS. So there will be a very minimum work required in case of COTS replacement or maintenance.

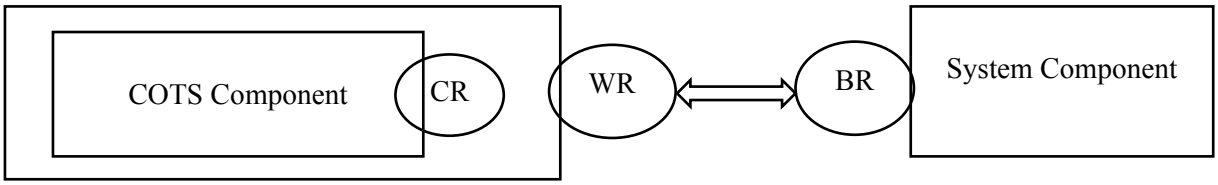

Figure 1: How COTS_Connector connects the system component and the wrapped COTS component

In figure 1, the major characteristic of COTS_Connector in addition to the basic characteristics it inherits from ACME Connector is that the connector has two sides. One points to the system component, it could be aspectual component or normal component, and the other side points to the wrapper component which also connects to the COTS component.

We have now three connectors in AspectualACME including the newly designed one to represent (i) normal connection between components, (ii) to represent aspectual concerns between normal component and aspectual component, and (iii) the new one to represent the connection between normal or aspectual component and the COTS component.

There are some basic differences between these three connectors irrespective of their common characteristics like properties and attachments, as described below:

ACME Connector: is the basic connector defined by ACME and it only connects two normal components without aspectual concerns

AspectualACME Connector: is the modified version of the above connector so it works for normal components and in addition it connects normal components with aspectual components by defining the crosscutting concerns connection

COTS Connector: is also the modified version of the ACME connector so it did work for connecting normal components and in addition it connects normal and aspectual components with COTS component through a wrapper.

To evaluate the designed solution, two systems are considered with two COTS components for each system. Both are used aspect oriented method in the implementation level.

i. Vehicle Management System (VMS) which is Developed for Addis Ababa Transport Bureau

ii. Integrated Market Actors Management Information System (IMAMIS) which is designed for Ethiopian Commodity Exchange Authority

The first system is explained below because it is the one used for prototyping.

Vehicle Management System

The Vehicle Management System is a medium sized software system designed to facilitate all the services given by transport bureau related to vehicles for owners and other stakeholders like Customs Authority to make sure whether the owner paid stamp duty or not and many others. The system is developed using aspect oriented software development method.

From the list of services provided by the system, we will explain temporary registration service and its prototype will be presented next. It contains a list of sub-services to accomplish the complete task. The sub-services or features are:

- Register Vehicle Information
- Register Owner Information
- Collect Service Payment
- Issue Temporary Plate

The COTS have the functionality to Register Vehicle Information sub-service, which has a list of activities in it:

- Connect to customs database based on the given connection information
- Check for vehicle availability based on a given vehicle motor no or chassis no
- Collect and provide vehicle information based on a given vehicle motor no or chassis no

When an owner is registered there are steps to be taken before saving the owner information. The workflow for owner registration is listed below:

1. Check if the vehicle exists
2. Collect the vehicle information from Customs and save it to Transports database
3. Using the registered vehicles code, register owner related to that specific vehicle

The COTS component will also integrate with the aspectual component (security) of the system. The integration syntax could be as follows:

```
COTSConnector Connector1 = {
                    Aspectual_Role
                    AspectualR1;
                    COTS_Wrapper_RoleCWR1
                    ;}
```

In the above syntax, the COTS connector is defined by two roles which are Aspectual_Role to represent the aspectual component's role and the COTS_Wrapper_Role to represent the wrapper role which holds the COTS wrapper inside it.

4. Discussion

The designed solution is evaluated and it is proven that the new artifact will solve or minimize the complexity of integrating COTS components into aspect oriented systems. It is also tested that the new artifact will make the changeability and maintenance of the COTS component because its integration is abstracted by the wrapper component. The integration of the system is being made directly with the wrapper component and not with COTS and the wrapper will communicate with the COTS. So to change or maintain the COTS component it will be just to modify the wrapper class without affecting the actual system.

5. Implementation

To demonstrate the real life applicability of the proposed solution, a prototype is designed and implemented in two scenarios for the Vehicle Management System. The first scenario is using AspectualACME syntax to demonstrate the proposed connector whether it is usable to represent the integration of COTS into an aspect oriented software. The second scenario will demonstrate the code level implementation of the architecture model represented by the ACME ADL syntax in the first scenario. This second scenario integrates two types of COTS into a system to show the flexibility of the proposed solution in changing the COTS. The two COTS types will

have a pre-compiled .dll library and a "web service". Both types will have the same functionality but in different internal ways of implementation.

The implementation of the model will be using Microsoft Visual Studio 2010 and Microsoft SQL Server 2005 products for coding the functions and developing the database backend respectively.

The components to fulfill Vehicle registration in VMS are RegisterVehicle (COTS Component), RegisterOwner (System Component) and the Wrapper component which hides the complexities of the COTS and present available and required features for VMS. The VMS by the help of the COTS will check if there is vehicle information available from the customs authority database and bring the information to the transport authority database if it exists when searching by vehicles Motor Number or Chassis Number which are unique identifiers of any vehicle worldwide. Then VMS will register the owners for the registered vehicle using RegisterOwner component developed in-house.

Through the process of checking information availability and registering of vehicle information is accomplished using the COTS component, the VMS shouldn't communicate to the COTS directly because it will make the integration so tight that any modification, maintenance or replacement of the COTS component will affect the VMS as a whole significantly. So to handle this integration problem, we implement wrapper component which mediates the COTS component and VMS.

To accomplish the above stated task, the wrapper component will have its own configuration document which made the mapping of features between the COTS component and VMS possible. Managing the mapping of features in configuration file will make the integration easy to maintain in case of any maintenance or complete change of the COTS component for any reason. There could be a different means of designing the wrapper component to fulfill the requirements other than using configuration file but in this demonstration the configuration file is chosen. To simulate and demonstrate the applicability of the wrapper component using its configuration file, the scenario described above will be implemented as follows.

Some parts of the configuration file of the wrapper component is presented below.

// here we have application setting required to map the location of the COTS component and different function names provided in the COTS component. The function names are required to create function mapping from the COTS and the complete system without changing the wrapper code in case of any function name or other change occurs in the COTS component. If any such change occurs in the COTS, it will be changed only on the configuration document without affecting the wrapper code.

```
<settingname="COTSLocation"serializeAs="String">
<!--this setting will be used if the above setting
    'COTSType' is set to 'dll'-->
<value>C:\Users\btolcha\Documents\Visual Studio
    2010\Projects\ThesisPrototype\RegisterVehicleInf
    ormation\bin\Debug\RegisterVehicleInformation.d
    ll</value>
<!--this setting will be used if the above setting
    'COTSType' is set to 'webservice'-->
<value>http://localhost:8732/RegisterVehicleService/
    RegisterVehicle.svc</value>
</setting>
<!--the following three settings are used in the
    wrapper to create delegate(representative) function
    based on their name in the COTS-->
<settingname="ConnectToDatabaseWrapper"serialize
    As="String">
<value>ConnectToDatabase</value></setting>
<settingname="IsVehicleInfoAvailableWrapper"serial
    izeAs="String">
<value>IsVehicleInfoAvailable</value></setting>
<settingname="GetVehicleInfoWrapper"serializeAs=
    "String">
    <value>GetVehicleInfo</value></setting>
```

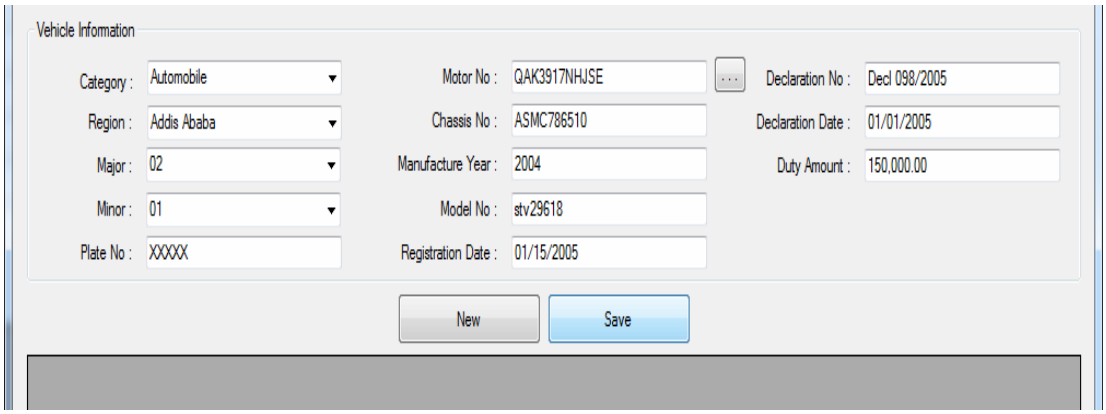

Figure 2: How the window looks when vehicle information is registered after completing vehicle information in addition to the information displayed using search from the customs database

After filling the rest of the information in addition to the information found from customs database including the owner information, the system will look like figure 2.

The new button in Figure 2 is used to search the vehicle from customs database using the COTS function through the use of wrapper components function. After the search is completed if the vehicle is available in the customs database by motor number or chassis number, it will be displayed in the space provided. Like for example, the 'Manufacture Year", "Model No", "Declaration Date" are displayed in the text area. The rest of the information will be filled by the system user and using 'Save' button on the 'vehicle information' section, the user can save the vehicle information into the transport authority database.

6. Contribution and Future Work

The main motivation of this paper is to design a way to integrate COTS components in an efficient and cost effective way in time of development and also in time of maintenance or change of the COTS component. The change may be required for many reasons like new requirement came up from the users or just the already existing COTS don't have the required feature.

The designed artifact is evaluated using two medium sized systems with aspect oriented implementation and it works well in hiding the direct contact between the COTS component and the system component. And also it doesn't require additional syntactic knowledge other than basics of ACME ADL elements.

As a future work, it would make the solution complete if the integration concept is handled for COTS with crosscutting concerns (aspectual components).

References

[1] Axel Anders Kvale, et al., "A Case Study on Building COTS-Based System Using Aspect-Oriented Programming", 2004.

[2] Wen Jing, et al., "AC2-ADL: Architectural Description of Aspect-Oriented Systems", 2009.

[3] Thaís Batista, et al., "Reflections on Architectural Connection: Seven Issues on Aspects and ADLs", 2005.

[4] Heikki Kontio, "Current Trends in Software Industry: COTS Integration", 2008.

[5] M. Morisio1, et al., "Investigating and Improving a COTS-Based Software Development Process", 1999.

[6] E. Kesseler, "Assessing COTS Software in a Certifiable Safety-Critical Domain", Information Systems Journal (18) 2008, pp 299 - 324.

[7] Alessandro Garcia, et al., "On the Modular Representation of Architectural Aspects", 2006.

Redesign the Integration of Movement Control and Maintenance Control Systems of Ethiopian Airline to Improve Reliability

Berhanu Bezualem
Ethiopian Airline Enterprise
berhanubezualem@yahoo.com

Mesfin Kifle
Department of Computer Science, Addis Ababa University, Ethiopia
kiflemestir95@gmail.com

Abstract

In general, as an enterprise grows, its IT challenges grow with it. One of these challenges is the integration of different legacy applications. Modern enterprise environment consists of numerous applications and systems that vary in size, complexity and age. Integrating legacy applications saves time, expense and development cost.

The same is true in the airline industry where the business is more and more dependent on IT technologies. There is a high demand of integrating applications. However, the integration solution is not without a challenge. In Ethiopian Airlines, there are two critical operational systems; movement control and maintenance systems that have to be integrated. The former is a legacy system whereas the latter is a web-based system. To integrate the two systems, interface software was developed in house and deployed for use but was not reliable as required; missing data and frequent failure. This work studied, analysed and recommended an alternative technology for improving the reliability problem of the existing integration software used to integrate these two applications by evaluating the different contemporary and current application integration technologies.

This paper discusses in detail the most common distributed technologies by considering major factors such as interoperability, coupling, serialization, reliability, ease of implementation and other factors deemed necessary. More emphasis was given to Web Service technology.

Having studied these different technologies, a mixed technology of MS MQ and Web service were selected as better solution to meet the business goal which is achieving reliability. Moreover, prototype was developed using the selected technologies to demonstrate the integration. Though WS RM is a best solution to meet this goal it was not selected in this study since the specification is relatively new and not yet implemented and tested for use in the real industry.

Keywords: Coupling; Integration; Interoperability; Message Queue; Operational systems; Reliable Messaging; Web service

1. Introduction

In general as an enterprise grows, its Information Technology (IT) challenges grow with it. One of these challenges is the integration of different applications. Virtually all enterprises have legacy applications and databases. Connecting to the legacy applications saves the time and expense of having to migrate them and provides a mechanism for tying together fragmented IT platforms. The importance of integration was very closely related with the sensitiveness of the data or application [1]. Furthermore, the distributions and analysis techniques that are commonly used are fairly specific to lifetime data. Different integration problems require different solutions, and it is important to use the right one. To address the diversity of applications and data that must be connected, different organizations based on their business targets can produce a range of integration products and technologies.

The airline industry is one of the major transport service providers. Its major business focus is providing air transport services for passengers and cargo. In most cases, it is expected that systems have to talk with each other. In addition, since products

acquired from different software vendors, the integration task mainly falls on the shoulder of the internal IT team of the airlines. The existence of this problem was a reality that was also observed in Ethiopian Airline. Among the applications that have to integrate are the two main operational systems: the Maintenance Control (MaC) and Movement Control (MoC) systems.

Ethiopian Airlines (EA) is one of the best airlines operating in Africa with its hub at Addis Ababa, Ethiopia [2]. EA has a number of systems which are mainly airline specific and other generic systems like System Application Program Enterprise Resource Planning (SAP ERP) to support its business. In addition, it has also in house developed applications mainly for the purpose of integration of various systems acquired from different vendors.

Different airlines use different products and the way they use the products differ according to the business model the airline follows, for example, some MoC products include MaC application as one module; however this module is recommended for airlines that do not have a big maintenance facility [3]. In EA the maintenance facility is too wide so this module does not support the business hence they need a complete package maintenance management solution. The two major operational systems in any airline operations are the Flight Following system also called the MoC and the MaC systems. MoC deals with tracking of the position of all aircrafts of the airline in real time while the major function of the MaC is developing maintenance schedule, producing maintenance tasks, and tracking of aircraft parts wear and tear. The latter system depends on the former system flight data feed like schedule origin or destination cities, schedule departure date and time and so on. These two systems use different technologies to exchange information. MoC provides only strictly flat file for data exchange while MaC avails Web service technology is for data exchange with the other system. It is within this technology constraint that the integration software was developed and deployed for use.

Because the newly developed integration software does not provide complete data feed, daily monitoring and manual intervention is required to maintain the completeness and accuracy of the data mainly for cost and safety reasons. The integration software developed in house by EA to enable operational data exchange between the two mission critical software solutions was not reliable as required. This is because as the daily monitoring of this interface software shows there were significant numbers of messages that were rejected by the other system and in addition the software frequently fails and needs restarting again and again.

Major problem is observed on the amount of messages failed or missed from the total amount of message sent in this mission critical business. The accumulated data shows that there is problem with the current interfacing application of EA.

The points of failure of these messages were Web service, Oracle and mapped drive. These missing data or incompleteness means the accumulated usage value assigned to a given aircraft parts do not show the real situation until it is corrected manually. Here, one can easily understand how much the manual tasks are cumbersome. Furthermore, the integration application has also a problem of non-functional requirement - reliability - which is the probability of the software executing without failure for a specific period of time.

The main task of this research was studying and analyzing in-depth, assess contemporary application integration technology architectures and recommend the best technology by considering major factors that drives the direction of IT. Furthermore, software quality attributes such as performance, reliability and scalability were evaluated in terms of business goals for design decisions.

The following methods were used to meet the objectives of this study. 1) Document review and analysis 2) Experiment with several scenarios to discover how the software behaves in real time situations. 3) Literature review on remote procedure call architectures, web service integration framework and software quality attributes. 4) Prototyping: Since

the main objective of this study was to improve reliability, a prototype was developed and used in order to validate the reliability improvement of the interfacing application by the use of selected integration technology for the business we have at hand.

There were no any modifications done on the two systems of MoC and MaC. In general, the scope of this study was integrating the two systems to improve the reliability problems that the EA was facing nowadays and recommending better integration technologies using different factors.

2. Literature Review

2.1 Major Distributed Computing Technologies Used for Application Integration

As pointed out in [5, 6] Web service, Distributed Component Object Model (DCOM), Java version of Remote Method Invocation (RMI), Common Object Request Broker Architecture (CORBA) and socket based technologies are the major distributed technologies used for application integration. As described in [6], currently Web service technology is the one that the IT industry embraces. The main difference between the Web services approach and traditional approaches (for example, distributed object technologies such as the OMG CORBA or Microsoft Distributed Component Object Model (DCOM) lies in the loose coupling aspects of the architecture. Coupling refers to the act of joining things together, such as the links of a chain: the degree to which software components depend [4]. In a distributed environment, to determine the degree of coupling in a system, one needs to look at different levels.

Instead of building applications that result in tightly integrated collections of components, the whole approach in modern distributed computing technologies (like in Web services) is much more dynamic and very adaptable to changes in general. Another key difference is that through Web services, the IT industry is tackling the problems using technology and specifications that are being developed in an open way.

As elaborated in [5] Web services is a term that can be understood in many different ways - some argue that it means the services built using new XML based standards and services like Simple Object Access Protocol (SOAP) and others consider it to mean a broader communication process or the new Web based service phenomenon. In either way, the idea behind Web services is a clear one. Until now, a majority of Web-based services have been targeted to users accessing them with Web browsers. The next logical steps are services provided by applications for other applications. Web services is an architecture that enables the building of loosely coupled distributed systems using technology based on open standards that do not force to lock-down to a particular programming language, component model or computing platform.

In this paper, it is tried to compare and contrast different competing technologies like CORBA, RMI and Web Services. As stated in [7], for example, RMI has significant features that CORBA doesn't possess - most notably the ability to send new objects across a network, and for foreign virtual machines to seamlessly handle the new objects. As stated in [8] RMI has a lot of potential, because of the flexibility of remote method invocation, it has become an important tool for Java developers when writing distributed systems. CORBA is gaining strong support from developers because of its ease of use, functionality, and portability across languages and platforms [9]. Before completing the comparisons with web services here, let us see it with DCOM too. Table 1 [10] summarizes the fact.

As shown in Table 1 Web services solve many of the problems in other technologies. It can be used to integrate the existing distributed technologies CORBA and DCOM. This role can be argued by considering the following arguments: 1) The nature of Web Services: it is XML, SOAP and HTTP based 2) Web Services support and endorse to the specification 3) Technologies integration, not replacement.

Table 1: Comparisons among CORBA, DCOM and Web Services

Characteristics	*CORBA*	*DCOM*	*Web Services*
Specification	Specification	Implementation/specification	Specification
Platform Support	Platform-Independent	Mainly Windows, but it can support other platforms	Platform independent
Data model	Object Oriented	Object Oriented Model	Message exchange
Client-Server Coupling	Tight-Coupling	Tight-Coupling	Loose -coupling
Language Support	Any language with an IDL binding	Any language with an IDL binding	Any language
Wire Protocol	Internet Inter-ORB Protocol (IIOP)	Object Remote Procedure Call (ORPC)	SOAP and (HTTP)
Internet friendly	No	No	Yes
Firewall traversal	Not firewall friendly	Not firewall friendly	Firewall friendly

2.2 Limitations of Existing Distributed Technologies

In addition to the tight coupling, the problem with the current technologies used in creating distributed software applications is the lack of interoperability. The other problem with applying these technologies is caused by the corporate firewalls: binary serialization unlike XML serialization and most of these technologies are not future proven. The difference between binary and XML serialization is that in XML serialization it is possible to change any common language runtime objects into XML documents or streams and vice versa. The XML serialization enables it to convert XML documents into such a meaningful format that the programming languages can process the converted documents with ease. But binary serialization converts the files into a binary format which is not a human readable format. Enabling communication with Web services, possibly dynamically discovered, requires loose coupling between a requester and the service to be used [11]. Web services bring many benefits to enterprise users who are willing to capitalize on the potential they provide. They can be used to reduce highly the application development complexity and costs by providing standard external interfaces to applications and whole systems [5]. In general, Web services have the following advantages: 1) Loose coupling 2) Interoperability 3) XML serialization (plain text), not blocked by firewall, and 4) Future proof.

2.3 Reliable Messaging Technologies

a. Message Queue

The Message Queue (MQ) technology offers so many advantages than the traditional technologies. In our case we use it because of the following two important benefits and others that really fit to solve the problem that we have at hand: 1) Avoid unnecessary reading and writing (see Figures 1 and 2) Avoid sorting of files: by doing so, we improve reliability.

b. Web Service Reliable Messaging

Web Service Reliable Messaging (WSRM) describes a protocol that facilitates the reliable delivery of messages between two web service endpoints in the presence of component, system or network failure [12]. The specification outlines two distinct roles viz. a source and a sink. WSRM provides support for various delivery modes such as exactly-once and at least once.

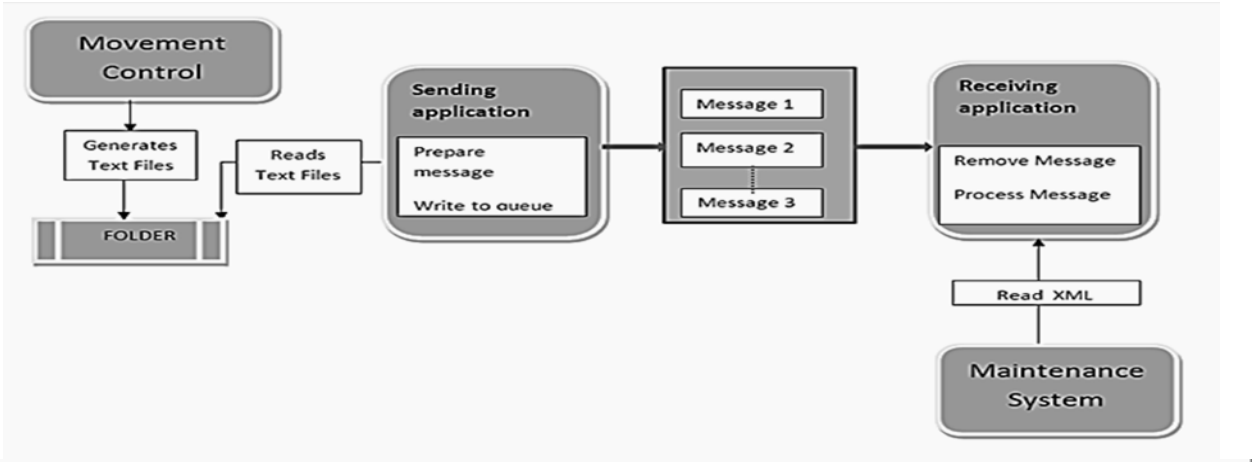

Figure 1: Microsoft Messaging Queue (MSMQ) Model in Web service

The delivery guarantees are valid over a group of messages, which is referred to as a sequence. Messages that are sent may never arrive, arrive too late, arrive multiple times or arrive out of order. After messages arrive, the service may crash and loose some messages. Application code can handle these failures with extra code, sometimes needing the cooperation of the remote endpoints using an enhanced message exchange pattern [13]. However, this approach can significantly increase the investment in the application as well as its time to market. It can also make the application more difficult and time-consuming to adapt to its changing requirements in the future. As an alternative, reliable messaging (RM) can be used as part of overall application failure recovery to improve application programmer productivity and time to market.

In general WS RM has the following major strengths: 1) Perform a task asynchronously, meaning it executes the next task without waiting for a result; 2) Performs in disconnected or connected environment (store and forward protocol) which means we don't need a dedicated connection. This improves the reliability in terms of communication which is the main aim of this study; 3) Message sequences are tracked and maintained; and 4) Interoperable which means it can perform in different platforms. However, RM is not yet fully tested and implemented.

3. Findings of MoC to MaC Applications Integration

3.1 Application Description

The current working system of the business is relying on the two operational systems; the MoC and the MaC systems [2]. MoC is the system where setting up the message format is performed. This is done specifically on Movement Control Administrator (MCA). In the existing business setting we have different message types which are generated from the MoC: a) Planned (NEW) Messages: There are two scenarios for these messages to get generated: when the flight schedule is processed and when there is a Tail change from one aircraft to another or when the tail change is from an aircraft with Tail to with no Tail (dummy Tail). b) Movement (MVT) Messages: These messages get generated when there is actual time insert or change in the Movement Control called OOOI and respectively to mean OUT, OFF, ON, and IN. c) Cancel (CNL) Messages: These messages get generated when the flight is cancelled or deleted in MoC. The second one is the MaC system which uses Web service.

3.2 Extent of the Weakness of the Existing Integration Software

In order to show the extent of the problem in the current integration interface, we try to collect data from system log. Table 2 shows this concisely.

Table 2: Summarized data collected from existing system log file

Date	*Total message Sent*	*Number of Undelivered Messages*	*% Failure Rate*	*Major Failure Reason*
Nov 11	17796	206	1	Web service Transport
Dec 11	20176	553	3	Web service Transport
Jan 12	17937	558	3	Web service Transport
Feb 12	12389	173	1	Web service Transport
Mar 12	14975	435	3	Web service Transport
Apr 12	14706	268	2	Web service Transport
May 12	18199	290	2	Web service Transport
Jun 12	10086	161	2	Oracle Connection
Jul 12	20228	535	3	Web service Transport
Aug 12	49970	575	1	Web service Transport
Sep 12	53342	1157	2	Web service Transport
Oct 12	53373	2853	5	Web service Transport
Nov 12	63446	2054	3	Web service Transport
Dec 12	64236	2418	4	Web service Transport
Jan 13	59948	2788	5	Web service Transport
Feb 13	69669	10509	15	Web service Transport

The data was collected from the log file of the existing integration software and from its technical documents. The software logs errors happening in each day. Moreover direct observation on the software was conducted to observe how the system behaves in a real time. This research was conducted by analyzing the data collected from the existing system log file. This log file was used to assess the negative impact it has indirectly on the business. In general the problem was found to be too difficult and the average 17 months (about 11/2 years) failure rate is 2.5% for this mission critical business which is directly related with safety and cost.

3.3 Findings and Analysis

The root causes of these failures were identified as: 1) Web service data communication, 2) Oracle connection: During the study there were few messages identified left undelivered, 3) Extensive string manipulation to handle message ordering and XML mapping. The other finding is that the software uses extensive string manipulation to sort data coming from source application in a sequential manner, and 4) Extensive data writing or reading tool from the database: This has a direct negative impact for messages not be delivered to the destination. To solve the problem the appropriate technology was identified by the study. The nature of the business requires reliability. According to the opinion of experts in the field, the error tolerance is zero for this business. However, the observed data indicates that the error rate is very high and needs a significant improvement on the existing integration software.

4. Solution Design

4.1 Factors in Integration Technology

The goal of integration here is to connect diverse pieces of software of EA. However, developing a comprehensive integration solution to integrate the applications from the two vendors by providing a unified solution to improve reliability is the main target although giving complete solution which is fully tested using today's latest integration technique like reliable messaging is a challenging one. Currently, the available product mappings for the

integration application and extended enterprise patterns focus on the use of the following technology options (see Table 3 [14]).

Table 3: Technology options across the patterns

Business Drivers	*Direct Connection*	*Broker*	*Serial Process*	*Parallel Process*
XML	✓	✓	✓	✓
Web services	✓	✓	✓	✓
J2EE Connector	✓			
Java Message Service (JMS)	✓	✓	✓	✓
Message-oriented Middleware		✓	✓	✓
Flow languages			✓	✓

Different data types are used to solve according to the problem at hand. Because this study was primarily analysis on the effect of lack of appropriate applications integration, selecting the right application integration technology and data was done in order to improve reliability. The data source for this research was from the movement control system of EA in the planned or NEW Messages.

To send a message to a queue requires only a few steps: First, obtain a reference to the appropriate message queue (using the queue's path, format name, or label), and then use the Message Queue object's Send method. Generally, that is all about specifying: 1) a queue by its path, 2) a queue by format name, and 3) a queue by label. In general, there are three main tasks that have to be done in order to implement MQ. These are: Creating Queues, Deleting Queues, and Sending Messages. For a message to be sent from application A to application B, there are two main tasks or procedures. Figure 2 illustrates message communication between two applications using the window service of MQ. The two main procedures are: 1) The Web Server of window service receives the messages from application A and makes queues (see Figure 2), and 2) We wrote a script called XML mapping that converts the text message into XML schema so that application B's Web service.

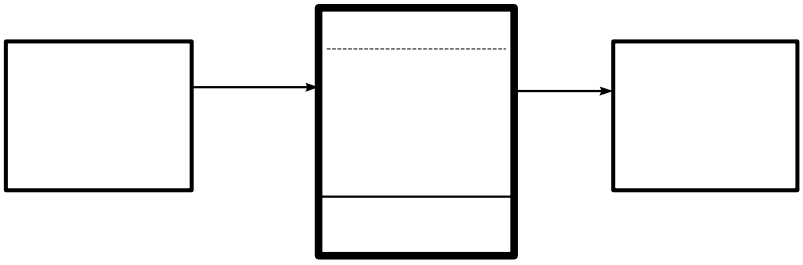

Figure 2: Communication of two Applications Using the Window Service of the MS MQ

4.2 Evaluation

The prototype was designed and developed based on the business decisions. Accordingly, the mixed technologies MQ and Web services were used to implement the prototype. This software was tested based on a copy of 24 hours actual data. The general observation was that: 1) It successfully works as message exchange software, 2) It stores the messages sent by application A (the source) until application B (destination) is up and running, 3) The software works without failure for 24 hours and this is the indication that the software can manage the processing of a day volume intensive task without any problem. This was achieved mainly unlike the existing software which uses a database extensive write or read manipulation for handling data exchange management, the

prototype software uses the MQ built in functionality of data exchange. This is also an additional contribution for achieving reliability.

5. Conclusion

The current working system of the airline business operation is relying on the two operational systems. In EA's case the MoC system uses Web service implemented in the MaC system to interact synchronously using Web service messaging technology. There are a lot of points to consider while deciding on the integration technology between applications like the current infrastructure, time to market, future expansion plans, reliability and transaction support.

In this paper, we tried to asses different software integration technologies. Accordingly, there are two broad categories of integration technologies: the traditional and current ones. From the current technology, we selected the two technologies: Web service and MQ. Using Web service and MQ for implementation, a prototype was developed using the script written in VB.Net in order to demonstrate the concept in integration. Moreover, the role of the database management technology was removed and the mapping of the message format from text to XML schema was managed without the need of the data management system.

In the future: 1) We recommend integration technology tasks should consider WS RM as a better integration technology option because it has a number of features which address the shortcomings of the current integration technologies and 2) Thoroughly assessing the complex features of MQ and Web Services as integration technologies also has to be conducted.

References

[1] Microsoft Corporation, Understanding Microsoft Integration Technologies, September, 2005, http://technet.microsoft.com/en-us/library.

[2] Ethiopian Airline Business Process, Movement Control and Maintenance system, November, 2012

[3] Sabre, Sabre Maintenance Control System, Texas USA, 2007.

[4] Sanjiva Weerawarana, Francisco Curbera, Frank Leymann, Tony Storey, and Donald F. Ferguson, "Web Services Platform Architecture", Crawfordsville, Prentice Hall PTR, 2005.

[5] Ron Schmelzer, Travis Vandersypen, Jason Bloomberg, Madhu Siddalingaiah, Michael D. Qualls, Sam Hunting, Chád Darby, David Houlding, and Dianne Kennedy, "XML and Web Services Unleashed", Sams Publishing, 2002.

[6] Timo Ahokas, "Using XML in Making Legacy Application Data Accessible Through Web Services", Helsinki University of Technology, Department of Computer Science and Engineering, Laboratory of Telecommunications Software and Multimedia, 2003.

[7] An Introduction to Complex Event Processing in Distributed Enterprise Systems, 2000, http://www.informit.com/articles/article.aspx

[8] Reilly, D., "Mobile Agents - Process migration and its implications", November 1998, http://www.davidreilly.com/topics/software_agents/mobile_agents

[9] Reilly, D., "Introduction to Remote Method Invocation", October 1998, http://www.davidreilly.com/jcb/articles/javarmi/javarmi

[10] Morgan, B., "CORBA meets Java", Java World, October 1998, http://www.javaworld.com/jw-10-1997/jw-10-corbajava.

[11] Gunjan Samtani, "Using Web Services for Application Integration", 2002.

[12] Shridee Pallickara, Geoffrey Fox, Beytullah Yildiz, Sangmi Lee Pallickara, Sima Patel, and Damodar Yemme, "An Analysis of the Costs for Reliable Messaging in Web or Grid Service Environments", Indiana University.

[13] George Copeland, "Web Service Reliable Message Programming", April 2008.

[14] Technology options for Application Integration and Extended Enterprise patterns. IBM Corporation, 2004.

Meningitis Symptoms Extraction from Published Conference Research Projects and Journals

Binyam Seyoum
binseyoum@gmail.com

Tibebe Beshah
School of Information Science, Addis Ababa University, Ethiopia
tibebe.beshah@gmail.com

Abstract

Meningitis is a potentially life-threatening infection of the meninges, the tough layer of tissue that surrounds the brain and the spinal cord. According to WHO statistics, every year bacterial meningitis epidemics affect more than 40 million people living in 21 countries of the "African meningitis belt" (from Senegal to Ethiopia). In this area over 800,000 cases were reported in the last years (1996–2010). Treating a meningitis disease is not hard if the symptoms are identified well, but identification of symptoms of meningitis diseases is a hard job when the data to be extracted is from unstructured source (e.g., PDFs, text files and word documents) mining of symptoms becomes tiresome.

This paper shows how content can be extracted from unstructured data, e.g., a word document or portable document format. Using PubMed and Google Scholar as a data source and pre-processing (data cleaning) techniques gained from information retrieval discipline combined with ontologies to extract content automatically from published conference research projects and journals.

The results of this work can highly benefit the domain experts in biological fields in identification of symptoms of meningitis which in turn can help in combating and eradicating meningitis from Ethiopia as stated by the Millennium Goals.

Keywords: Meningitis; Text Mining; Ontologies

1. Introduction

The proliferation of large amounts of data available on the Web, on corporate Intranets, on news wires and on Life-science journals is overwhelming. Life-science journal publishing has undergone a digital revolution in the last decade. These life-science publications embody a store of knowledge and information of interactions and relations among biological entities, which is very important for the understanding of biological processes. With biomedical literature increasing at a rate of several thousand research projects per week, it is impossible to keep abreast of all developments, although the amount of data available is constantly increasing, ability to absorb and process this information remains constant. Search engines only exacerbate the problem by making more and more documents available in a matter of a few key strokes. Text mining (TM), Natural language processing (NLP) and Information Extraction (IE) have shown the most promising techniques in making biological literature more accessible and easy to retrieve information and associations from thousands of documents [5].

Text mining, NLP and Information Extraction as a new and exciting research area [1], try to solve the information overload problem by using techniques from data mining, machine learning, information retrieval (IR) and knowledge management. This shared basic techniques involve the pre-processing of document collections (text categorization, information extraction, term extraction), the storage of the intermediate representations, the techniques to analyse these intermediate representations (such as distribution analysis, clustering, trend analysis, and association rules), and visualization of the results.

On the other hand according to WHO [1], Ethiopia is at higher rate of outbreak of meningitis. Because of this there is an active research in Alert Hospital to eradicate meningitis from Ethiopia. Every year, bacterial meningitis epidemics affect more than 400 million people living in 21 countries of the "African meningitis belt" (from Senegal to Ethiopia). In this area over 800 000 cases were reported in the last years (1996–2010). Of these cases, 10% resulted in deaths, with another 10–20% developing neurological squeal. During the 2010 epidemic season (weeks 1–26) 22,831 cases were recorded in 14 countries under enhanced surveillance. Among these 22,831 cases there were 2,415 deaths [2, 3]. The most affected countries in the region are Burkina Faso, Chad, Ethiopia, and Niger. Burkina Faso, Ethiopia, and Niger were accountable for 65% of all cases in Africa. In major epidemics, the attack rate range is 100 to 800 people per 100,000. However, communities can have attack rates as high as 1,000 per 100,000. During these epidemics, young children have the highest attack rates. The meningitis in these regions has caused many deaths every year at an estimated economic cost of huge amounts of money [4].

Consequently, this paper is aimed to help biological researchers in Ethiopia by giving the tool, to quickly and efficiently to find and extract information (the symptoms of meningitis) from published conference research projects and journals. For this purpose the tools and techniques of text mining and information extractions have been used.

The problem with unstructured documents is even worse in biological literatures, if we look at Medline database [7], which maintains the abstracts of research projects in the field of biomedical research, had a growth of 500,000 new research projects in 2004 per year. In 2010, this has become two research projects per minute [7]. This availability of huge textual resources provides the scientist with the chance to search for correlations or associations such as protein–protein interactions, gene–disease associations, disease symptom association, disease-cause association and new findings about the research area [8]. Nevertheless these huge number of documents with more and more information in them are kept untouched because there is hardly a tool that can mine the knowledge that is hidden, many researchers currently are extracting information manually and their discovery is as good as their processing power.

Currently an active research is being held on meningitis symptoms at Alert Hospital in Addis Ababa, as WHO states that Ethiopia is currently at the pick of meningitis outbreak [1] and researchers at Alert Hospital are doing a research on the symptoms of meningitis.

The problem with unstructured documents is also faced by Alert Hospital researchers. Using techniques and tools of text mining, this paper will contribute its share by extracting symptoms of meningitis from published conference research projects and journals. Text mining is the process of discovering new, previously unknown information, by automatically extracting information from a usually large amount of different unstructured textual resources by computer

The general objective of this paper is to identify the symptoms of meningitis disease from meningitis literatures, using the basic techniques and algorithms of information retrieval combined with ontologies and to produce a prototype.

2. Related Work

The volume of published biomedical research, and therefore the underlying biomedical knowledge base, is expanding at an increasing rate. While scientific information in general has been growing exponentially for several centuries, the absolute numbers specific to modern medicine are very impressive. The MEDLINE 2004 database contains over 12.5 million records and the database is currently growing at the rate of 500,000 new citations each year. With such explosive growth, it is extremely challenging to keep up to date with all of

the new discoveries and theories even within one's own field of biomedical research.

Starting with a collection of documents, a text mining tool would retrieve a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted. Three text analysis techniques are presented: collection, preprocessing and extraction or analyzing, but many other combinations of techniques could be used depending on the goals. The resulting information can be placed in a management information.

Many attempts have been made to get the most out of the increasing biological literatures and different scholars have been proposing and implementing their work on their respective field. One pioneer work is the work of Don Swanson. This paper uses the preprocessing of documents the same as Swanson since some books place the works of Swanson as of concept extraction or text mining and some put it as different technologies [9].

The work of Swanson is categorized in the Concept Linkage, because in his work he actually found a cure. Swanson pioneered the research of knowledge discovery from text by exploring the benefits of inferring associations in a series of experiments using simple semi-automated methods to aid human discovery. Titles from MEDLINE were used to make connections between seemingly dissociated arguments: the connection between migraine and magnesium deficiency, which has been subsequently validated experimentally; between indomethacin and Alzheimer's disease, and between Curcuma longa and retinal diseases, Crohn's disease and disorders related to the spinal cord.

Swanson's work is different from the work in this paper due to 1) event though both works use basic preprocessing techniques the category is different; the work in this paper is more of information extraction but Swanson's work is Concept Linkage. 2) There is help of visualization in the work in this paper which holds visualization of the final work. 3) The tools, procedures and algorithms are different. This work is inclined to information extraction, since only symptoms are extracted from the collected documents.

Another work in the area of text mining is the work of Mathiak and Eckstein [9]. They stated that their work of text mining has five parts: text gathering, text preprocessing, data analysis, visualization, and evaluation. The aim was to analyse the different methods applicable to the five steps and to add their own results if possible [9], and to present the most feasible way of text mining. It presents a framework from collecting the texts to visualizing the result. In the procedures used and the overall steps involved, their aim was to see different methods and check the applicability of the methods to the five steps.

3. The Proposed Solution

Data collection is the first step in information extraction, by choosing Google scholar and PubMed as source and tool. The reason behind is accessing a journal or conference proceeding from Google Scholar is quite different from PubMed.

First, a keyword was selected that can represent the theme of the documents that was needed for the research project and following were selected "meningitis", "meningitis symptoms", "new findings of meningitis" and "meningitis 2013". The keyword "meningitis" is believed to represent the documents that are related to meningitis and "New findings of meningitis" is selected because in the new findings of meningitis the symptoms are also included. Since we are looking for new findings regarding symptoms, the keyword "meningitis symptoms" is self-explanatory. As it can be seen from Table 1, it is the central theme of the documents that are used since 40% of the documents are gathered form this keyword search. With "meningitis 2013" some results are displayed on both search engines. This term is suitable for the desired search because symptoms up to 2012 are already known and what is

needed is the 2013 findings on meningitis focusing on symptoms.

Table 1: Search results from Google scholar and PubMed

Keyword	*Website*	*Results*
Meningitis	PubMed	61,475
	Google Scholar	672,000
New findings of Meningitis	PubMed	2,687
	Google Scholar	252,000
Meningitis Symptoms	PubMed	32,770
	Google Scholar	283,000
Meningitis 2013	PubMed	1,710
	Google Scholar	131,000
Total		*1,436,642*

Using the above keywords in PubMed, PubMed central (PubMed for only full articles) and Google scholar the following result was obtained.

As can be seen from Table 1, Google Scholar presents a large amount of result when compared to PubMed with the same keyword but this doesn't mean that all the results are relevant since some junk results are also included. Therefore, the next step was to filter the above results to more relevant documents by limiting the number of search pages and restricting results from only certain trusted websites and also can provide freely for review the document. Some trusted websites for biological research were added after searching the web like CDC, WHO, and the Red Cross foundation. When using this websites for the given search dramatically, the Google search results were decreased as shown in Table 2.

Table 2: Google Scholar result using specific websites

Keyword	*Result*
Meningitis	15,000
New findings of Meningitis	8,000
Meningitis Symptoms	15,600
Meningitis 2013	10,000
Total	*48,600*

Finally results that are presented from the first to the third pages are considered since unrelated contents are usually displayed after the third page.

At last 6,553 papers that qualify the above filter methods were downloaded both from Google Scholar and PubMed. During the gathering process from Google Scholar and PubMed 2,210 articles have been removed due to redundancy from both sources. A total of 4,343 papers were collected and made available for the next step of preprocessing.

To successfully employ text mining on PDF encoded paper, it would be advantageous to start with customizing the conversion process in order to be able to optimize on all levels. A java API was implemented for the conversation process to meet the requirement. A free API provided by Apache called PDFBox is used to convert the PDFs. Using Apache PDFBox API, the collected 4,343 pdf files where converted to text file.

Text Tokenization is the processing of breaking or chopping a continuous character stream into meaningful constituents called tokens. In text mining specifically in term extraction, the presence of words/terms and their statistical distribution play a significant role rather than the sequence of the terms, this is called the bag-of-words approach. In bag-of-word approach to make a statistics of words, tokenization is applied.

After tokenization the tokens (words) were about 4,893,520. These tokenized words have too much redundancy either by having the same word in different form as in the past and the future or having words that are irrelevant called stop words. In order to decrease the dimensionality of the words, terms that are grammatically close to each other (like "cell" and "cells") are mapped to one term via word stemming. Some authors like [6] put stemming after tokenization in order to decrease the dimensionality of the tokenized words.

Stemming is used to map a word to its root word. Porter's algorithm is implemented for stemming since it is a well-developed and maintained stemming

algorithm for English language. A free Java API by Porter is used to stem the tokenized words.

The tokenized words were successfully stemmed modifying Porter's algorithm. During stemming about 600,000 words have been stemmed and has decreased the dimensionality of the words from 4,893,520 to 4,293,520. But in a bag-of-words approach tokenization and stemming are not enough. Still there are some stop words so a stop word dictionary is used which is available in the national center for text mining (NCTM) and is believed to be a standard dictionary for stop words. However the dictionary has its downside; it doesn't include stop words for biological documents About 15 stop words have been added using word count frequency. The list and their frequency is shown in Table 3.

Table 3: Biological Added stop words

No.	*Word*	*Count/frequency*
1	Cell	3451
2	Tissue	3008
3	DNA	2972
4	RNA	2900
5	Portion	2900
6	mRNA	2680
7	rRNA	2500
8	Mitotic	2322
9	Cyclin	2010
10	Yeast	1800
11	Genome	1525
12	receptor	1010
13	Gene	896

After selecting the stop words list, a simple Java class was built that can match between words that are presented as stop words and words that are stemmed. The implemented program matches the words and removes similar words. In this process almost a third of the stemmed words have been removed due to stop word and redundancy.

In all from the initial of 4,893,520 word preprocessing (stemming and stop word removal) greatly reduced the dimensionality by 46% and the cleaned words are 2,512,300. Then comes term extraction. Term extraction first removes white spaces and commas and put the words in a collected form. When the documents are closely seen only symptoms and other relevant words remain in the list.

The ultimate goal of this paper is to extract the symptoms from literature. This is where ontologies came in play. Ontologies as a conceptual framework possess different concepts in a tree structure, and these concepts are expressed using a term. If ontology of symptoms is used all symptom terms are presented on a given clinical symptoms ontology. So that ontology of clinical symptoms from open biological ontology (obo) which is a free and trusted ontological provider is downloaded and then the file is exported to plain text to resolve the issue of filtering only symptoms from the pre-processed files by matching and cross referencing with the ontology just like stop word removal. The overall task at this point is the symptom terms that are filtered using ontologies are the symptoms of meningitis and can be presented as one, but further filtering is required since this work is intended to present the new symptoms rather than presenting all (old and new) symptoms together.

The filtering of terms using ontology to get a collection of words that are used to describe symptoms gives excellent result. But further analysis is required since the end users need the recent symptom not a collection of symptoms.

Phrase is used for an expression that consists of one or more words. Sometimes symptoms can be phrases that constitute more than one word, for example, "stiff neck" is a symptom that consists of two words. This work uses bag of words approach. In this approach the position of words is ignored and focuses on the word level techniques. The problem is if a symptom is a phrase and if tokenization is applied the meaning is lost. Therefore a mechanism should be implemented to incorporate phrases.

A couple of techniques have been implemented to deal with phrases.

To lower the ontology to the word level or actual clinical names, e.g., hair loss will be mapped to alopecia (correct clinical term for hair loss). There are many phrases that fall in to this category. Out of 767 types of clinical symptoms 278 phrases can be mapped to the original medical term.

After conversation and before tokenizing the collected documents other phrase symptoms were collected that much the ontology phrases. These are around 62 out of 489 symptoms. If tokenized the relation between the two words will be lost and never discovers the symptoms.

Using the above two methods the symptoms that are phrases and words of symptoms were extracted. Still there are some issues that are not covered by the above two techniques, e.g., if papers use nonscientific terms to describe the symptoms, the matching will be ineffective but this is a rare case since the publications are for the scientific community.

4. Discussion

The objective of Information Extraction (IE) is recognizing and extracting certain types of information from unstructured or semi structured documents. Ontology based information extraction is more precise in extracting because the machine is not extracting blindly rather by using definition of all the words provided by the ontology. Information extraction with the guide of ontology has three parts:

- Process natural language text documents.
- Present the output using ontologies.
- The information extraction process is guided by the ontology to extract things such as classes, properties, instances and terms.

In this paper, Java open source tools and preprocessing technique is used to gain information retrieval and also build a system more like a prototype that can take input of texts and convert, tokenize, stem, remove stop words and cross reference with the ontology provided and then display the result. The program has successfully extracted information from nearly 4,343 conference and journal research papers. However these results are very much dependent on the ontology provided. The ontology used is from open biological ontology (OBO). Therefore the result is expected to incorporate all clinical symptoms.

The evaluation for this work is done in two ways. First is checking the relevance of the tools and techniques. Then we have to check if the results are as expected. Choosing and using the right tools and techniques can lead to the right outcomes. The goal of this work is to identify the symptoms of meningitis from meningitis literature using content mining techniques and algorithms and to develop a prototype. Alert Hospital as a testing environment for this work conformed the results as valid. The results will soon be considered as known symptoms when another vaccine is issued for meningitis. The symptoms that were discovered manually were 13 in number and with this work they were 10. Two symptoms were displaced when dealing with phrases and it was a tolerable error as a starting work.

The prototype is developed using Java and coded using NetBeans IDE. The prototype can potentially convert the given pdf files, preprocess (tokenizing, stemming, stop word removal), load ontology in text file to match and display the result of the finding. The User Interface consists of file buttons to browse PDF files, to pre-process the documents, load ontology and load known and an extract button to load known symptoms and extract.

Symptoms of meningitis are essential in the domain of biology, medical research and patient diagnosis. Knowing the symptom of a disease helps researchers in understanding the underlining principles of specific diseases and this powerful knowledge can help in developing a cure or vaccine for the specific disease or even for further research on combating the existing one.

The prototype, the underneath principles and the results were explained to the domain experts in Alert Hospital, then they were free to test the prototype with their own data. More than 95% of the results

matched with the experts test. The 5% error was caused by the data pre-processing step which skips some redundant words.

5. Conclusion and Future Work

This research attempted to show the possible application of term extraction using IR data pre-processing steps and ontology to increase the precision of the terms (symptoms) extraction. The data collection followed by file conversion from PDF to text file for more flexible preprocessing was then cleaned using the IR data preprocessing steps that include tokenization, stop word removal and then after term extraction using ontologies stemming.

The data collection and preparation were major tasks due to the uncleanness of the data collected from PubMed and Google Scholar. This is also due to higher volume or size of the data in the database. After keywords were selected by consulting domain experts the two websites were queried for any relevant documents. The search results were numerous and some unwanted junk were in it. Therefore some cross referencing and trusted website source filtering were applied to limit the search result. After successfully filtering and acquiring the desired documents, the documents were converted to text file to successfully preprocess using open Java API called PDFBox.

Here the data is ready for preprocessing so, tokenization, stop word removal and stemming were applied to clean the data and make available for the general objective which is finding the symptoms of meningitis from published research papers. Since the data is clean and only relevant terms are there in the documents, but clinical symptoms are hard to find in the mixed words. Therefore, using clinical symptoms ontology, all technical terms of symptoms were extracted from the cleaned documents.

The work of this research can benefit researchers in Alert Hospital on finding the current symptoms of meningitis. Also scientist who investigate the treatment of any disease will use it for reference to cross check effect of meningitis for their patients.

References

[1] Pubmed, NCBI, http://www.ncbi.nlm.nih.gov/pubmed, Last Accessed on 18 September 2013.

[2] "wikipedia.org," 12 September 2013, Available at http://en.wikipedia.org/wiki/Meningitis.

[3] "webmd," 12 September 2013, Available at http://children.webmd.com/vaccines/tc/meningitis-topic-overview.

[4] "Mafricar," menafricar.org, 15 September 2013, Available at http://www.menafricar.org/meningitis-and-africa, Last accessed on 17 September 2013.

[5] "WHO," WHO, 16 September 2013, Available at http://www.who.int/mediacentre/factsheets/fs141/en/, Last accessed on 18 September 2013.

[6] "chealth.canoe.ca," canoe, 12 September 2013, Available at tp://chealth.canoe.ca/, Last accessed on 15 September 2013.

[7] R. Feldman, "The Text Mining Handbook," Israel: Bar-Ilan University, Israel, 2007.

[8] WHO, Control of epidemic meningococcal disease, Vol. 3, No. 3, pp. 70-80, 1998.

[9] Wiki, "http://en.wikipedia.org/wiki/Google_Scholar", Available at http://en.wikipedia.org/wiki/Google_Scholar.

Mining Patterns from Investment Data: DSS in Support of EIA's Investment Policy Advocacy Roles

Biruk Worku Kote
birukworku13@gmail.com

Tibebe Beshah
School of Information Science, Addis Ababa University, Ethiopia
tibebe.beshah@gmail.com

Abstract

Discovering patterns which can assist policy advocacy tasks of investment promotion experts at the Ethiopian Investment Agency (EIA) is found to be vital. In this research, Class Association Rule (CAR) modeling technique was used in pattern discovery. Inspired by the concept of constraint based rule post-processing concept, this paper also discovered a mining model evaluation and rule reduction technique that is biased towards the task of investment policy advocacy. The evaluation and rule reduction technique has three rule post processing constraints called "consistency", "inclusiveness", and "diverseness".

The research result showed that in terms of "consistency", the Apriori generated mining models and in terms of "inclusiveness" and "diverseness" the Predictive Apriori models were found better. Therefore, due to their bias to the task of investment policy advocacy, the final model was chosen among the three Predictive Apriori mining models; and based on it, a Decision Support System (DSS) that assists the policy advocacy tasks undertaken by investment experts at the EIA was prepared.

Keywords: FDI; EIA; DSS; CAR; Inclusiveness; Consistency; Diverseness

1. Introduction

Investment Promotion Agencies (IPAs), like the Ethiopian Investment Agency (EIA), usually collect large volume of data from their international customers in the form of "required information". International experiences suggest four main roles for an IPA. However, IPAs are advised to take their policy advocacy role as the most valued one than the rest three. A World Bank survey in [2] discovered that the strength of an IPA's investment policy advocacy role within government has been more critical to winning investment than the rest three IPA roles.

If the investment climate is sound, investors will come and the need for promotion, servicing and targeting are reduced. However, good investment policy advocacy requires good treatment & management of Data, Information, System and Knowledge (DISK) [3]. As public sector, IPAs are in a position to identify problems in the investment environment through their working relationships with international and domestic companies. This means, they may act like the chief advocates which can influence the decisions of investment policy makers and by doing so, IPAs can save a great deal of Foreign Direct Investment (FDI) to their countries [4].

A country may invite investment into one or more of its economic sectors. An economic sector into which the country invites investors is commonly referred to as "economic sector of investment". In the case of Ethiopia, such economic sectors are classified into three as: Primary (Agriculture), Secondary (Manufacturing), and Tertiary (Service). Economic sectors of investment are encapsulation mechanisms for industries. Under each economic sector of investment, there can be a number of industries which are also referred to as "investment sectors", which in turn are further classified into sub-sectors. The choice of investment sector (specific industry of investment) by a foreign-country-experienced

investor can be made only after the selection of economic sector of investment is carried out.

In [5], it states that; not only countries, but also locations within countries are selected for investment on the basis of various criteria. In [2], it is explained that the consideration of which economic sectors are attracting for investment, how many jobs are created or how much technology transferred can assist in judging whether investment incentives and tax holidays are providing appropriate value for money.

The hypothesis of this research is: "By applying data mining techniques on investment dataset, as collected form the EIA, it is possible to explore novel patterns and prepare a prototype to a predictive DSS that can assist the EIA's investment policy advocacy roles."

2. Related Work

These issues of "spillover effects of FDI and foreign-country-experience of investors to an investment hosting state" have been controversial among scholars. After literature review on the relationship between FDI & economic growth, Wan [6] found that some literatures depict a positive while others depict the existence of a negative relationship between FDI & economic growth. The author finally concluded that there are conflicting predictions concerning the growth effects of FDI in the existing literatures.

In general, three views have been reflected in the literatures as: 1) FDI & foreign country experience of investors is something beneficial for the general societal and institutional development of the hosting state through spillovers. 2) It is a "no gain" experience, and non-spillover. Rather it contradicts with environmental wellbeing and negative impacts on the domestic firms. 3) Contrary to the above two, however, other scholars refer FDI and its spillover as beneficial to a hosting state but as far as it is treated with care and supported with intensive studies and research.

A good work by Gorge and Greenway [7] provides a comprehensive evaluation of the empirical evidence on productivity, wages and exports spillovers in developing, developed, and transitional economies. Mello [8] also investigated that; while the motivation and strategy of the investing Multi National Enterprise (MNE) is of importance, so is the scope for efficiency spillovers to domestic firms which depend also on the capabilities of the local firms to absorb the new knowledge. Blonigen [9] made empirical examination of literatures mostly about list of exogenous factors which determine FDI decision making by Multi-National Enterprises (MNEs). In [10], the application of data mining by the research endeavors in the investment domain is generally advised. However, empirical data mining research work which specifically raises the issue of investment policy advocacy from pattern discovery point of view could not be found. Looking for a pattern discovery and data mining solution to investment policy advocacy problems, this research can, therefore, be considered a ground work for future researches of its kind.

3. The Proposed Solution

3.1 Preprocessing

The EIA dataset contains 20+ years (Jan. 1992- Present date) of investment data; having 61, 059 individual investment records. For recency and data reduction purposes, this research is based on the last five years (Jan. 2008- Dec. 2012), which contains 35, 366 individual investment records only. All records contain equal number of data explaining attributes, that is 13; and all except one, are categorical. Each record represents an individual investment project. However, not all the 13 attributes were found important in explaining the investment business. Therefore, only the useful ones, which are 4, are considered in this study and the rest 9 are dropped out through the data cleaning process. The 4 business explaining attributes are: Type_of _Investor (Domestic, Wholly_Foreign, Joint_with_Domestic, Public, Ethiopian_by_Birth, Foreign_but_Domestic); Investment_Type (Domestic, Foreign, Public);

Form_ of_Company (Sole, PLC, Share, Public); Economic_Sector (Primary, Secondary, Tertiary).

Each record in the EIA dataset is an investor's record who registered to invest either in local or in foreign currency (FDI) in Ethiopia. However, as this research addresses the investment policy advocacy problem, only from the FDI and foreign-country-experienced investors' side, all records with values of "Domestic" and "Public" were wiped out as they are not FDI. Finally, the count of all investment records which explain the investment business with the foreign-country-experienced investors in Ethiopia reduced to 5,635.

Another pre-processing task was the task of additional attribute derivation. A study of this kind, which is related to foreign-country-experiences of investors investing in a host country, can be studied better if it considers the international investment agreements between the countries where the investors came from (i.e., the investors' countries of experiences) and the host state (in the case of this research, it is Ethiopia). Such agreements include international trade relationships, international economic agreements, and investment agreements concluded between the investor's countries of experience and the host state.

Ethiopia has entered into three international investment agreements and treaties; the Bilateral Investment Treaties (BITs), the Double Taxation Treaties (DTTs), and the Economic Partnership Agreements (EPAs), with a number of countries of the world. Based on this fact, and in order to reflect the effects of international agreements into the research; three new attributes; "Is_BIT", "Is_DTT", and "Is_EPA", were derived. The attributes were derived in order to check whether or not an investor is from a country with which Ethiopia has entered into one ore more of the three international agreements. The three newly created attributes are: Is_BIT (BIT, Non_BIT); Is_DTT (DTT, Non_DTT), and Is_EPA (EPA, Non_EPA).

3.2 Experimental Settings

The purpose of the data mining phase, in this research, is developing mining models through application of various experimentations. The final cleaned investment dataset with 5,635 investment records was divided into two as: training-set (75% * 5,635 = 4,226 records) and test-set (25% * 5,635 = 1,408 records). Despite the difference in data size, all experimentation setups applied on both sets were exactly the same. Three separate CAR experimentation setups were applied to both sets.

Due to lack of data mining experience and absence of a data mining research in the domain area, the domain experts at the EIA could not be able to provide minimum support (Min_sup), minimum confidence (Min_conf), and predictive accuracy (n) values for the research. This problem has called for statistical derivation of rule interestingness measures, from the dataset itself. Therefore, in this research, three combinations of rule interestingness measures were derived from the dataset, using statistical techniques and by avoiding statistical biases to the extent possible (see Table 1).

Table 1: Statistically derived and selected rule interestingness measures

No.	*Min_sup*	*Min_conf*	*Predictive accuracy(n)*	
			For training set experiment	*For test set experiment*
1	4.42%	12.88%	488	483
2	7.45%	21.87%	265	272
3	3.69%	26.6%	405	468

As it can be seen in Table 1, the values of Predictive accuracy (n) used for training set have variations with that of the test set. The reason is, as it is explained in [11], in Apriori, Min-sup and Min-conf values are required percentage values on an unseen data whereas in Predictive Apriori, support & confidence are combined and merged in to a single

measure called "predictive accuracy". This measure is expressed in the form of the number of rules required (n).

In this research, Predictive accuracy values were set to be equal to the count of rules as a result of a corresponding experiment using the Apriori algorithm. For example, in the first Apriori experiment, Min_sup=4.42% and Min_conf=12.88% were applied on the training dataset. The resulting count of all CARs generated was 488. Here, the number 488 is a Predictive accuracy value and used in the corresponding Predictive Apriori experiment on the same training dataset (the arrows in Table 2 & 3 represent this). This was done from the intention of generating any possible number of best Predictive Apriori rules up to the number of rules actually found as a result of a previous, corresponding, Apriori experiment on the same dataset. Therefore, while the Min_sup & Min_conf percentage values can be same for both the training and test sets, the values of predictive accuracy can vary between the two sets.

3.3 Experimentation and discussion of results

Once the rule interestingness measures were prepared, as discussed above, six sets of experiments (three experiments using the Apriori and another three experiments using the Predictive Apriori algorithm) were prepared from each of the training and the test sets. This resulted in the creation of a total of twelve CAR mining models (six training and six test models). Prepared by applying either of the Apriori and the Predictive Apriori algorithm on the two datasets, each training model was set to have a corresponding test model which was later used as an evaluation control model. The CARs were generated from the three experiments on the training dataset, using Apriori and Predictive Apriori algorithms. Experiments made on the training dataset are named as training experiments (or training models) whereas those made on the test dataset are called test experiments (or test models). The result of the six training model creation experiments is presented in Table 2 whereas that of the six test models is presented in Table 3.

Table 2: Results from the mining experiments on the training dataset

No.	*Experimentation setup*		*CARs*	*Rules distribution among class values*			*Range of support & confidences in the rules (%)*		
	Algorithm	*Min_sup & Min_conf*		*Primary*	*Secondary*	*Tertiary*	*Primary*	*Secondary*	*Tertiary*
1	Apriori	4.42% & 12.88%	488	101 (21%)	152 (31%)	235 (48%)	15-34	21-41	38-80
2	Predictive Apriori	n = 488	346	107 (31%)	136 (39%)	103 (30%)	35-72	13-82	27-94
3	Apriori	7.45% & 21.87%	265	33 (13%)	72 (27%)	160 (60%)	22-29	22-39	38-80
4	Predictive Apriori	n = 265	264	59 (22%)	102 (39%)	103 (39%)	25-72	25-82	27-94
5	Apriori	3.69% & 26.60%	405	20 (5%)	128 (32%)	257 (63%)	27-38	27-41	35-80
6	Predictive Apriori	n = 405	346	107 (31%)	136 (39%)	103 (30%)	35-72	13-82	27-94

Table 3: Summary of results from the mining experiments on the test dataset

No.	*Experiment setup*		*CARs*	*Rules distribution among class values*			*Range of support & confidences in the rules (%)*		
	Algorithm	*Min_sup & Min_conf*		*Primary*	*Secondary*	*Tertiary*	*Primary*	*Secondary*	*Tertiary*
1	Apriori	4.42% & 12.88%	483	88 (18%)	150 (31%)	245 (51%)	38-14	56-21	91-36
2	Predictive Apriori	n = 483	329	98 (30%)	119 (36%)	112(34%)	86-2	91-7	99-28
3	Apriori	7.45% & 21.87%	272	34 (13%)	72 (26%)	166(61%)	33-22	46-22	91-36
4	Predictive Apriori	n = 272	271	55 (20%)	104 (39%)	112(41%)	86-25	91-25	99-28
5	Apriori	3.69% & 26.60%	468	57 (12%)	149 (32%)	262(56%)	38-27	56-27	91-30
6	Predictive Apriori	n = 468	329	98 (30%)	119 (36%)	112(34%)	86-2	91-7	99-28

As it is presented in Table 2, in all the three Apriori generated training models, there is a trend that the largest number of CARs are generated for the "Tertiary" sector whereas the smallest are for the "Primary" ones. Contrary to this, in two out of the three training experiments with Predictive Apriori, the largest number of rules are generated to the "Secondary" sector whereas the smallest are to the "Primary" ones.

The largest number of most dependable CARs (in terms of confidence level), for all the three Apriori generated training models, are associated with the "Tertiary" sector whereas the least dependable ones belong to the "Primary" one. The same is true with the rules from the Predictive Apriori generated training models (in terms of predictive accuracy).

Whenever predictive accuracy (n) increases, in the case of the three Predictive Apriori generated training models, the number of CARs are found either increasing or remained unchanged. When "n" increases from 265 to 405, the number of CARs also increases from 264 to 346. However, when "n" is raised from 405 to 488, the number of CARs remained unchanged (=346). This means all Predictive Apriori generated CARs from training experiment #4 (with n = 265) are subsets of the rules from training experiments #2 & 6 (each with n = 346). However, a kind of trend is not noticed with the Apriori generated CARs. Since the Apriori generated models are determined by a combination of two factors, that is Min_sup & Min_conf; rising the Min_sup level form 3.69% to 4.42% and then to 7.45% or the Min_conf level from 12.88% to 21.87% and then to 26.60% individually, doesn't merely make one model to be a subset of the other. Rather, the three Apriori models could only have some rules in common.

The number of Predictive Apriori generated CARs, under all the three training experiments are less than that of the Apriori ones. In one of the training experiments, however, this difference becomes very insignificant. In training experiment #3, the Apriori generates 265 rules whereas in the training experiment #4, Predictive Apriori produces 264 CARs after generating with predictive accuracy of 265 CARs. Even if the number of rules generated using the two algorithms under training experiments #3 & 4 is almost equal, the two have only 9 CARs in common.

All the three training experiments which were made using the Apriori algorithm have their 9 best rules in common. On the other hand, all the three Predictive Apriori training experiments have their 264 rules in common. Furthermore, the 9 rules which are intersections to the three Apriori generated training models are also members of the 264 Predictive Apriori generated rules.

3.4 Model evaluation and selection

The six predictive models, which were built on training data set, were evaluated for predictive strength. Ultimately and based on a chosen model, a prototype to a DSS that can assist policy advocacy tasks of investment promotion experts at the EIA was developed. The goal of this evaluation step was to select one dependable CAR mining model which is a concise representative of the EIA dataset, a complete and lossless one as much as possible, and quality in terms of its rule strength.

The methodology followed for evaluation is by using the test experiments as controls which can help in testing and evaluating their corresponding training models (experiments) through comparison of the results between the two.

Throughout this evaluation, the research uses the terms "best rules", "best CARs" & "best replicas" and they all represent one same thing, that is, a set of N CARs, where $\{N| N > 0\}$, form a training or a test model, the measure of rule strength of which (i.e., confidence level/Predictive Accuracy) is greater than or equal to 70%. The 70% accuracy level is chosen intentionally based on the concept of constraint based association rule filtering explained in [1].

One constraint driven pattern discovery technique discussed in [1] is rule post-processing after the actual data mining task is completed. That is, the filtering out of any unwanted patterns by placing some constraint conditions suitable for the course of the business so that coming up with the required type of rules.

In this research, the final model to be selected was required to be a set of CARs which are considerably enough to be taken as supportive for the task of investment policy advocacy. For that, the rules must have good quality, both in terms of the numeric rule strength measures (i.e., confidence level/predictive accuracy) and other qualitative measures like being non-confusing in its prediction. In most literatures, association rule experiments are carried out for rule strength higher or equal to 50%. Therefore, in this final model selection process, 70% is a reasonably good constraint to be considered as a minimum numeric rule strength measure which helps for rule post-processing.

This research uses a new mechanism for rule evaluation and best model selection by defining three evaluation constraints: "consistency property", "inclusiveness property", and "diverseness property".

Consistency property of a model refers to a mining training model's ability to replicate its own best CARs on its counter test model.

Assume set T is a predictive training model built on the training dataset and S is its counter test model built on the test dataset. Also assume that all the experimentation setup applied on both S & T are the same. If C, represented as {C| C C L and C ≠ Ø }, is a non-empty subset of L where L is the set of all "best class association rules" of T, and L' is the set of all "best class association rules" of S, then the relation R from T to S is said to be consistent iff C is the subset of L' and is represented as R = { T→S| C C L and C C L'}. The consistency property represents stability so that dependability of the rule regardless of a change in dataset.

Inclusiveness property of a model refers to a mining training model's ability to include best CARs of other models which are developed on the same training set as its own members.

Assume Q & J are two predictive training models. If C, represented as {C| C C L and C ≠ Ø}, is a non-empty subset of L where L is the set of all "best class association rules" of Q and L' is the set of all "best class association rules" of J, then the relation Z from Q to J is said to be inclusive iff C is the subset of L' and is represented as Z = {Q→J| C C L and C C L'}. The inclusiveness property represents the lossless behavior of a model. If a model can make other model as its subsets, the model losses nothing. Furthermore, it represents the other models as well as it may include better rules than what is included in its subset models.

Diverseness property of a model refers to the property of a model to incorporate best CARs which have rare items at their consequent side, where a 'rare item' refers to either of the two rare values, "Primary" and "Secondary", of the class attribute "Economic_Sector".

Assume N is a predictive training model and M is its counter test model. If A is the set of rare items, represented as A = {"Primary", "Secondary"} and Y is an element of A, represented as {Y| Y E A}, and if C, which is represented as {C| C C L and C ≠ Ø}, is a non-empty subset of L where L is the set of all "best class association rules" of N having Y at their consequent side and L' is the set of all "best class association rules" of M having Y at their consequent side, then a relation V from N to M is said to be diverse iff C is the subset of L' and is represented as V = {N→M| C C L and C C L'}.The diverseness property represents the completeness and usability of a model. In some situations, a best model may not be the one that is strong in predicting only one common thing and ignores other predictable aspects of the problem at hand. For that matter, a common thing does not require a prediction. Rather, a complete model which can predict every predictable aspect of the problem to some reasonably acceptable degree of

accuracy can be considered a better model. In this regard, the diverseness property can be a key property in addressing "the rare item problem" as it has the capability to select models which support CARs that point to rare consequents.

The result of the comparative evaluation of the consistency and diverseness of the six predictive training models, in comparison to their counter test models, is depicted in Tables 4 and 5.

Table 4: Apriori models for rule consistency & diverseness

Model No.	*Apriori generated training & test models*					*Rule consistency and diverseness property*
	No. of CARs generated from the training dataset (A)	*No. of all best CARs replicated on the counter test model (B)*	*Proportion all "best replicas" (B / A)*100*	*No. of rare CARs replicated(C)*	*Proportion of rare CARs replicated (C /B)*100*	
1	488	55	11%	0	0%	A consistent, non diverse model.
3	265	46	17%	0	0%	A consistent, non diverse model.
5	405	64	16%	0	0%	A consistent, non diverse model.

Table 5: Predictive Apriori models for rule consistency & diverseness

Model No.	*Predictive Apriori generated models*					*Rule consistency and diverseness property*
	No. of CARs generated from the training dataset (A)	*No. of all best CARs replicated on the counter test model (B)*	*Proportion all "best replicas" (B / A)*100*	*No. of rare CARs replicated(C)*	*Proportion of rare CARs replicated (C /B)*100*	
2	329	2	0.6%	0	0%	A consistent, non diverse model.
4	271	16	6%	5	31%	A consistent and diverse model. 5 out of 16 rules (i.e., 31% of replicated CARs) are belong to sectors other than the default.
6	329	2	0.6%	0	0%	A consistent, non diverse model.

As it can be seen from Tables 4 and 5, all the six rules are consistent, but to different degrees. However, the only diverse model found is the second Predictive Apriori generated model. The other five models were found to be non-diverse. Concerning the five none diverse models, all of their replicated rules belong to the default "Tertiary" sector. On the other hand, out of the 16 "best replicas" of the only diverse model, the 5 (i.e., 31%) are found to be rare rules.

In general, the Predictive Apriori generated models were found to be best in terms of "diversity" and "inclusiveness" whereas in terms of "consistency", the Apriori generated models were found best. The interpretation of this is that the Apriori generated models are stable. However, their prediction stability is up to predicting about the default "Tertiary" sector. Generally, they can dependably predict which combination of an investor's attributes can bring to know the investor's tendency to choose the default "Tertiary" sector. However, these models are dull about prediction towards the rare sectors.

The Predictive Apriori models on the other hand are better predictors of investors' choice about rare sectors. Furthermore, they are good representatives of other models generated by a lesser volume of dataset, that means they are lossless. However, especially towards the default sector, the Predictive Apriori models are not as stable predictors as the Apriori models.

The first thing that requires a decision here is therefore; a choice between "diversity +

inclusiveness" versus "consistency", that means, choosing between a Predictive Apriori or an Apriori model. However, this decision by itself is up to answering the question "which kind of model is considered best for the task of investment policy advocacy?" and answering the question in return needs referring a good definition of knowledge discovery data mining in [12] as a non trivial extraction of implicit, previously unknown, and potentially useful information from data.

A predictive model for policy advocacy shall be consistent, all round, as well as strong. Even if a compromise is required in any of these, the model shall be complete. This is because, advocators must feel confident that what they are lobbying about the country's investment climate brings at least something better, and by no means worse, than how they are currently doing it. Furthermore, they must feel that the model can tell them something novel, which they cannot address intuitively and through their day-to-day observations. Finally, they need a complete model which has the ability to predict "everything", where "everything" means all predictable trend given the combination of facts. Otherwise, the usability of the model falls in doubt. Therefore, the final model is generally better to be more diverse and more inclusive about the novel trends, and less accurate towards "predicting" a trivial information which can be reached upon by default. This can generally be achieved when the model is selected among the Predictive Apriori generated ones.

The second decision is easier. Once which type of model to select is known, which model to select is a matter of comparing among the models in that selected class. Therefore, among the three Predictive Apriori generated predictive models, the forth model (which is the one generated with n = 272) is best in terms of "consistency" and "diverseness" whereas the other two are equally best in terms of "inclusiveness". However, even if the two are best in terms of inclusiveness, their inclusiveness is a complete one. This means they include all the 271 rules of model #4 as their subsets. As these models are Predictive Apriori generated models and inclusiveness is tested on the bed of the same dataset, all the best rules of model #4 are also among the best rules of the other two models. Even if the two include additional rules over what model #4 (which is their subset and intersection) does include, the difference is a set of weakest rules.

Having these logical grounds, Predictive Apriori mode #4 (with n = 272) was found better and selected as the final model of this research. Finally, selecting all the "best rules", the predictive accuracies of which are >= 70%, out of the chosen model, a prototype to a DSS which is capable of predicting foreign-country-experienced investors' choice of economic sector of investment was developed.

1. `Invr.Txt="Foreign but domestic" && DttTxt="NON_DTT IS_EPA=EPA" ==> Economic_Sector=Tertiary 62 acc:(0.94766)`
2. `Invr.Txt=Foreign but domestic IS_EPA=EPA 68 ==> Economic_Sector=Tertiary 64 acc:(0.93632)`
3. `Form_of_Company=Sole Invr.Txt=Foreign but domestic IS_BIT=BIT IS_DTT=NON_DTT 68 ==> Economic_Sector=Tertiary 62 acc:(0.8941)`
4. `Form_of_Company=Sole Invr.Txt=Foreign but domestic IS_DTT=NON_DTT 100 ==> Economic_Sector=Tertiary 90 acc:(0.87941)`
5. `Form_of_Company=Sole Invr.Txt=Foreign but domestic 112 ==> Economic_Sector=Tertiary 98 acc:(0.85223)`

3.5 Model Deployment

Once the models were developed, evaluated, and a final working model is selected, the best 25 CARs were taken out of the selected final model and a prototype to a predictive DSS was prepared. As the system requirements are identified as a result of the research output, the object oriented software analysis, design, and development approach was followed to analyze the requirements, to design them and finally to develop it. Figure 1 shows the GUI for the EIA sector predictor DSS.

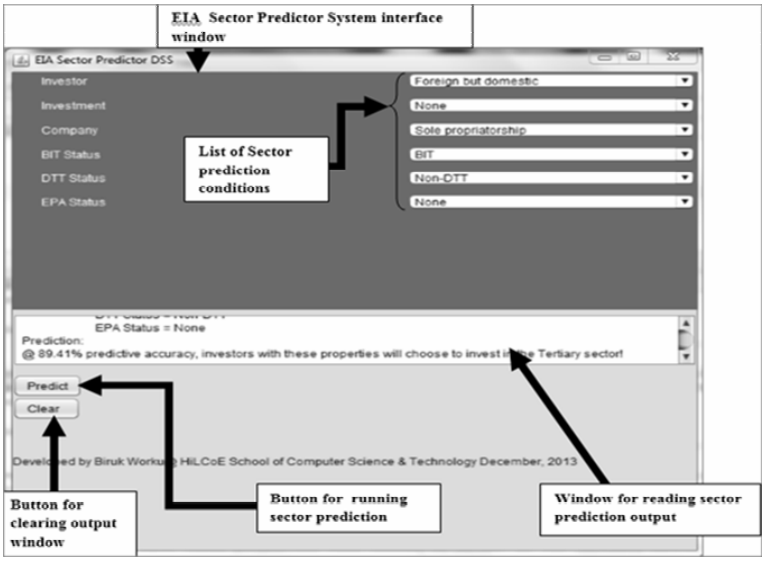

Figure 1: Graphical User Interface of the EIA Sector Predictor prototype system

4. Conclusion and Recommendations

In this research, the investment dataset, as collected from EIA, was explored from data mining point of view. Investment data have diverse features. They have a number of data explaining attributes that can be utilized for various useful investment business goals. If mined, patterns from investment datasets have a lot to provide for a better investment climate. Mining rules that assist the task of policy advocacy at the EIA, which is addressed in this research work, can be considered a good evidence to show such possibility.

Applying the Apriori and Predictive Apriori algorithms on the EIA's investment dataset, six data mining experiments were made and six CAR models were generated. Furthermore, this research explored new model evaluation techniques which are biased to the task of investment policy advocacy. The context aware model evaluation techniques, explored and applied in this research, can be used whenever the domain experts are unable to evaluate the created mining models. Moreover, the same concept and technique can be applied in other business areas which have similar problem.

Finally, as a solution to the decision making problem in the business domain, and based on the results from the mining models, a prototype of a decision support system was developed. The developed DSS can assist the EIA's investment policy advocators in their policy advocacy decisions.

Through data mining researches and application of data mining tools and techniques, good investment policies can be devised. Furthermore, current computer science and data mining technologies can be integrated with the tasks of investment in general and investment policy advocacy roles undertaken by IPAs in particular.

As an IPA, the EIA should strive for the better of its policy advocacy roles. This in turn can be achieved only when more research and exploration to the huge accumulation of data at the agency's custody is undertaken. To this regard, any number of data mining researches and decision support tools cannot be said enough.

As discussed earlier, the most important one among the tasks of an IPAs is the task of policy advocacy. This task is most important because it is a powerful arm for the generation of FDI and favorable spillovers through policy implementations. As it can be noted from the output of this research, the knowledge about the various attributes and combination of attributes of foreign-country-experienced investors provides for technology supports to the advocacy roles.

References

[1] S. Kotsiantis and D. Kanellopoulos, "Association Rules Mining: A Recent Overview," (GESTS) International Transactions on Computer Science and Engineering, Vol. 32, pp. 71-82, 2006.

[2] OECD, "Chapter 2. Investment Promotion and Facilitation," in Policy Framework for Investment: Users' Tool Kit, OECD, 2011.

[3] UNIDO, "Guidlines for Investment Promotion Agencies ", Vienna, 2003.

[4] A. Allen. (2011, September 16, 2013). Catalytic Philanthropy: Investing in Policy Advocacy, Available at http://www.atlanticphilanthropies.org/news/catalytic-philanthropy-investing-policy-advocacy

[5] M. Proksch, "Selected Issues on Promotion and Attraction of Foreign Direct Investment in Least Developed Countries and Economies in Transition," Investment Promotion and Enterprise Development Bulletin for Asia and the Pacific, pp. 1-18.

[6] X. Wan, "A Literature Review on the Relationship between Foreign Direct Investment and Economic Growth", International Business Research, Vol. 3, pp. 52-56, January 2010.

[7] H. Gorge and D. Greenway, "Much Ado about Nothing? Do Domestic Firms Really Benefit from Foreign Direct Investment?", IZA Discussion Paper No. 994, November 2003.

[8] L. R. De Mello, "Foreign direct investment in developing countries and growth: A selective survey", Journal of Development Studies, vol. 34, pp. 1-34, 1997.

[9] B. A. Blonigen, "A Review of the Empirical Literature on FDI Determinants ", NBER Working Paper 11299, April 2005.

[10] I. Staff. (2009, September 22, 2009). Data Mining for Investors, Available at http://www.investopedia.com/articles/basics/03/053003.asp

[11] M. Shweta and K. Garg, "Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and Comparative Analysis of Various Association Rule Algorithms," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, pp. 306-312, June 2013.

[12] U. Fayyad, et al., "From Data Mining to Knowledge Discovery in Databases," American Association for Artificial Intelligence, 1996.

Knowledge Discovery from Agricultural Survey Data: The Case of Teff Production in Ethiopia

Bruk Legesse
brucklyn21@gmail.com

Berhanu Borena
PhD Candidate, Addis Ababa University, Ethiopia
berhanuborena@gmail.com

Abstract

This paper examines classification and association pattern of agricultural productivity survey data of Teff in Ethiopia. Although different approaches of Statistic, Technology, Metrology and Geology were applied to identify factors contributing for improvement of Teff productivity, there remains a lot of work to bring overall change on productivity of Teff. This research focused on identifying relationships between attributes of agricultural productivity survey data of Teff with input mechanisms and techniques to clearly understand the nature of production of Teff in Ethiopia. In order to conduct this research we followed the approach called CRoss-Industry Standard Process for Data Mining (CRISP-DM). The results of this research using classification and association rule have discovered that the technique of data mining is applicable to generate knowledge from agricultural productivity survey data of Teff in agricultural production. Association algorithms: Apriori, Tertius, and FilteredAssociator were used to select the features/attributes that have direct and indirect relationship with the target class in addition to the expert judgment. Subsequently the classification algorithms namely J48, RandomForest, REPTree used to generate rules. The generated model is represented in terms of IF THEN RULE.

Keywords: Teff Productivity; Data mining; Classification; Decision Tree

1. Introduction

Agriculture is the foundation of Ethiopia's economy, accounting for half of the gross domestic product (GDP), 83.9% of exports, and 80% of total employment [1]. Ethiopia's agriculture is plagued by periodic drought and soil degradation caused by inappropriate agricultural practices and overgrazing, deforestation, high population density, undeveloped water resources, and poor transport infrastructure [2].

Teff is one of the cereals believed to have originated in Ethiopia between 4000 BCE and 1000 BCE. Teff accounts for about a quarter of the total cereal production in Ethiopia [4]. Teff has been widely cultivated and used in Ethiopia, which is used for making Injera. It is now raised in the U.S., in Idaho in particular, with experimental plots in Kansas, American customers including people from traditional Teff consuming countries and also people desiring to eat Teff to avoid glutens that irritate celiac disease [5]. Teff is native to Ethiopia where it accounts for one quarter of the total cereal production.

With the growth of population consuming Teff the need to improve its production is felt and given due attention. In this effort Although different approaches of statistic, technology, metrology and geology were applied to identify factors contributing for improvement of Teff productivity, there remains a lot of work to bring overall change on productivity of Teff Teff productivity were not thoroughly investigated using data mining approaches.

In this research we employed data mining approach to uncover previously unknown patterns related census data and Teff productivity.

2. Related Work

The literatures covered below are taken from the perspective of what kind of data input used, amount of instance, method or approach used, tool and algorithm used, output data and deployment.

Revathi and Hemalatha [9] presented a brief idea of some of the widely used data mining techniques over cotton growth and development. In this work some of the data mining techniques are discussed and presented such as Naïve Bayes, Decision tree, Multilayer Perception. The cotton seed dataset consists of 500 instances, with 24 attributes of varying quality - Good, Average and Bad. Their experiment shows that J48 decision tree predicts the best performance from other classifiers used for experiments. They also found the time taken to build the J48 and Random Tree take less seconds from other classifiers.

Similarly Peyakunta and Singaraju [6] compared the effectiveness of the classification algorithms Genetic algorithm, Fuzzy classification and Fuzzy clustering. They found that the accuracy rate of Fuzzy Classification rules is higher than Fuzzy C-means algorithm; Fuzzy C-means algorithm is more suitable for Unsupervised Fuzzy soil data.

The work of Rao and Das [7] was also on the classification of herbal garden to classify the herbal gardens information. The data was collected from different herbal gardens and developed a database and uploaded online at www.herbalgardenindia.org. This website at present has a total of 71 herbal gardens from all over the country that are registered as members in this network. A total of 1024 species are available in these herbal gardens from which the majority of the species present in the garden are herbs of plant type. It is noted that the total number of quantity of planting material available in the entire registered herbal garden is 9,060,426 cuttings [7]. The initiative behind this experiment was the question that what type of habit of species is present in which location. They modeled these questions to find the relationship between location of species and their habit.

Another work by Gholap *et al.* [8], proposed an analysis of soil data using different algorithms and prediction techniques such as Naïve Bayes, J48, and JRip with the help of the WEKA data mining tool. They found that J48 is a very simple classifier to make a decision tree, and it gives the best result in the experiment.

3. Method

3.1 Data Collection and Preparation

More than 80% of researchers working on data mining projects spend 40%-60% of their time on cleaning and preparation of data [10].

Before data preprocessing data understanding is needed. In order to accomplish the study the data is collected from FDRE Ministry of Agriculture and CSA (Central Statistical Agency) database and library. The initial target dataset contains a total of 135,524 records with originally 44 attributes.

After the initial data collection, the dataset produced is converted into CSV (Comma Separated Variable) file format to allow them to be used in WEKA.

The activities during the data preparation phase included data cleaning, data selection, attribute or feature selection, transformation and aggregation, integration and formatting.

Data cleaning is the process of examining data and determining the existence of incorrect characters and mis-transmitted information [3]. The activities during this phase includes deleting attributes and data, filling missing values as advised by a domain expert, identifying and removing outliers and resolve inconsistencies that occur in all field attributes.

Irrelevant or unneeded data are usually eliminated from the data mining database before starting the actual data mining function. Other criteria for excluding data may include resource constraints, cost, restrictions on data use, or quality problems [3].

We deleted 15 attributes including attributes that don't have direct relationship with the productivity of Teff as the experts judge.

3.2 Attribute/feature selection

Instances are evaluated and classified based on the values of their attributes. Theoretically, decision tree could determine relevant attributes for classification automatically using the concept of

information gain or entropy without manual efforts. Of course, some of the attributes of an instance may be irrelevant to the process of classification as well as association and thus should be excluded.

Missing data is a common problem in statistical analysis and Data Mining. Rates of less than 1% missing data are generally considered trivial, 1-5% manageable. However, 5-15% requires sophisticated methods to handle, and more than 15% may severely impact any kind of interpretation [11].

Accordingly the recommended data size we can deal with is attributes having up to 20% of missing values. Attributes more than 20% missing value were removed. These include:

1. Weight of improved seed (WTNISEED)➔ Number of deleted instances=1288
2. Source of water for irrigation (SIRRG)➔ Number of deleted instances=2275
3. Number of trees (TREES) ➔ Number of deleted instances=93
4. Number of trees of bearing age (TREESBA) ➔ Number of deleted instances=91
5. Improved seed cost (COSTIMPS) ➔ Number of deleted instances=1281
6. Type of natural fertilizer (D23)➔ Number of deleted instances=14,193

Table 1 shows the selected attributes among the original and total dataset.

Table 1: List of selected attributes and their description

Field name	*Data type*	*Description*
HHSEX	Nominal	Head Sex
FLDTYPE	Nominal	Field Type
OWNTYPE	Nominal	Owner Type
EXT	Nominal	Is the field under Extension program?
IRRG	Nominal	Is the field used Irrigated?
SERRO	Nominal	Is the field affected by soil erosion?
MERRO	Nominal	Measure taken for Soil erosion
SEEDTYPE	Nominal	Seed type used
WTNISEED	Numeric	Weight of non improved seed(Kg)
DAMAGE	Nominal	Is the Field damaged?
DREASON	Nominal	Damage reason if field damaged
DPERCENT	Numeric	Damage percent if the field was damaged.
DMEASURE	Nominal	Measure taken to prevent damage
DMTYPE	Nominal	Type of damage prevention
DMCHEM	Nominal	Chemical Used?
FERT	Nominal	Fertilizer Used?
FERTTYPE	Nominal	Type of fertilizer used?
D22A	Nominal	Type of chemical fertilizer used?
D22B	Numeric	If chemical fertilizer used, quantity in Kg
D24	Nominal	How many times do the farmers produce crops?
D26	Nominal	What was the field used for?
AREAH	Numeric	Area measured in hectare (Hectare)
PRODQ	Nominal	Production in Quintal per Hectare Area

4. Results and Findings

4.1 Feature/Attribute Selection by the Association Rule

We used expert's advice to select the features/attributes that have direct and indirect relation with the target class which is production of Teff in quintal per hectare. In addition to this commonly known procedure, we also tried to select and categorize those attributes as directly related or not by applying Association rule.

As we can see from the test runs conducted for the selected association rules, the result shows that there are attributes which have direct and indirect relationship or have association with the target class which is yield of Teff in quintal.

Though there are some differences the outcome of the entire three association rule test runs share most common points with the experts' advice.

The Apriori and Tertius algorithms in the above test runs indicate that most attributes has association with the productivity. Such as

1. Irrigation used (IRRG)
2. Extension (EXT)➔
3. Seed type (SEEDTYPE)
4. Any damage? (DAMAGE)
5. Damage reason (DREASON)
6. Any measure to prevent damage (DMEASURE)
7. Damage percent (DPERCENT)
8. Fertilizer used (FERT)
9. How many times do you produce crops (D24)
10. What was the field used for? (D26)
11. Field type (FLDTYPE)

Those attributes listed in above are indicated by the association rules assuming that they have a relation with the productivity as confirmed by the expert. Even though there are also other directly related attributes confirmed by the expert but not by the experiment such as

1. Type of chemical fertilizer used (D22A)
2. If chemical fertilizer, quantity in kg (D22B)
3. Area in hectare (AREAH)
4. Type of damage prevention (DMTYPE)
5. Chemical used (DMCHEM)
6. Fertilizer type (FERTTYPE)
7. Soil erosion (SERRO)
8. Measure taken to soil erosion (MERRO)
9. Weight of non improved seed (WTIMSEED)

The time taken to generate 100 rules by Apriori is 2.38 second, Tertius is 3.09 second, and FilteredAssociator is 1.25 second. Tertius algorithm is the slowest of all compared with all rest algorithms testes in this research.

4.2 Build and Test the Modeling for Classification using Decision Tree

The experimentation is performed through decision tree algorithms. The algorithms selected are J48, REPTree, and Random forest. The decision tree software employed for the purpose of this research was the Weka software package, which contains several classifiers, clusters and association algorithms.

4.2.1 Generating Rules from Decision Tree (J48)

Table 2: Confusion matrix for J48

Actual	*Predicted*		*Total*
	High	*Low*	
High	11714	3430	*15144*
Low	3845	17990	*21835*
Total	*15559*	*21420*	*36979*

The confusion matrix in Table 2 depicts that out of the total records provided to the program, 11,714 (77.35%) and 17,990 (82.39%) records were classified correctly in the class of High and Low respectively.

4.2.2 Generating Rules from Decision Tree (REPTree)

Table 3: Confusion matrix for REPTree

Actual	*Predicted*		*Total*
	High	*Low*	
High	11722	3422	*15144*
Low	4061	17774	*21835*
Total	*15783*	*21196*	*36979*

The confusion matrix in Table 3 depicts that out of the total records provided to the program, 11,722 (77.40%) and 17,774 (81.40%) records were classified correctly in the class of High and Low respectively.

4.2.3 Generating Rules from Decision Tree (Random Forest)

Table 4: Confusion matrix for Random Forest

Actual	*Predicted*		*Total*
	High	*Low*	
High	11324	3820	*15144*
Low	3663	18172	*21835*
Total	*14987*	*21992*	*36979*

The confusion matrix in Table 4 depicts that out of the total records provided to the program, 11,324

(74.77%) and 18,172 (83.22%) records were classified correctly in the class of High and Low respectively.

4.3 Evaluating the classification model

Evaluating the performance of a data mining technique is a fundamental aspect of machine learning. Evaluation method is the yardstick to examine the efficiency and performance of any model. The evaluation is important for understanding the quality of the model or technique, for refining parameters in the iterative process of learning and for selecting the most acceptable model or technique from a given set of models or techniques.

We used the available ones which was present on Weka tool such as k-fold cross validation, F-Measure, ROC and Confusion Matrix (precision and recall).

Table 5: Comparison of Algorithms for their Efficiency

Algorithm Efficiency Evaluator	*J48*	*REPTree*	*Random Forest*
F-Measure	0.804	0.798	0.797
Precision	0.804	0.799	0.797
Recall	0.803	0.798	0.798
ROC	0.857	0.858	0.87

As shown in Table 5, the three algorithms have shown to be very closely efficient with the efficiency evaluations. But there is a slight difference as shown in the result. ROC curve shows that REPTree and J48 are almost the same and RandomForest has a slight difference. Precision, F-Measure and Recall shows that J48 is more efficient.

Table 6: Comparison of Classification Algorithms for their performance

	J48	*REPTree*	*Random Forest*
Total Record	36,979	36,979	36,979
Number of Correctly classified instances	29704	29,496	29,496
Accuracy in Percentage	80.32%	79.76%	79.76%
Time taken to build the model	1.33 sec	1.51 sec	3.06 sec

As shown in Table 6 each algorithm takes the same numbers of instances which is 36,979 (100%) and gave us the knowledge. Though they take equal number of instances, the correctly classified numbers of instances were different due to the difference in algorithms.

5. Recommendation and Conclusion

This research attempted to study the application of classification rule mining to find relationship and interesting patterns between attributes of census or survey data and Teff productivity. Data collection, selection and cleaning were major tasks which took most of this research.

We attempted to select relevant records and attributes based on the objectives of this research and practices in data mining. The total number of attributes in the original target dataset was 44. We attempted to select the relevant attributes for the data mining task using statistical and agricultural experts' advice and data mining algorithm.

A dataset totaling 36,979 records was used in generating classification and association rules. Initially, all of the records were given with 24 attributes to Apriori, Tertius, and FilteredAssociator, which are association rule mining algorithms in order to identify directly related attributes with Teff productivity. Subsequently the records were submitted to J48, RandomForest, and REPTree for classification rule mining.

The result from the association rule indicates that from the total dataset used in this research 95.45% of Teff production was cultivated on pure field type

which is dedicated only for Teff production using damage prevention mechanisms and this has an association with the use of fertilizer which is about 94.71% of the total dataset.

The result from the classification rule also indicates that from the total dataset used in this research high productivity of Teff is associated with area in hectare, damage preventive measure, use of extension service, fertilizer used, , type of seed, sample weight of seed and type of head sex.

On the basis of the findings of the study and the experience gained from the research, the following recommendations are suggested. More attributes should be added to allow complete analysis of the productivity of Teff. Possible values of an attribute should be less ambiguous.

There are attributes which have directly and indirectly related attributes and not included in this research due to different reasons during data preprocessing. There are also attributes having missing values more than the accepted range value. Thus to enhance such studies we recommend collecting and using data that have a better quality and size.

Although the findings of the research are not conclusive, they, however, can be considered as insight-giving to the bigger picture of the phenomena of Teff Production against inputs and methodology being used. Accordingly, the findings can be incorporated for the advocacy and awareness raising efforts of CSA and the Ministry of Agriculture, preferably after being substantiated and supplemented by qualitative research.

References

[1] Wikipedia, the free encyclopedia, Agriculture in Ethiopia, 2012, Retrieved from: http://en.wikipedia.org/wiki/Agriculture_in_Ethiopia, Last accessed on June 5,2012

[2] Background note: Ethiopia, 2012, Retrieved from http://www.state.gov/r/pa/ei/bgn/2859.htm, Last accessed on June 1,2012

[3] Fayyad, Usma, Piatetsky-shapiro, G. Smyth, and Padharic, "From Data Mining to Knowledge Discovery in Database", 1996, Retrieved from: http://citeseer.nj.nec.com.fayyad96from.html, Last accessed on June 8, 2012

[4] Gabre-Madhin and Eleni Zaude, "Market Institutions, Transaction Costs, and Social Capital in the Ethiopian Grain Market", Washington, DC: International Food Policy Research Institute, 2001.

[5] Teff for Gluten Intolerance, 2012, Retrieved from: http://www.matr.net/article-6172.html. Last accessed on June 6, 2012.

[6] Bhargavi Peyakunta and Jyothi Singaraju, "Soil Classification Using Data Mining Techniques: A Comparative Study", International Journal of Engineering Trends and Technology, July to Aug, 2011.

[7] Nukella Srinivasa Rao and Susanta Kumar Das, "Classification of herbal gardens in India using data mining", Journal of Theoretical and Applied Information Technology, March 31st, 2011.

[8] J. Gholap, A. Ingole, J. Gohil, S. Gargade, and V. Attar, "Soil Data Analysis Using Classification Techniques and Soil Attribute Prediction", 2011.

[9] P. Revathi and M. Hemalatha, "Efficient Classification Mining Approach for Agriculture", International Journal of Research and Reviews in Information Science, June 2011.

[10] DV Kalashnikov, S Mehrotra, and Z Chen (2005), "Exploiting Relationships for Domain-Independent Data Cleaning", SIAM Data Mining (SDM) Conf.

[11] Edgar Acuna1 and Caroline Rodriguez (2004), "The Treatment of Missing Values and its Effect in the Classifier Accuracy".

Selecting Appropriate System Development Methodology to Develop Human Resource Management System for Ethiopian Center for Disability and Development

Fanuel Adugna Geche
fadugna@gmail.com

Workshet Lamenew
School of Information Science, Addis Ababa University, Ethiopia
workshet@gmail.com

Abstract

Selection of an appropriate System Development Methodology to develop an information system is one of among the most essential steps in any System Development. Finding and choosing of this methodology to do the work that follows for a Human Resource Management System (HRMS) falls also in that category. The Human Resource Management System by itself has much functionality and can be considered as fairly a large system. Therefore the choosing of a method among many should consider many factors and a wrong choosing on this stage might sicken the development process and may lead to failure.

Rational Unified Process (RUP) was the selected method which the researcher found to be helpful and constructive. It is well documented and is a complete methodology. It has a higher level of reuse and has reduced integration time and effort. Most importantly the ability of the methodology managing the change of requirements of the customer has added more value for selection.

The selected method's implementation has its own steps and dynamism which can be benefited from and its constraints also help to minimize the risks. The Human Resource Management System developed for Ethiopian Center for Disability and Development here has firm grasp on the business requirement and was developed entirely by using Rational Unified Process. The object oriented approach here for the definite realization on the functionalities recognized earlier made the execution of the project much more attainable and change insensitive. The interleaved nature of the method, the development environment chosen and the developed system being able to be realized in modular manner made a remarkable coherence to develop and deploy this Human Resource Management System.

Keywords: Selecting Appropriate System Development Methodology; Human Resource Management System; Small Enterprise

1. Introduction

It is common nowadays to see companies going for a partial or whole automation of their systems. The driving force behind this move may partly be attributed to a strong aspiration for higher efficiency and performance. Ethiopian Center for Disability and Development (ECDD), for which this paper come to life, is no exception.

Due to the fact that the organization's operations are project-based, there are regular recruitment, deployment and termination processes that are undertaken by the organization's Human Resource Department along with other in-process human resource activities that the department normally does. The motivation behind this paper is, therefore, the inefficiency of the manual record keeping and a time consuming handling of these routine activities.

A System Development Methodology refers to the framework that is used to structure, plan, and control the process of developing an Information System. Software methodologies are also purely a means to reduce project risk. "No methodology" is

still a methodology, just not a very good (repeatable, predictable, improvable) one.

Many methodologies and techniques may be used in the development of information systems. Choosing the right methodology and the technique that goes with it need a deeper understating of what the enterprise is involved in and a correct interpretation of the business procedure. When it comes to selection of methods, many facts should be involved for the success of any project. Keeping this in mind method choosing is the most important and complex part of any development.

Though there are many automated HR systems on the market, from the researchers' point of view, almost all of them have too much functionality and require a lot of customization to fit ECDD's requirements. All the available applications whether it is an open source or not, needs further analysis to fit it in the current HR department's activities. Analyzing it in this case may likely include customizing it as per the requirement developed earlier and that by itself will need an additional resource like skilled manpower on the perspective technology chosen, additional time is also required since we are studying other people's developed application and extra cost will more likely be incurred. Not to mention all the available HR systems are developed outside of this country, they greatly have a shortcoming on organization's business rules interpretation. Though some of them possess a great technological advantage and standardize the HR department functionalities, the cost of purchasing and deployment and maintenance of these applications are very high.

2. Related Work

There are no one-fit-all development methodologies. Some companies have their own unique customized methodology for developing products or services; others simply use standard commercial off-the-shelf methodologies. With the incorrect methodology, discovering, designing, building, testing and deploying a project is chaotic. The project methodology that is chosen represents merely the framework for the real work to be done and indicates where creativity is needed. There is no guarantee that the team will deliver if it just follows a chosen methodology. Clients seldom complete requirement specifications because their requirements are constantly changing. The most logical solution is to simply evolve the products as the client's needs change along the project development process. This shows the need for a methodology more flexible than a formal waterfall approach. In fact, the trend is shifting to the more iterative or incremental style of methodologies [1].

We all know that more than forty methodologies are there to use for development of a new application call it a standalone application or a web enabled one. All these methodologies have their own way of deployment techniques for them to become successes and produce an acceptable product for the user.

Attempt to actually *evaluate the available methodologies* has always given questionable results. As mentioned in [2], there are two general trends as to how various criteria for evaluation are organized. There are relatively ad hoc lists of criteria for evaluation, and systematically organized frameworks, which generally provide more authoritative assessment results. However, the frameworks investigated are too generic and disproportionate in their emphasis on certain parts of a method [2].On their conclusion also random lists of assessments criteria to evaluate methods do not tend to provide conclusive answers. Frameworks, which are based upon more rigorous logical foundations, deploy a more systematic questioning approach. The existing approaches investigated were theoretically sound, but there were deficiencies in pragmatics.

Selection of a suitable software model to use within an organization is critical for overall success of a project. The selection of one model over the others is driven by project size, budget, team size, criticality of the project and a lot of other factors. The different aspects which need to be kept in mind while selecting a suitable process model can be summarized as follows.

Reuse of existing Components? Component Based approach may be ideal choice.

Is it a very large project with High risks or high cost of failure? A Spiral process Model may be the best choice.

Although the customer has defined business goals but the requirements are not freeze yet Agile (light Weight) Model will have the advantage over others as it has the flexibility to change the requirement at any stage.

All the developers are experts? And if the project is small enough, an agile approach may work.

Want to keep stakeholders involved? An Incremental process Model may be what is needed.

Don't have an on-site stakeholder to sit with the developers? Agile development model is not suitable.

Developers are not highly skilled? A Waterfall or Spiral process Model can help keep them on track and out of trouble.

3. The Proposed Solution

The objective of this project was to Select the appropriate System Development Methodology and develop an automated Human Resource Management System (HRMS) (beta version) with the view to experiment on the selected Methodology. To achieve the above stated objective, the researcher will be performing the following activities: Selecting Appropriate Methodology, Conducting System Development using the selected Method, Experiment the selected Methodology, Deploy the system with in the intranet of the company. The scope of this project is limited to the Human Resource Management of ECDD that will be deployed in the intranet, which is handling the following modules only: - System Administration, Employee information management, Employee self-service management, Document Management, Notification and Alert management, Security Management and Report Management.

The activities carried out under the method were: Data gathering, System development Methods, Development Environment and Programming tools and training and testing procedures.

For easing the evaluation for the methodologies the majority of scholars agree that they fall under two categories that are Heavyweight and Lightweight. Heavyweight one is also known as traditional methodologies and their focus are detailed documentation, inclusive planning and demonstrative planning. Lightweight methodologies focused mainly on short iterative cycles and rely on the knowledge within a team.

There is also another assessment to evaluate a methodology from the project manager point of view. Decision must be made by the project manager on selecting a methodology. The table below which is taken from [1] show how the project manager can select a methodology.

Table 1: Requirements for selecting a Methodology

Requirement	*Rationale*
Budget	Methodologies require money and effects schedule
Team Size	Number of staff to be managed is required.
Project Criticality	The urgency of the project decides the methodology.
Technology Used	Hardware such as computer servers, composite materials, or electronics may be needed.
Documentation	The methodology needs documentation.
Training	Effective training to key support staff and project managers is required.
Best Practices/ Lessons Learned	Past lessons learned and good practice should be available.
Tools and techniques	Tools and techniques must be available.
Examination of existing processes	The maturity of existing processes will influence the pace at which a project will progress.
Software	Methodologies require software as part of their design.

When we come to actually choosing from the methodologies we have, we reviewed documents for the selection criteria.

There are criteria for selecting a Software Engineering Tools (SETs). Though the paper's objective were to select SETs for Small and Medium Enterprises (SMEs), the criteria used here for selecting SETs were much more sounding and can be used for this project too. Accordingly there are four (4) conditions used for the selection purpose.

1. Flexible Project Management
2. Development Process Support
3. Tool Usability
4. Communication and Collaboration

Table 2: Methodologies valuation based on the criteria

Method	*Flexible Project Management*						*Development Process Support*			*Tool Usability*			*Comm. & Colla.*	
	PRTE	*Planning*	*PT*	*Risk*	*Effort*	*AC*	*ATP*	*Modeling*	*VC*	*DF*	*IVF*	*TU*	*DC*	*TC*
XP	H	M	H	M	M	H	H	H	M	L	M	M	L	H
Scrum	H	H	H	H	H	M	H	H	H	H	H	H	M	H
Crystal	M	H	M	M	H	M	H	H	H	M	H	H	M	H
DSDM	H	M	H	L	M	L	L	M	L	L	M	M	M	H
RAD	M	H	H	L	H	L	L	M	L	L	M	M	H	M
Adaptive	H	L	M	L	H	M	M	M	L	L	M	M	L	H
RUP	H	M	M	M	H	M	H	H	M	H	H	M	H	M
FDD	H	M	H	L	M	H	H	H	M	L	H	H	M	H

Legend: H= High, M= Medium and L= Low

Criteria Description

PRTE: Project Resource and Time Estimation

PT: Project Tracking

AC: Attaching custom metrics

ATP: Adjustment to Processes

VC: Version Control

DF: Documentation Flexibility

IVF: Information Visualization Flexibility

TU: Tools Usability

DC: Documentation Collaboration

TC: Team Communication

The unquestionable actions of an organization are to become much more efficient in its responsibilities and accommodate numerous features in its existence. ECDD is no different. Non-governmental Organizations (NGO) activities are specific and target oriented. For the accomplishment of these tasks skillful personnel are required and that is what the organization has. What it lacks is a standardized and customized Human resource management system to make the allocation and job success within the company to its highest expected level. As we all know a Human Resource Management System (HRMS) has many functionalities and features. One fit for this company's Human Resource Department is very much hard and want more asking price in every aspects of its appliance. By dividing the HRMS's functionalities and processes and dealing with these tasks one at a time makes system development process for the Human Resource Department much more easy and attainable. Among all the methodologies available and seen before the viable choice for this task is using the Rational Unified Process (RUP). RUP's reputation and acceptance level by the Development community is very much recommendable for fairly large scope projects. Also for a chunk of that very large scope project development we found it to be comfortable and suitable. The following points are some of the points why this methodology was preferred:

- The interleaved nature of its steps works well with our way to develop a customized application

and greatly helps to integrate any change from the client.

- The fact that its documentation level is detailed makes for the later upgrading or changing of its functionalities by any Individual/team much more preferable.
- The fact that it is open, public and training materials are readily available also makes it more acceptable to implement.
- The style used by this methodology in development helps the developer to see what might happen in the future and makes some remarks for the anticipated attitudes.

Rational Unified Process (RUP) is an object-oriented and Web-enabled program development methodology. According to Rational (developers of Rational Rose and Unified Modeling Languages), RUP is an online mentor that provides guidelines, templates, and examples for all aspects and stages of program development. RUP and similar products - such as Object-Oriented Software Process (OOSP), and the OPEN Process - are comprehensive software engineering tools that combine the procedural aspects of development (such as defined stages, techniques, and practices) with other components of development (such as documents, models, manuals, code, and so on) within a unifying framework.

RUP establishes four phases of development, each of which is organized into a number of separate iterations that must satisfy defined criteria before the next phase is undertaken. The phases are Inception, Elaboration, Construction, and Transition.

4. Discussion

Here we will explain what was done and how it was done. The scope on the methodology is depicted in Figure 1.

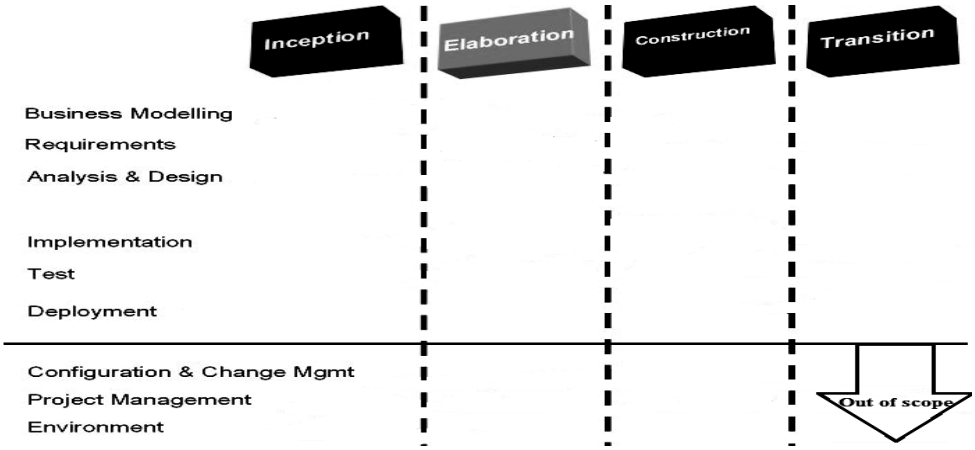

Figure 1: RUP's Methodology Understanding

The methodology as it was stated before has a nature of interleaving steps in order to attain the dynamic requirement needs of the customer. The Table 3 specifies what was done on what stage and what the outcomes were.

Table 3: Summary of Methodology used

	Inception	*Elaboration*	*Construction*	*Transition*
Business Modeling	By using the Business use case and the Business Object model we tried to communicate the business engineering process with the software engineering one.	The final stage for this step was here. The output this step is a document which is called Business use cases which are analyzed to understand how the business should support the business process.	Non for Business modeling	Non for Business modeling
Requirements	A complete listing of the main features and functionalities for this version of the product.	Identifying and listing of all the driving requirement that will helped me in choosing the architecture to be used and the quality of the product that am developing. Mostly the above points were forced by the system's non-functional requirements.	Non for Requirements	Non for Requirements
Analysis& Design	Actor Identification for the system.	Selecting the appropriate architecture for the system and identifying the system features was done here.	A full business process model that will describe the use case of the system and the activity diagram which will help for the document preparation of the *Problem* and *Solution* domain of the system was done here.	Distribution of classes of the system in to large scale by using the component view.
Implementation	Not applicable at this stage	The Subsystem overview and the persistence model which helps us to identify the data model of the system	Actual coding and User Interface preparation was done.	An Entity Data Model that is used for linking the business and Data model.
Test	Not applicable at this stage	Unit Testing on the data model and normalization of the tables existing.	Unit Testing, Regression Testing.	Unit testing, Regression testing, System Testing.
Deployment	Not applicable at this stage		Primary stage of distinguishing the deployment area was done.	Setup Server & configure, Configure Client work stations, if needed Configure internet

5. Conclusion

As a conclusion any software development process must follow a methodology for a complete success and accomplishment of a project. The right methodology for a project has a high degree of acceptance with both the developer and the customer which are involved. Method usage by itself will give well-studied steps and tasks to get to where one wants to go by making the communication and collaboration interactions between the customer and developer a lot easier and will also be goal oriented. Not to mention choosing the right method will greatly help on planning, risk reduction, effort exerted to complete a task and effective time usage, it will also make it more flexible to changes made and team communication will be easy.

The Human Resource Management task as it was stated many time in this paper is one of the most important and projecting part of any organization. By bringing this part of an organization to a very well acceptable standard stage, the organization itself could make the best out of it personnel set. ECDD is no different. HR has many functionalities, and I strongly believe that it should be split to a more manageable and separate modules where later on all could be put together to make one complete customized system that serves well within the company.

Coming to our method usage on the HR, we found the choice of method which is RUP to be best suited for this small enterprise. During the selection phase, we tried to make a brief and short amalgamation between the RUP methodology, HRMS and the ECDD and why it was worthwhile going this way.

In our opinion Rational Unified Process (RUP) usage for development was suitable. Though the level of expertise needed by this method is high, we found it to be more interesting and sensible to actually go though it to produce a web based application for ECDD. Its high business value deliverance and its object realization in each step made the work much more easy and thrilling. The greater advantage that we have noticed on this method is the high and demanding need of the client involvement though it is a heavyweight made it more suitable for this application development.

References

[1] Jason P. Charvat, “Project Management Methodologies: Selecting, implementing, and supporting Methodologies and Processes for Projects”, 2003.

[2] Peter Bielkowicz, Preeti Patel, and Thein Than Tun, “Evaluating Information Systems Development Methods: A New Framework”, London Guildhall University, UK, 2002.

Knowledge Sharing Architecture for UNDP Ethiopia

Genet Awlachew
UNDP-Ethiopia Country Office, Addis Ababa, Ethiopia
genetawlachew@yahoo.com

Mesfin Kifle
Department of Computer Science, Addis Ababa University, Ethiopia
kiflemestir95@gmail.com

Abstract

Knowledge sharing is a process of exchanging skills, experience, and understanding among people. This phenomenon is increasingly becoming critical within the United Nations Development Programmee (UNDP). UNDP had introduced its own motto of becoming a 'knowledge-based organization' has and since 2008 developed a comprehensive Knowledge Management Strategy (KMS). Part of this strategy is a web-based knowledge sharing tool called Teamworks (TWs) currently used in many of UNDP's Country Offices including in Ethiopia. The problem, however, is that staff often shy away from using TWs due to poor motivation, absence of proper understanding about the full benefits of the tool, as well as other technical problems.

Taking the experience of the Country Office (CO) in Ethiopia, this study presents information and discussion on the development and implementation of the UNDP's knowledge sharing practice. In conducting this brief research, wide literature review was carried out as well as data was gathered through a set of questionnaires presented to different group of staff. Information obtained through these methods was further supplemented by both desk study and authors' participatory observation. The objective of the research focused on identifying barriers and enablers in using TWs as a knowledge sharing tool within the CO in Ethiopia. The outcome of the study showed that despite the rhetoric, TWs is still only moderately used at the CO level and that its full utilization by staff is dependent on several factors. These factors included whether there exist personal benefits in choosing a particular platform, supervisors' recognition, impact on annual assessment of staff performance, and individual own awareness to certain aspects of social media interfaces. At the core of the finding was that the availability of any of the above factors could be an enabler while the absence becomes a barrier. On the basis of that, the research recommended both system level and technological solutions.

Keywords: Knowledge Management; Knowledge Sharing; Social Networking; Social Media, Teamworks; Knowledge Sharing Architecture; Barriers and Enablers

1. Introduction

In modern business operations organizations want to remain competitive and successful. To that end they must create, capture, keep, and share knowledge and expertise efficiently and effectively. According to Boer [1], this is the main reason why many organizations are increasingly showing interest in the field of managing and sharing knowledge. Alarbi *et al.* [2] argue that the continued advancement in the field of Information and Communication Technology (ICT) has further accelerated the chances for continued use of knowledge sharing practices in knowledge creation and acquisition. One such organization is UNDP.

As a global organization, UNDP works in 177 countries including Ethiopia with thousands of staff working with numerous partners and helping billions of people around the world. This organization possesses immense knowledge through its diverse staff whose experience have to be systematically captured, organized, and made accessible to all. On that basis, UNDP has globally positioned itself to become a knowledge-based organization. Since 2008 it has developed a new KMS aimed at creating human and technical infrastructure. Accordingly the UNDP's CO in Ethiopia has introduced a web-based

knowledge sharing and social networking platform known as TWs to its staff as well as partners. TWs is a secure, globally available Web 2.0 Extranet platform developed to enable UNDP to leverage the collective knowledge of communities, individuals, programmes and projects [3].

Within the forgoing context, the concept of knowledge sharing architecture depicts a design through the components of people, knowledge objects, technical infrastructure and knowledge management processes enabling the process of sharing knowledge [4]. For such architecture to establish a successful knowledge sharing design and to function properly, any organization need to fulfill four steps, namely, identifying the drivers for the knowledge sharing process; establish the knowledge sharing community; strategize to implement the initiatives; and to ensure sustainability [5]. This shows that the main components of knowledge sharing architecture include the knowledge components itself, the KM process, the IT and the organizational aspects.

At the UNDP Ethiopia, many aspects of the above components of the KS architecture are still evolving. As such the UNDP's KMS acknowledge that the introduction of knowledge sharing practice is a critical institutional success factor. One pillar of the strategy is, therefore, the consolidation of the organization's existing knowledge networks, including a particular emphasis on outward facing networks, into a unified global platform managed via the TWs networking system. However, the use of such system is rarely appreciated and staff still tends to use more of the previous email based knowledge sharing. Many shy away from using the new web-based system to share their knowledge, expertise and best practices with their colleagues, UN family and partners.

There are several factors that impede the utility of new initiatives, such as TWs. These factors, which we call in this paper as barriers and enablers, need closer scrutiny. Taking the TWs as a knowledge sharing platform within the UNDP CO in Ethiopia as an example, therefore, this study will explore these enabling and/or impeding factors. The authors presume that though the design of TWs embodies the approach required in mitigating wastage of knowledge and information within the organization under discussion, at the same time insufficient institutional motivation, lack in knowledge sharing culture, as well as technological barriers make the current knowledge sharing processes difficult. Consequently we try to recommend what should be done to enhance the organization's knowledge sharing practices.

2. Related Work

The concept of knowledge is far complex than the mere availability of information. Collison and Parcell [6] argue that knowledge is much richer than data or information not only in its content but also in aspects of its usefulness. According to them, knowledge exists in a hierarchy in which value is added as one progresses from capturing to creation to sharing and application of the concept. In the field of Computer Science, knowledge is often confused with other concepts such as data and information, which, in turn, are regarded as the basis of knowledge. Theirauf [7] delineates the three inter-related components as follows: data is the lowest point or an unstructured collection of facts and figures; information is the next level, and it is regarded as structured data; whereas knowledge is "information about information". Zins [8] claims that both the three elements are part of a sequential order. Data are therefore the raw material for information, and information is the raw material for knowledge thereby depicting that data through time evolves to information which sequentially transform to knowledge.

For the purpose of this study, the authors apply the definition stated by Davenport and Prusak [9] who depict knowledge as "a fluid mix of framed experience, contextual information, values and expert insight that provides a framework for evaluating and incorporating new experiences and information".

Such knowledge could exist in different formats: in documents or knowledge repositories as well as in human minds. According to Gamble and Blackwell [10] knowledge originates and is applied in the 'mind of the knowers'. When it comes to organizations, it often becomes embedded not only in documents or repositories, but also in organizational routines, practices and norms. From the KM perspectives, knowledge could be explicit often found in documents or tacit which is found in human minds as unwritten, unspoken or hidden knowledge. According to Nonaka and Takeuchi [11] these two forms of knowledge are strongly linked and cannot be managed independently.

Irrespective of in what form it exists, knowledge, in order to be useful, it should be properly managed, transmitted and expanded. Senge [12] contends that sharing knowledge is not simply about giving people something, or getting something from them. According to him, the concept also denotes helping to solve problems together. Hence, knowledge sharing takes place only "when people are genuinely interested in helping one another develop new capacities for action; it is about creating learning processes" [12]. The measure of knowledge value is, therefore, whether it is shared or made easily assessable to its users. Moreover, knowledge sharing needs time and a common point or a platform where people meet, connect and share.

Historically, the practice of sharing knowledge is as old as human existence while the management aspect that is very recent. Even though some literature uses these concepts interchangeably, the two are different. It is the advent of knowledge management field that has helped knowledge sharing in getting significant attention. Duhon [13] states that knowledge management is "a discipline that promotes an integrated approach to identifying, capturing, evaluating, retrieving, and sharing all of an enterprise's information assets". Whereas for King [14] knowledge management is "the planning, organizing, motivating, and controlling of people, processes and systems in an organization to ensure that its knowledge-related assets are improved and effectively employed". Yet Awad and Ghaziri [15] try to depict that the concept of Knowledge Management as "a newly emerging, interdisciplinary business model that has knowledge within the framework of an organization as its focus".

Despite varying emphasis, all these actors feature the existence of key components in any KM discourse; which are: people, processes, and system. These components in turn form the KS architecture. In that context neither such literature nor the experience under discussion at the CO in Ethiopia level shows the lack of which specific factors enable or impede the organization's KS effort. Due to that gap both management and staff remain unable to clearly establish the usefulness and contribution of KS tools in general and TWs in particular, albeit only in very broad terms.

3. The Proposed Solution

This section proposes ways that help to enhance knowledge sharing architecture for the UNDP CO in Ethiopia. These recommendations are presented in the form of both system level and technological solutions based on the current research.

3.1 Structural Solutions

There are motivational methods to encourage staff's knowledge sharing behavior within a given organization. In such a situation, according to Bock *et al.* [16], however, the challenge is in keeping people motivated and interested in sharing what they know and experience. At the structural level, therefore, organizations like UNDP CO Ethiopia should consider the following approaches:

- Introduce structures and motivating methods to encourage staff not to keep their knowledge for themselves and refrain from sharing with others. To that goal, management could evaluate staff annual performance by recognizing individual contribution in knowledge sharing and include it in the RCA.

- Knowledge sharing success should be seen in its totality, not only in just setting up the various platforms but also whether such efforts involve appropriate knowledge exchange must be an equal concern.
- Conduct periodic knowledge audits, tailor-made trainings complemented by initiating of discussions on targeted thematic issues. That will further help to socialize the knowledge sharing among its staff, management and wide stakeholders.
- Aiming beyond the technical and operational aspects both management and staff must work in unison and commit themselves to the principle of 'Delivering as One'. That helps in bringing a comprehensive understanding of knowledge sharing as neither a onetime assignment nor an isolated activity.

3.2 Technological Solutions

The introduction of Web based technology and social networking platforms continue to play significant role not only in connecting people but also in empowering individuals in sharing their knowledge and experience. The following technological solutions are therefore recommended as enablers in the specific case of TWs' utilization for particular needs of the CO in Ethiopia:

- TWs should preferably be open like an outlook interface so that the logging is not required whenever staff look for or share for information/knowledge. To this end the application of Single Sign-On (SSO) which is a session/user authentication process that permits a user to enter one name and password in order to access multiple applications [16] is recommended. Then SSO can be synchronized with the existing UNDP's tools and systems such as e-mail, Intranet, and Atlas (UNDP's internal work-tool for finance, projects and HR matters). SSO encourages staff to use TWs more frequently than the current intermittent levels.
- Due to its collaborative environment, UNDP's working system requires teamworking which includes writing reports, concept notes, Terms of References (ToRs), etc. in groups. To facilitate this, the introduction of additional features such as Workflow technology would enhance the TWs platform as a knowledge sharing tool.
- TWs consumes high Internet bandwidth which otherwise constrains the long-term functioning of the KS system. This requires changing or transforming the existing TWs's design to have reduced page. Given the slow Internet connectivity speed in the host country (Ethiopia), optimizing the content of the pages should reduce the time that takes to open or download pages, making it efficient and more user-friendly. To resolve the connection/ bandwidth problem, adopting the use of a mirror server to the nearest place could be also helpful.
- Additional features such as Groupware or Instant Messaging could be added to the existing TWs's interface to show who is online so that users can communicate in a real time situation. This will increase TWs usefulness and familiarization by the general staff.
- Periodic Knowledge Audit (preferably once in two-years) is recommended to be explored by the TWs developer or staff at Headquarters to identify feasibility and suitability to the CO's available resources.

4. Discussion

A survey was conducted focusing on the various factors and circumstances that affect or simulate staff to participate in the knowledge sharing practice within the Country Office. Questions were posed on how efficiently staff used the different knowledge/ information sharing tools, such as Facebook, Twitter, LinkedIn, in the office. 57% said that they are extremely efficient while 28% are moderately efficient with the same method. Respondents said

they are comfortable and efficient in using the TWs for knowledge and information sharing and the response rate was 78% and out of these, 14% are extremely efficient while the rest (69%) are just efficient. The aggregated result showed that out of the total respondents exactly half (50%) of them preferred face to face discussion and the other half preferred a technologically assisted mechanism.

In another part of the survey participants were asked if they participate in one or more than one of the knowledge practice networks. The answers obtained were varied in which over 70% said that they have participated/contributed their knowledge, experiences and expertise to the UNDP community through several networks; while 28% of the respondents stated that they have never participated the E-discussions launched or query posed through these networks. Moreover, a particular question was raised to identify which specific factors discourage the respondents from participating in the current knowledge sharing practice. Majority of them (76%) said that lack of time was the most inhibiting factor while a lesser number (24%) said that fear of providing wrong or faulty information/knowledge prevents them from participation.

Regarding the question on how often respondents login to TWs, results show that 50% of them login at a moderate frequency level - that is few times per month while the remaining 36% and 14% login very infrequently and very frequently respectively. 57 % of respondents contribute to the knowledge sharing practice using TWs sometimes and 43% contribute hardly ever. No respondents were found contributing to the TWs on a daily basis. The top most frequently used among the four and popular knowledge sharing and social networking tools, is Facebook followed by Teamworks, Tweeter and LinkedIn.

Respondents were further asked to evaluate TWs as a knowledge sharing platform and its features. Analysis of the responses depicts that over 40 % of respondents rated as "good" for the specific arguments of "User friendliness of the features". 21% of the respondents rated the user friendliness of TWs as "Moderate". Only 7% of the respondents rated it as "excellent". Further, participants of the survey were requested if they have participated in any of the UNDP's series of the TWs training sessions organized during and after the TWs' prototype phase. Responding to that question, 62% of respondents said they attended the training and the remaining 38% didn't take part in any of the sessions.

Overall, the survey showed that majority of staff have a good knowledge and skill on how to use TWs and its various features. However staff participation in using TWs as a knowledge/information sharing and social networking platform are very low. It was found that there is a higher use of e-mail and other social networking tools for sharing knowledge and information than TWs. Face to face discussion is the most widely preferred type of knowledge sharing mechanism while TWs is the least preferred. This shows despite the high number of those who claim to have been participating, the level of activism or frequency still remained significantly low.

The above statement is supported by observations on the system carried out during of the span of the study time (June- –August 2013). The system dashboard being open and displaying all the global TWs users' contributions shows who has done what in terms of knowledge sharing. Constraints such as time and fear of providing wrong information or knowledge were the top most frequently cited barriers affecting respondents from engaging in knowledge sharing at their current position. Almost all respondents reported that they have access and are familiar with similar kinds of knowledge and social networking tools including TWs.

Analysis into available data further showed that though TWs utilization is very low, majority of respondents were found to be members of the global Practice/Knowledge networks. But their contributions to the E-discussion or query posed by these networks members and facilitators through TWs are almost zero or minimal. As indicated earlier, however, the knowledge sharing concept is beyond reading and viewing but helping,

collaborating, advising and sharing knowledge, experiences, expertise, ideas, etc. Hence, being a member to a given knowledge sharing platform does not necessarily mean someone is a contributor. Analysis on other data indicate that majority of respondents reported that their contributions will increase if knowledge sharing practice is highly linked with staff RCA and supervisors' recognition. As the researchers' participatory observation further supplemented, this practice is already in place but need to be further and applied at the CO level.

The authors made further investigation and empirical study to identify the reason why TWs is not highly utilized by its users for sharing knowledge/information and networking as well as documents and files repository. TWs were compared with other public domain social Media – such as Facebook, Tweeter, and LinkedIn. These three each has millions of users who are very much attached to them on a regular basis. So the researchers supplemented the study by login into the three public social media and UNDP's TWs simultaneously and tried to analyze the pros and cons of each and every feature in these tools in a real time evaluation/exercise.

The results of the investigation on the above showed that staff contributions to the knowledge sharing practice is dependent mainly on three factors: a) receiving personal benefits; b) recognizing that staff contribution is valued by colleagues; and c) noticing that knowledge sharing is reflected on staff job description as one of the competencies. Similar enquiries also showed where increased motivation came from staff to contribute to the knowledge sharing practice in the office namely supervisors giving positive feedback for contributing; seeing staff contributions reflected on staff annual performance Result Competency Assessment (RCA)and noticing that other staff as well as supervisors contributing to the knowledge sharing practice. On the other hand, lack of time and fear of sharing wrong information are the most inhibiting factors affecting respondents from participating.

5. Summary

This study was conducted on existing knowledge sharing practices at the UNDP CO in Ethiopia. Data have been obtained from a survey and these data were geared to answer the questions posed in the beginning. The analysis made on the available data served as a basis for drawing conclusions and subsequently to forward recommendations. To evaluate the recommendations the authors sent out formal request for a group of seven KM experts who shared their views. The problem with a knowledge management/knowledge sharing strategy for the UNDP Country Offices, including in that of the Ethiopia Office, is mainly tied to the business process. Ideally a knowledge management or knowledge sharing strategy should sit within a wider KM/KS framework and use common systems, tools, and protocols.

The easiest ways to get going in any business related knowledge management system or knowledge sharing practice including that of TWs require visioning from the management side and motivation from the part of general staff to make use of it.

Many validated the argument that people refrain from sharing their knowledge because they do not have motivation, lack time to do so or are unaware of the value of the knowledge they share. Moreover, people may not be conversant enough with the technology used. Hence the need to make the TWs more user-friendly lies on sharing the secret for the success of those social media like Facebook: simplicity.

As guiding principle, knowledge sharing should be user-based for the identification of needs to guide the processes and structures that will be required to its implementation. The answers why knowledge sharing becomes necessary lay in accountability, effectiveness and efficiency and enhance learning, evidence-based decision making, etc. Linkages to these objectives are essential.

Overall, some of the ideas raised here could be instrumental in provoking further similar undertakings (see below future works).

6. Conclusion and Future Work

6.1 Conclusion

UNDP's organizational productivity is closely tied with the production and dissemination of knowledge. The introduction of TWs as a knowledge sharing tool helps to fulfill this mission and becoming a knowledge-based entity. In this paper, the authors investigated existing knowledge sharing practice at UNDP taking the specific experience of the CO in Ethiopia. An understanding was drawn that people who share knowledge are more likely to solve problems and to collaborate for common purposes and interests. It is of the highest advantage for an organization to see its KS system and practice through its employees' ability helps to share knowledge, insights, ideas and solutions. In turn, this indicates that knowledge increases and becomes more useful when it is shared.

This paper has as its goal to explore and propose a structural level knowledge sharing architecture with the aim to improve utilization of TWs as a convenient knowledge sharing platform. Accordingly, to examine the various barriers and enablers, the authors presented information and discussion pertaining to the development and implementation of UNDP's KMS and within that the knowledge sharing practice. Using literature review together with information gathered from primary sources (questionnaires and participatory observation), the research examined into the practical experience at the Ethiopian CO level. In all that, Teamworks was particularly selected as a subject of closer scrutiny. The outcomes of the research showed that despite the rhetoric, TWs as a knowledge sharing tool is still only moderately used and that its full utilization by staff still faces several barriers.

It appears that expressed enthusiasm by management and staff within the organization towards the knowledge sharing initiative is not without challenges. Much depends on whether there exist personal benefits associated in choosing a particular platform, supervisors' recognition, impact on annual assessment of staff performance, and individual awareness to certain aspects of social media interfaces. The availability of each of these factors could be an enabler while their absence could be a barrier. In addition to the inconsistent motivational factors, fear not to make mistakes from the part of employees could further constrain the sharing of information. On the basis of that, the research was able to recommend both structural level and technological solutions. That requires commitment and the will to apply this enthusiasm in a day to day work environment.

In due course, though, it is reasonable to assume that TWs has the chance to succeed if the barriers identified are properly addressed and the enablers that exist at various levels are adequately exploited. At the same time, the indicators of success should go beyond simple sharing but also to include knowledge generation, storage and dissemination. The system must also be able to win the interest of all staff and involve those at the leadership level for TWs to modestly contribute to UNDP's motto of becoming a 'knowledge-based organization'. However, both the barriers and enablers identified here are neither exhaustive nor permanent. Thus, only the combination of efforts to lessen these bottlenecks and strengthen the good practices would make the knowledge sharing effort eventually a productive undertaking.

6.2 Future Work

Knowledge sharing is an ever expanding body of knowledge within the field of knowledge management. That implies the need for continued research and updating of the knowledge sharing practices. In light of this general observation and the limitations listed above, we would like to suggest the following key areas for future work:

a) Investigate what constitutes a fully developed knowledge sharing architecture? Does the

current practice at the UNDP CO level in Ethiopia meet those requirements? If not, anything missing need to be examined.

b) Evaluate and validate the recommendations suggested in this study; particularly the applicability of the technical solutions.

c) Further investigate would TWs utility continue, in an organizational setting, to be equally important like that of the other social media tools such as Facebook, Twitter, etc.?

Even after these works, however, one time research would not be adequate to fill the gaps in the above areas. Instead, it requires cumulating on the experiences gained by the Country Office in the process of knowledge sharing and improve the usage of the TWs platform in a step-by-step fashion.

References

[1] N. I. Boer, "Knowledge Sharing within Organizations: A situated and relational Perspective", Erasmus School of Economics, Rotterdam, 2005.

[2] Z. Alarbi et al, "Knowledge Sharing Culture in Libyan Organizations", in Proceedings of ICMESS International Conference on Management, Economics and Social Sciences, Bangkok, December, 2011.

[3] Knowledge Strategy: Enabling UNDP to share and leverage its knowledge and experience (2009-2011), https://undp.unteamworks.org/login?destination=node%2F44690, Last accessed on August 16, 2013.

[4] C. M. Pompiliu and I. Cristescu, "Knowledge Management Architecture – Principles and Tendencies" in Proceedings of Academy of Economic Studies, Bucharest, 2008.

[5] J. Holm, "Creating an Architecture to Deploy Knowledge Management at your Organization", NASA/KM Asia proceedings by California Institute of Technology, Los Angeles, July, 2002.

[6] C. Collison and G. Parcell, "Learning to Fly: Practical Knowledge Management from Leading and Learning Organizations", Capstone, Minnesota, 2004.

[7] R. J. Theirauf, "Knowledge Management Systems for Business", Quorum Books, Connecticut, 1999.

[8] C. Zins, "Conceptual Approaches for Defining Data, Information and Knowledge", http://www.success.co.il/is/zins_definitions_dik.pdf, Last accessed on September 3, 2013.

[9] T. H. Davenport and L. Prusak, "Working Knowledge: How Organizations Manage What They Know", Harvard, 2000.

[10] Gamble and J. Blackwell, "Knowledge Management: A State of the Art Guide", Kogan Page Publishers, London, 2001.

[11] Nonaka and H. Takeuchi, "The Knowledge-Creating Company: How Japanese Companies Cerate the Dynamics of Innovation", Oxford University Press, New York, 1995.

[12] P. Senge, "The Fifth Discipline: The Art and Practice of the Learning Organizations", Random House, New York, 1999.

[13] B. Duhon, "It's all in our Heads", Journal of Applied Knowledge Management/Inform, Vol. 12, No. 8, September 1998, pp 8-13.

[14] W. R. King, "Knowledge Management and Organizational Learning (2009)", University of Pittsburgh. http://www.uky.edu/~gmswan3/575/KM_and_OL.pdf, Last accessed on September 16, 2013.

[15] E. M. Awad and H. M. Ghaziri, "Knowledge Management", Princeton Hall, New Jersey, 2007.

[16] G. Bock, Y. Kim and J. Lee, "Behavioral Intention Formation in Knowledge Sharing: Examining the Roles of Extrinsic Motivators, Social-Psychological Forces, and Organizational Climate", MIS Quarterly, Vol. 29, No. 1, October 2005, pp. 87-111.

Online Service Delivery of Geo-information Data: The Case of Ethiopian Mapping Agency

Ghebrealif Assefa
Ethiopian Mapping Agency, Addis Ababa, Ethiopia
alef1995@yahoo.com

Sebsibe Hailemariam
Department of Computer Science, Addis Ababa University, Ethiopia
sebsibe2004@yahoo.com

Abstract

This paper explores the use of online service delivery of geo-spatial data for Ethiopian Mapping Agency (EMA). The main objective of this research is to automate online service delivery system on Geo-Information data for customers of the agency. Customers seek to get geo-spatial data timely and adequately but those living far away from Addis Ababa have been coming physically to EMA to be served without knowledge of what services and products are availed in the agency. This caused customer dissatisfaction. To boost the service delivery system of the agency, need assessment was undertaken from EMA and from 16 identified stakeholders of the agency using primary and secondary data collection methodological approach. As a result, a web portal was developed based on the requirements of EMA and its stakeholders to support the online service delivery system.

Keywords: Online service; Geo-information data; Service delivery

1. Introduction

Ethiopian Mapping Agency (EMA) is the authorized Agency in Ethiopia mandated to produce and disseminate geospatial information products and services. To accomplish this mission, the agency has been using tedious and time taking process to identify map details from hard copy of topographic maps. Due to long process of service delivery system in EMA, some regional states including most of the governmental sectors are forced to produce their own geo-information data spontaneously without legal autonomy. The result of this disintegrated geo-information production causes data incompatibility and unnecessary duplication of data which causes economic loss at national level. To avoid this problem online service delivery system and automated retrieval of map details is devised in this paper by addressing the following research questions.

1. How an existing spatial search engine (SSE) is integrated with the proposed system to support online and offline services?
2. What mechanisms are used to build a secured geo-information database?
3. What techniques are used to develop user friendly prototype and system architecture that satisfies customers of EMA?

To address these research questions, relevant literatures were reviewed to understand and share experience from various countries. New architecture was also designed to be integrated with the existing system but the SSE was adopted from Environmental Spatial Research Institute (ESRI) software products for visualization and printing of map details that can be used for Intranet and Extranet. To span the gap between demand and supply of geo-information data, we used semi structured interview with marketing officers of the agency by purposive sampling technique. Requirements were also analyzed from the existing process as well as from stakeholders to build an improved service delivery system and results were summarized from the data collected by majority respondents' reaction to the existing service delivery system.

This paper is organized as follows. Section 2 presents literature review of related works of online service delivery system used in other countries.

Sections 3 and 4 deal with the analysis and design of the developed system, respectively. Section 5 deals with prototype development. Section 6 presents discussions and finally conclusion and recommendations are briefly given in section 7.

2. Related Work

To understand various techniques of online service delivery systems and spatial data retrieval, a number of literatures were reviewed. Most of the literatures were valuable to obtain additional knowledge to the study. Experience of online service providers of geo-information data in other countries was also reviewed to be improved. By comparing six technologies, we selected ArcGIS server for Spatial Search Engine (SSE) to adopt on this system because it is cost effective, compatible with the stakeholders' data and interoperable with ArcGIS software. We also reviewed eleven previous related works such as: Geospatial Portal from Intergraph [1], The Atlas of Canada [2], Maps of India.com [3], Caribbeean-on-line.com [4], Google Maps [5], Yahoo Maps Local [6], Maps of the World United Nations Cartographic Section [7], Standfords [8], Microsoft's Mappoint [9], Perry-Castañeda Library [10] and The United Nations Educational, Scientific and Cultural Organization (UNESCO) [11] and we compared with the new system using four parameters and in all parameters we conclude that the new system is the best of all other systems because it is rich in Ethiopian gazetteer, download-ability, search-ability and suitability to the existing data of Ethiopian Mapping Agency.

3. The Proposed Solution

A new system is proposed to serve customers online and offline. The new system has a web portal to serve as extranet (network outside EMA) with detail Metadata descriptions so that the external user can have an idea of what geo-information production and services are delivered by EMA. To design online service delivery system, need assessment was made by distributing questionnaires and by making semi structured interviews for the purpose of identifying functional and non-functional requirements.

3.1 Functional Requirements

In software engineering, a functional requirement defines a function of a software system or its component that explains and describes the interaction between the system and the users [12]. The new system provides the following functionalities.

- Support to search any specific locality in Ethiopia.
- Support to zoom in and zoom out detail map features.
- Support to print selected area.
- Gather complaints and feedbacks from the public online.
- Enable to initiate customer for online sale order.
- Support to register customers online.
- Track to view users' information.
- Support to approve sales by Marketing Officer.
- Support to edit users' information.
- Support to select geo-information data by category.
- Support to download resources.
- Authenticate users according to their roles.
- Support to upload resources of EMA.

3.2 Non-Functional Requirements

This is a software requirement that describes not what the software will do but how the software will do it. Program needs to be able to run, but which doesn't relate to the actual functionality of the program [12].

Performance of the OSDOGID system emphasized on the challenges of hard disk space, speed, memory and bandwidth requirements. Images of raster data consume hard disk space and bandwidth to publish on the web service. To avoid these constraints we have changed the raster (.tif) files to small sized (.pdf) files for easily downloading the spatial data. The web service is configured with the following performance issues.

1. *Pooling:* This service should be pooled and used repeatedly by many clients simultaneously with maximum number of 200,000 instants and minimum instant of 1.
2. *Time-Outs/Response time:* the client can use the web service with maximum time of 600 seconds and waits to get service in 60 seconds with an idle instance can be kept running at 1800 seconds.
3. *Memory Requirement:* the system is expected to be run on a powerful server that can be accessible by many clients to search crop and the desired area for printing in large format printer. This was suitable in the intranet system.

Sixteen use cases were identified in the system with three Actors for geo-information service delivery processes as shown in Figure 1.

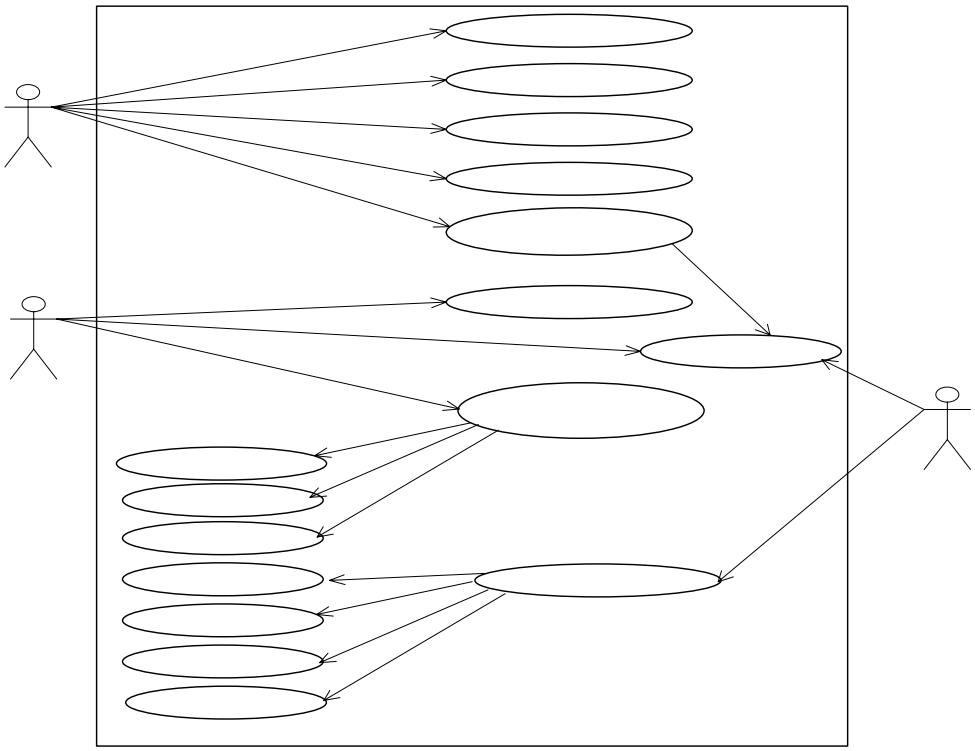

Figure 1: Use case diagram of the new system

Table 1: Sample Use Case Description for Initiate Order

Use Case Id	**UC-02**	
Use Case Name	**Initiate Order**	
Description	This use case shows how the client initiates map resources	
Participating Actors	Customer	
Entry Condition	The customer must activate the system home page	
Success Scenario	Step	Action
	1.	The customer navigates his preference links.
	2.	OSDOGID displays a map product category page.
	3.	The customer selects product by category.
	4	The OSDOGID displays the selected product
	5. A	The customer clicks initiate order. [Alternate A]
	5. B	The customer clicks on “cancel” button. [Alternate B]
Exit conditions		The customer submits his request.
Alternate flow A.	6. A	OSDOGID displays the “New Customer Registration” page.
	7. A	The customer fills the registration form

	8. A	The customer clicks the "Submit" button
	9. A	OSDOGID displays successful massage.
Exit conditions		The customer clicks "ok" button.
Alternate flow B.	6. B	OSDOGID returns to metadata page.
Exit conditions		User can view another product or exit to the home page and get

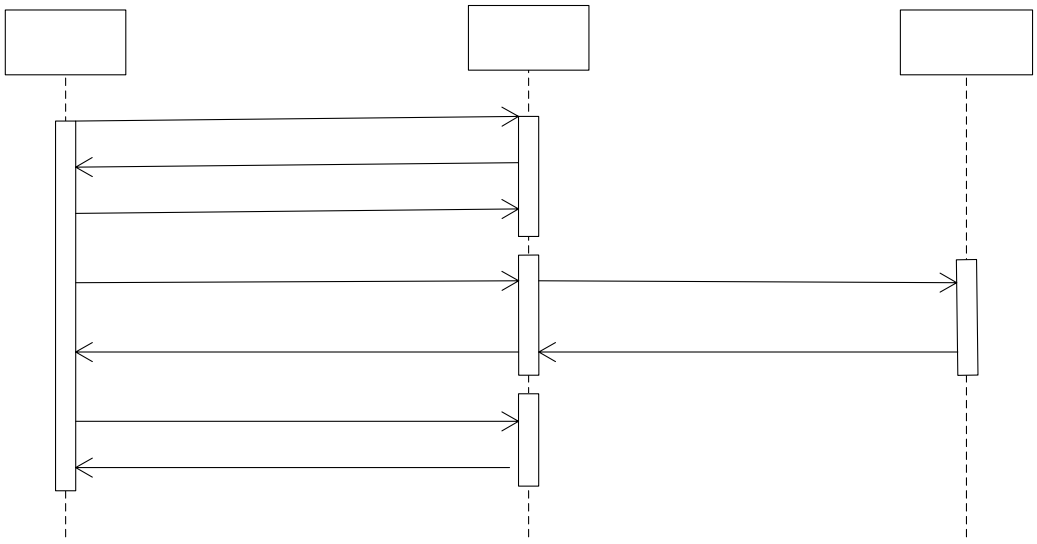

Figure 2: Sample Sequence diagram of map selection by category

4. System Design

After the determination of the requirements, it is the design that follows. The design is all about stating the design goals of the system and subdividing the system into smaller parts so as to tackle the problem in a modular approach. The output of this phase includes description of each subsystem and the deployment of the subsystems.

4.1 Architecture of the System

It is the architecture that determines the type of interactions that the components are going to have.

To provide online service including intranet to be used in EMA, multi tiered (5-tiered) architecture was designed based on the requirement of the agency. The system consists of the following servers:

- *Geo-database Server and Web Server*: to support spatial searching offline in EMA and online geo-information of Ethiopia.
- *Mobile Information server:* is a server in the Commercial Bank of Ethiopia (CBE) that is responsible to manage payment information of customers and alerts marketing officers by mobile banking system which is already launched by CBE.
- *GSM Server:* is a server of ethio telecom responsible to facilitate communication between customers of EMA and Marketing Officers with CBE by mobile banking system.
- *FTP Server:* responsible to upload thematic data from OSDOGID system.

Marketing Officer's Mobile: is a dedicated mobile used to receive information of payment dueled from customers by mobile banking system. The officer can approve customers to download resources based on the payment requirement.

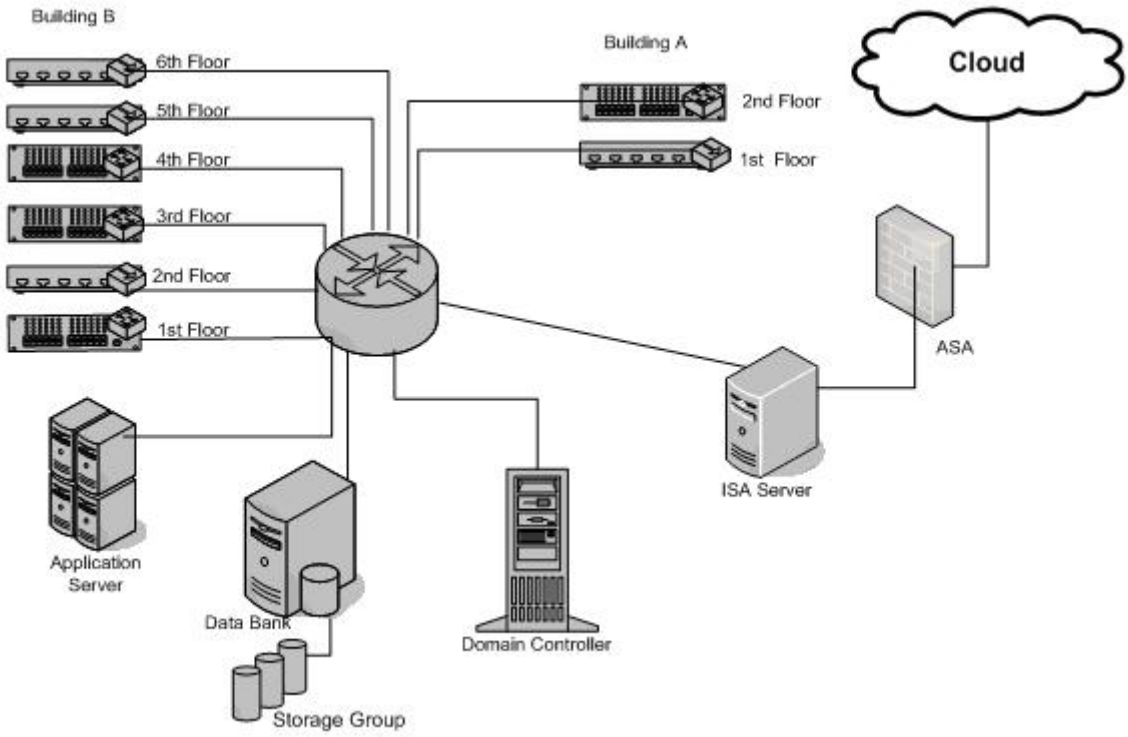

Figure 3: Existing Architecture of EMA

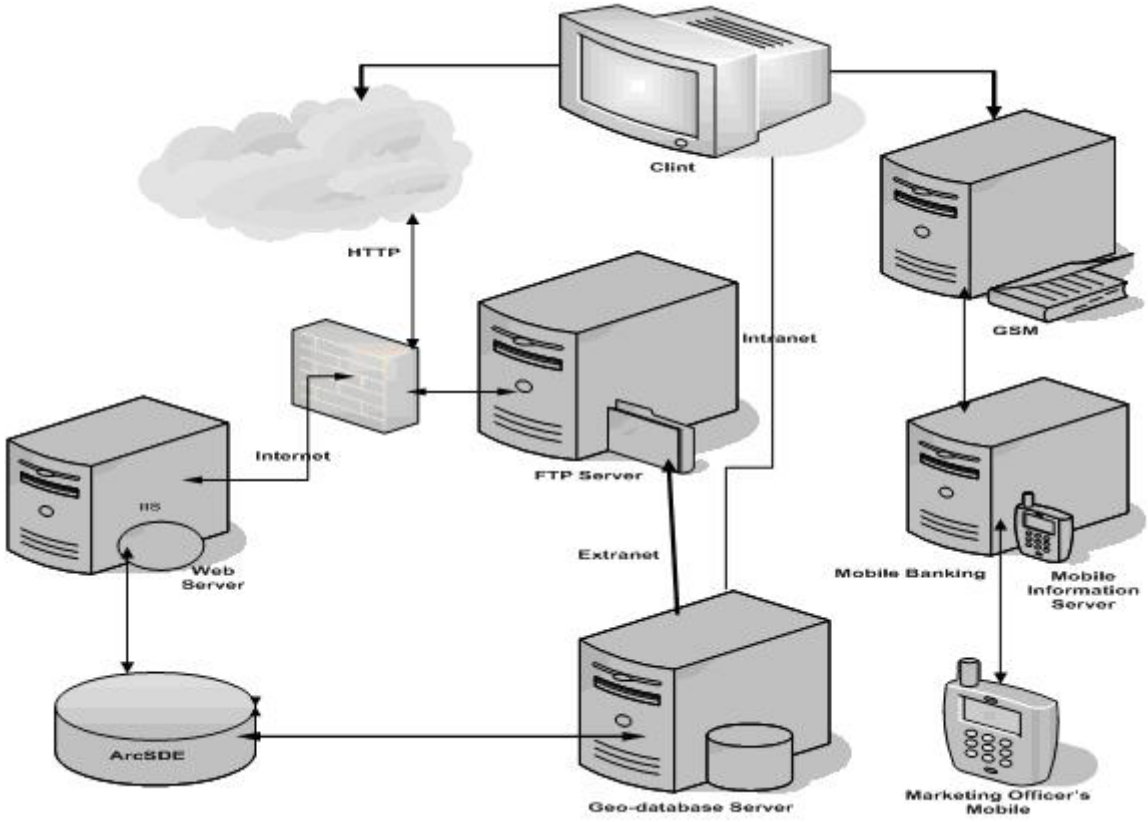

Figure 4: New System Architecture of EMA

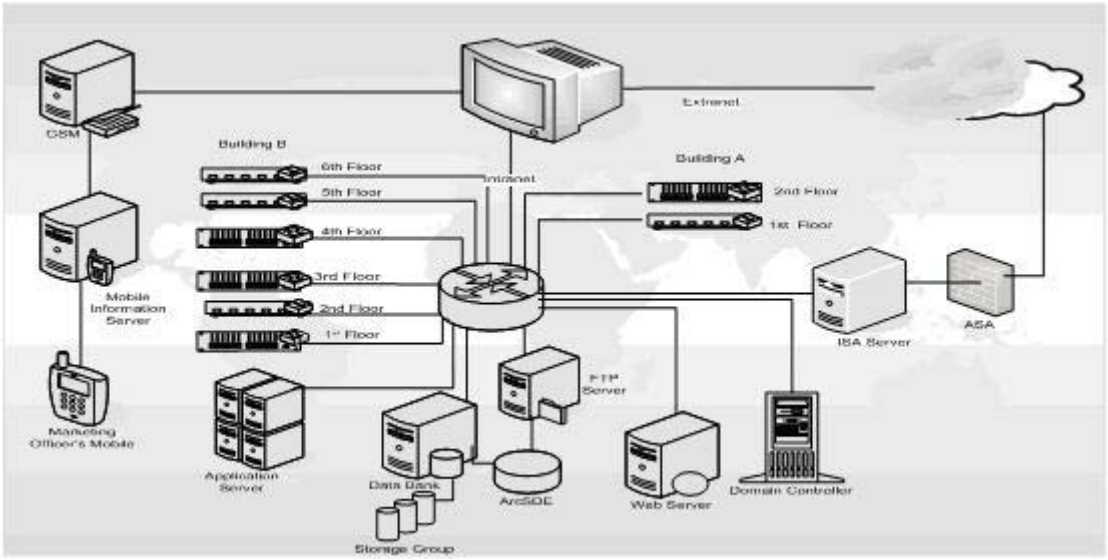

Figure 5: Integrated Architecture of EMA

4.2 Subsystem Decomposition

In order to simplify and minimize complexity of the solution domain, the system has been divided into three subsystems. These are "Spatial Searching Subsystem", "Information Subsystem" and "Administration Subsystem". The Spatial Searching Subsystem is responsible for providing geo-information data by selecting specific area and is also responsible to edit map, zoom in/zoom out features and print the selected map(s) in large, medium and small size format. The information subsystem is responsible in providing the detailed information of EMA's products and services and the Administration subsystem enables the administrator to manage user accounts.

5. Prototype

This subsection of the implementation phase was about how to go through this work by developing a web portal. The portal has sixteen interfaces which were categorized into five main interface components. The system consists of the search area, Initiate Order, Customer Registration, Manage Customer Information, and Upload Resources.

The portal consists of both dynamic and static pages. The dynamic pages are integrated with the database at the backend that can be modified dynamically when new information is inserted by appropriate users such as customer information, order management, resource management, etc., whereas, the static pages are persistent information that are not integrated with the database such as: home page, search area, map products, contact us, our site, directorate profile, view information, map viewer, etc.

6. Discussions

To make sure that this system is feasible and successfully implemented in EMA, we tried to assess the existing enablers and opportunities found in EMA and in its stakeholders from the collected data. The majority of respondents ascertained that online service delivery system is important for time and cost minimization. The developed system is also valuable to the geo-information seeker society in prevailing standard at national level. All stakeholders have Internet connections with more than 2Mbps bandwidth whereas EMA has currently 8Mbps. Because of these enablers and opportunities, we believe that the developed system is adroit to work effectively and efficiently without bottleneck of bandwidth in data sharing activities at national level.

We have investigated the existing process of EMA related to service delivery system. Even though there were efforts in production of topographic and thematic maps, we identified the cause of customer dissatisfaction from 16 stakeholders by using purposive and cluster sampling technique. To

mitigate the gap, we developed online service delivery system for the agency that could play a positive role to contribute the sustainable development in Ethiopia.

We have also developed a web based portal that gives detailed information about the availed geo-information data in EMA. The system can serve offline and online service delivery system of geo-spatial data of Ethiopia. However, since we have used web technology, the location of users is not limited to the country. Users of web-based systems could be anywhere in the world. This means those Ethiopian and non-Ethiopian living in the country as well as outside the country can be served by the system.

We believe that this project would have a remarkable solution for data integrity and mitigating economic loss by avoiding duplication of work at national level; because, production of geo-information data at different sectors is prone to data incompatibility due to accuracy and projection parameter errors.

To meet the general objective of this research, results were summarized for each specific objective that shows advantage of the new system comparing to other systems and its result is shown in Table 2.

Table 2: Summary Process on Development of OSDOGID to Meet Each Objective

No.	*Specific Objective*	*Method used*	*Advantage of OSDoGID over other systems*	*Result found*
1.	To integrate SSE for visualizing and printing of map details that can be used as intranet and extranet.	▪ Similar literatures were reviewed. ▪ ArcGIS server was used for publication and for intranet ▪ Web based portal was developed for extranet	▪ Detail gazetteer data was integrated for specific area searching ▪ Able to edit and print in three map size format ▪ Can be zoomed in/out for detail features ▪ Used for intranet and extranet	Automated SSE used for intranet and extranet was nicely developed.
2.	To build geo-information database that comprises spatial and non-spatial data.	▪ 1:250,000 scale of Mosaic topo maps, road networks, gazetteer data, spot satellite imageries and political map of Ethiopia were used as input data ▪ MS SQL server used for non spatial data handling ▪ ArcSDE used for spatial data handling	▪ Seven volumes of geographic name database were integrated ▪ Full coverage of 1:250,000 scale topographic maps of Ethiopia ▪ Full coverage of spot satellite imageries were integrated for the whole Ethiopia ▪ Metadata of EMA production and services were clearly described.	Robust dynamic pages
3.	To develop prototype and illustrate system architecture	▪ Primary and secondary data were collected from EMA and their stakeholders. ▪ C# and Silverlight were used at front end ▪ Mobile banking was used for online sales to illustrate architecture	▪ Prototype developed based on requirements of EMA's and its stakeholders' ▪ Used to select multi-purpose geographic layers of Ethiopia ▪ Full information about products and services given by EMA ▪ Online customer registration	User friendly online service delivery of geo-information data was developed for EMA

7. Conclusion and Future Work

This paper contributes to improve geo-information data delivery system of Ethiopian Mapping Agency. It enables to provide geo-information data online to the public without physically coming to EMA. This will help customers to buy geo-information products and services without wastage of time and money. Furthermore, it has established full information about the availed products and services of the agency so that governmental and non-governmental sectors do not need unnecessary resource wastage of time, material and human expertise due to duplication of work at national level. Prevailing of standardization was also another advantage of this work.

We believe that the application of the proposed system OSDOGID will change EMA's service delivery one step ahead. The geo server part of the system can serve as Intranet with the existing LAN without the need of Internet. EMA's staff can use the system to search any area located in Ethiopia. The user can crop the desired area and print if large format printer is attached with his/her computer. Business development experts could be able to use the system to visualize the demand area of a customer by LCD projector or by standalone flat monitor. Customers can easily identify consecutive topographic maps to be purchased by zooming in and zooming out the detail map features.

This work can be considered as gate opener on dissemination of geo-information data in Ethiopia by showing only 1:250,000 scale topographic maps. Other geo-information data is not addressed in this work due to time and resource limitations. So, the area is still open for further research on how to deliver very huge data to geo-information seekers.

E-commerce is not matured in Ethiopia. So service charge handling is also a gap in the area. Future research work on this area could come up with solutions for the above gap and related ones.

It is our recommendation for the agency involved in geo-information data production and the ENSDI members to integrate their own spatial data in one central database. To avoid duplication of work and wastage of resources at national level, ENSDI members can utilize and upgrade this work for data sharing, integration and standardization issues of geo-spatial data.

References

[1] Okar Service Plc., "Experience the Power of the Intergraph 2013 Geospatial Portfolio, Geospatial Portal", Addis Ababa, Ethiopia.

[2] The Atlas of Canada, http://www.atlas.nrcan.gc.ca, published by Natural Resources Canada, last accessed on 12th Jan 2013.

[3] Maps of India.com, retrieved from, www.mapsofindia.com, last accessed on 3rd February 2013.

[4] Caribbeean-on-line.com, www.caribbean-on-line.com, accessed on 12th Jan 2013.

[5] Google Maps, retrieved from http://www.maps.google.com, accessed on 2nd February 2013.

[6] Yahoo Maps Local, retrieved from, http://maps.yahoo.com, accessed on 22nd Feb 2013.

[7] Maps of the World, retrieved from, www.mapsofworld.com, accessed on 2nd Feb 2013.

[8] Standfords, retrieved from, www.stanfords.co.uk, accessed on 21St Feb 2013.

[9] Microsoft's Mappoint, www.mappoint.msn.com, accessed on 2nd Feb 2013.

[10] Perry-Castañeda Library, Map Collection from University of Texas at Austin, www.lib.utexas.edu, last accessed on 18th Feb. 2013.

[11] UNESCO, the United Nations Educational, Scientific and Cultural Organization, http://www.portal.unesco.org, last accessed on 21st Feb 2013.

[12] Mark Hansen, Staurt Madnick and Michael Siegel, "Data Integration Using Web Services", MIT – Sloan School Management Working Paper No 4406-02, CISL Working No. 2002-14, 2002.

Traffic Accident Analysis from Drivers Training Perspective Using Predictive Approach

Hiwot Teshome
hewienat@gmail

Tibebe Beshah
School of Information Science, Addis Ababa University, Ethiopia
tibebe.beshah@gmail.com

Abstract

Road traffic accidents are among the top leading causes of deaths and injuries, especially in the developing world. Ethiopia in particular experiences the highest rate of such accidents. One of the solutions to reduce the problem of traffic accident is finding the causes through research studies. In this work, factors for the relationship between driver's training and road traffic accident are identified using data mining classification technique. Using WEKA as a tool, different rules are generated so that different domain experts get new insights related to road traffic accidents.

Keywords: Accident; Data mining; Classification; WEKA

1. Introduction

Traffic accident is a special case of trauma that constitutes a major cause of disability and untimely death [1]. As reported by WHO [2], around 1.24 million people die each year as a result of road traffic crashes. Without action, road traffic crashes are predicted to result in the deaths of around 1.9 million people annually by 2020.

Traffic accident is the result of multiplicity of factors and it is often the interaction of more than one variable that leads to the occurrence of accident among which are driver, road and traffic characteristics [3]. For the last five years records have shown that over 87% of all road accidents in Ethiopia are attributed to driver error [4].

Driver factors in road traffic accidents are all factors related to drivers and other road users. This may include driver behavior, visual and auditory acuity, decision making ability and reaction speed. Drug and alcohol use while driving is an obvious predictor of road traffic accident, road traffic injury and death. Speeding, travelling too fast for prevailing conditions or above the speed limit, is also a driver factor that contributes to road traffic accidents [5].

The capital city of Ethiopia, Addis Ababa, shares 65% of the total accident in the country. Pedestrians are the most vulnerable ones in Addis Ababa; over 81% of accident fatalities are of accident type "car hit pedestrian". Moreover, 81% of crashes in Ethiopia are attributed to driver error [6].

In an attempt to prevent road accidents one role that can be played is finding the factors through research. Researches are done using different tools and technologies. In line with this, data mining is one of these research technologies which is growing rapidly through time [7].

In order to combat this problem, various road safety strategies have been proposed and used. Research on road safety has been conducted for several years, yet issues like drivers training, driver experience still remain undisclosed and unresolved. Besides that, research on road safety using data mining tools has been conducted for several years also. The aim of this research is identifying the relationship between drivers training and road traffic accident using classification technique of data mining.

The previous researches so far, directly or indirectly, try to identify the factors of road traffic accidents on the side of road, drivers and so on but even if the main factor fall on the shoulder of the driver as per the author's knowledge from the

training perspective of drivers is not considered yet. If that is the case mining from the driver's training point of view is worthwhile.

Through the research attempt has been made to answer the following research questions:

- ✓ What are the determinant attributes from driver's training related factors that have impact on the occurrence of traffic accident?
- ✓ What interesting patterns or rules can be generated?
- ✓ Which data mining technique performs well in developing a model that can explain the relationship between drivers training and road traffic accident?

The primary data were gathered using questioners. 1000 questioners were distributed for different drivers that have different status, knowledge, occupation, age and gender. To have a better understanding of the problem the sampling was taken from different associations that work on drivers like taxi drivers association groups, heavy truck drivers' association and the data available from Ethiopian Road Authority. As per the author's knowledge, since this kind of research was not done before literatures were not used as a baseline to develop a questionnaire. Because of that a questionnaire was developed using the data that was analyzed during data understanding and domain understanding and also with the help of experts who work in the driver training area. As a pilot test 20 questionnaires were distributed for 20 drivers; 10 males and 10 females with the average age from 20-45 and their feedback was collected and helped the questionnaire to be refined again.

To prepare the data collected from the questionnaire into a suitable form for data mining analysis, data preparation activities like attribute selection, data cleaning, and data formatting have been done. After automating the data on Excel sheet, the first task was selecting attributes that have direct relationship with the objective. After the feature selection of the data using WEKA tool, the preprocessing stage which includes cleaning inconsistent and incorrect data, incomplete records, predict missing values, correct erroneous and anomalous data is performed. After the pre processing steps the data contain the selected predictors in a format which is ready to use for modeling.

In this research classification of accident occurrence is done attempting different classification techniques like J48, PART and Naïve Bayes to predict the occurrence of accident. Among the tools WEKA software is used to conduct the research.

2. Related Work

Many data mining application researches on the analysis of road traffic accident have been done globally and a few locally. Reviewing this related works that were done using data mining tools and techniques at different place and time in the same problem domain gave the author an in-depth insight for this research.

As shown by Tibebe [8], data mining is one technique in order to conduct a research on historical road traffic accidents to investigate analysis accident severity where the data comprising a dataset of 4,658 accident records at Addis Ababa Traffic Office. In this work various classification models using the decision tree technique by applying Knowledge SEEKER algorithm of the Knowledge STUDIO data mining tool were built to help in decision-making process at the traffic office. The methodology adopted had three basic steps namely data collection, data preparation, and model building and validation.

The model classifies accident severity into four classes as fatal injury, serious injury, slight injury and property-damage. Accident cause, accident type, road condition, vehicle type, light condition, road surface type and driver age were identified as the basic determinant variables for injury severity level. The classification accuracy of the decision tree classifier was found to be 87.47%.

Zelalem [9] conducted a data mining research to classify driver's responsibility on a given accident in

Addis Ababa. The research uses decision tree and multilayer perception (MLP) neural network data mining techniques to analyze the accident data. The study focuses on predicting the degree of driver's responsibility for car accidents and identifying the important factors influencing the different levels of responsibility using the road traffic accident dataset of Addis Ababa Traffic Control and Investigation Department (AARTCID).

The researcher used WEKA data mining tool to build the decision tree (using the ID3 and J48 algorithms) and MLP (the back propagation algorithm) predictive models. Rules representing patterns in accident dataset have been extracted from the decision tree indicating important relationships between variables that influence driver's degree of responsibility such as age, license grade, level of education, driving experience, and other environmental factors. According to the author, the accuracies of the models were 88.24% and 91.84% respectively. In addition the research reveals that the decision tree model is found to be more appropriate for the problem under consideration.

On the other hand Getnet [10] investigated the potential application of data mining tools to develop models supporting the identification and prediction of major driver and vehicle risk factors that cause RTAs. The researcher used WEKA tool to build the decision tree (using the J48 algorithm) and rule induction (using PART algorithm) techniques. Performance of the J48 algorithm was slightly better than that of the PART algorithm. The license grade, vehicle service year, vehicle type, and experience were identified as the most important variables for predicting accident severity.

In this research determinant factors of drivers and road that cause traffic accident are identified which will help health policy makers in planning health programs and it will also support the Traffic Control Division of Addis Ababa in taking proper action such as revising the existing traffic rules against vehicle accidents.

As pointed out by Tibebe and Hill [11], developing adaptive regression trees to build a decision support system is one of the methods to handle road traffic accident. The study focused on injury severity levels resulting from an accident using real data obtained from Addis Ababa traffic office.

Yang *et al.* [12] used Neural Network approach to detect safer driving patterns that have less chances of causing death and injury when a car crash occurs. They performed the Cramer's "V" Coefficient test to identify significant variables that cause injury, therefore, reduced the dimensions of the data for the analysis.

Sohn and Lee [13] applied data fusion, ensemble and clustering to improve the accuracy of individual classifiers for two categories of severity (bodily injury and property damage) of road traffic accident. They applied a clustering algorithm to the dataset to divide the data into subsets of data, and then used each subset of data to train the classifiers (Neural Network and decision trees).

Table1: Summary of Researches

Author	*Objective*	*Method*	*Key Result*
Tibebe (2005)	To investigate analysis accident severity	decision tree technique	Accident cause, accident type, road condition, vehicle type, light condition, road surface type and driver age as the basic determinant variables for injury severity level.
Zelalem (2009)	To classify driver's responsibility	decision tree and multilayer perception (MLP) neural network	Age, license grade, level of education, driving experience influence driver's degree of responsibility

Author	*Objective*	*Method*	*Key Result*
Getnet (2009)	To identify and predict major driver and vehicle risk factors that cause RTAs	decision tree	The license grade, vehicle service year, vehicle type, and experience are most important variables for predicting accident severity.
Tibebe & Hill (2010)	Mining Road Traffic Accident Data to Improve Safety	adaptive regression tree	road-related factors are identified

As shown in Table 1 different scholars did their researches on road traffic accident using data mining techniques and they have proposed their solutions that they identified as factors for the occurrence of an accident. On the contrary, this research tries to identify factors of traffic accident with respect to the training view of drivers; that is to mean those attributes that are generated as factors on drivers training instead of being general factors for road traffic accident.

3. The Proposed Solution

In addressing the above mentioned problem the data mining experiment process and the resulted prototype are presented in this section.

The underlying problem that initiated this project is the fact that road traffic accidents are among the top leading causes of death and injury of various levels in Ethiopia. During the document analysis it was found that the data source for this research could not be the traffic accident data kept at AARTCID because majority of the data that is collected from did not correspond to the research objective. In order to come up to the solution the first thing that has been done was preparing and distributing questionnaire to use it as a data set for this research.

As can be found from data of Ethiopian Road Transport Authority the total number of vehicles that are found in all regions of Ethiopia are 474,143 and in Addis Ababa in particular is 219,217.

From all types of vehicles, public transport, automobile, trailers dry (*Derek chenet*) and trailer liquids (*fesash chenet*) have been selected with their respective values of 48%, 32%, 12%, and 8% which is taken as the total population for this research.

Using proportional stratified sampling technique 1000 samples were taken to distribute the questionnaire for the four categories of vehicle type as shown in Table 2.

Table 2: Number of Distributed questionnaires

Type of car	*Number of questionnaires distributed*
Automobile	300
Public Transport	500
Derek chenet	100
Fesash chenet	100

After the distribution and collection of questionnaire the second task was encoding it into an Excel sheet so that it will get prepared for data the cleaning step. The encoded data set that is generated from the questionnaire have 17 columns (or attributes) of text and numbers. On the data cleaning steps noises, missing values, and redundant values were treated. Finally 10 attributes remained through feature selection for the modeling purpose as shown in Table 3.

Table 3: Final Selected attributes with their descriptions

No.	*Attribute Name*	*Description*
1.	DriverAge	Age
2.	DriverSex	Sex
3.	DriverEducational status	Educational level of the driver
4.	DriverYear of licence	The time to get the license
5.	DriverCartype	Category of the driver's car
6.	DriverExperience	Driving experience
7.	OccurredAccident	Occurred accident
8.	DriverBesttraning	Training that the driver chooses
9.	Type of traningForBehaviour Change	Type of the training which cause behavioral change
10.	TrainingTypeToCauseInfluence	Type of training to influence accident

During modeling several experiments are made using the J48, PART and Naïve Bayes to come up with a meaningful output with two test options. Major factors for the relationship between accident and training are identified and rules are generated using J48 decision trees and rule induction (PART algorithm). The comparison of the models using WEKA's experimenter showed that PART slightly outperforms Navies and J48 algorithms with an accuracy of 92.4%, 91.2% and 90.9% respectively. The results of the experiments carried out in this research show driver age, driver educational status, and car type are determinant factors for drivers training which can help to predict accident occurrence.

A prototype is developed using rules from PART and J48 algorithms. This helps to have a better understanding from the two rules generated. From both algorithms different rules were generated and the rule which can determine driver related factors can be shown as fines and new rules. Making the prototype to predict differently as per the algorithm helps to analyze the output in different perspectives.

4. Discussion

Three experiments are conducted using 10 attributes that are filtered during the data preparation phase with the selected algorithms (J48, PART, Navies). Based on the selected algorithms different experiments have been done using two test options (tenfold cross validation 66% percentage split) and different results were obtained.

In the series of experiments, evaluation of models is done based on accuracy of models and confusion matrix. All the three classifiers performed well and almost similarly with respect to the number of correctly classified instances. Evaluating the models helps to identify a relatively better model. Table 4 shows the evaluations of the three algorithms compared with respect to these efficient checking algorithms.

Table 4: Summary of the result of the three experiments

No.	*Classification Models(classifiers)*	*Number of Correctly classified instances*		*Accuracy in Percentage*	
		10 fold cross validation	*66% percentage split*	*10 fold cross validation*	*66% percentage split*
1	J48	912	304	91.2%	89.4118
2	PART	924	304	92.4%	89.4118
3	Naïve Bayes	909	303	90.9%	89.1176

As mentioned earlier we can clearly see from Tables 5 and 6 that there is a relative better model prediction in the case of PART incorrectly identifying the dataset. The ROC area of the PART indicates 0.93265 with small significance difference with ROC area in Naïve Bayes indicates 0.9334 and also much better than the J48 ROC area. The overall model accuracy of PART is 92.4% which shows it has better prediction than the accuracy of J48 and Naïve Bayes. For the given data set under this study PART with 10 fold cross validation has shown better accuracy.

Table 5: Confusion matrix Predicted category

Actual Category	*Predicted category*		
	J48	*PART*	*Naïve Bayes*
Yes	782	783	758
No	66	55	45

Once the model has been trained and tested, we need to measure the performance of the model. In data mining comparison is done by using accuracy and ROC (Receiver Operating Characteristic).

Table 6: Comparison of Algorithms using ROC

Algorithm Efficiency Evaluator	*J48*	*PART*	*Naïve Bayes*
ROC	0.8878	0.9365	0.9334

A prototype is developed using rules from PART and J48 algorithms. This helps to have a better understanding from the two rules generated. From both algorithms different rules were generated and the rule which can determine driver related factors were fines and new rules. Making the prototype to predict differently as per the algorithm helps to analyze the output in different perspectives.

5. Conclusion and Future Work

This research attempted to study the possible application of classification techniques of data mining to predict the relationship between drivers training and road traffic accident. The study followed CRISP DM model and WEKA data mining tool has been used to implement the Naïve Bayes, J48, and PART algorithms.

The results of the experiments carried out in this research show driver age, driver educational status, car type are determinant factors for drivers training which can help to predict accident occurrence. A prototype is developed using rules from PART and J48 algorithms. This helps to have a better understanding from the two rules generated. From both algorithms different rules were generated and the rule which can determine driver related factors were fines and new rules. Making the prototype to predict differently as per the algorithm helps to analyze the output in different perspectives.

In general, encouraging results are obtained by employing both decision tree and rule induction techniques and the rules generated by J48 and PART algorithms are easily understandable by subject experts. Thus, the results obtained in this research have proved the applicability of data mining in road traffic accident preventing and controlling activities especially on the area of drivers training. More specifically it will provide valuable help in developing new methods, rules and regulations to increase the quality of the drives training now a day.

This work focused only on drivers training using classification techniques. The next step will be applying different techniques like clustering and associating data mining techniques integrating with driver information for better predictions, and to find interactions between the different attributes.

References

[1] Dipo T. Akomolafe Akinbola Olutayo, “Using Data Mining Technique to Predict Cause of Accident and Accident Prone Locations on Highways”, American Journal of Database Theory and Application 2012, 1(3): 26-38.

[2] WHO, World report on road traffic injury prevention, Switzerland, Geneva, 2013.

[3] William Eckersley, Ruth Salmon, and Mulugeta Gebru, “Khat, Driver Impairment and Road Traffic Injuries: A View from Ethiopia”, Jerusalem Children and Community Development Organization, Addis Ababa, Ethiopia, January 2010.

[4] Ethiopian Road Authority, "How safe are Ethiopian roads?", Unpublished road safety report, Addis Ababa, Ethiopia, 2005.

[5] Road Traffic Accidents in Nigeria: A Public Health Problem, 2013.

[6] A. Persson, "Road traffic accidents in Ethiopia: magnitude, causes and possible interventions", April 2008.

[7] The CRISP-DM consortium, "Step-by-step Data Mining Guide Available", August, 2000.

[8] Tibebe Beshah, "Application of data mining technology to support RTA severity analysis at Addis Ababa Traffic Office", Unpublished Master's Thesis, Addis Ababa University, 2005.

[9] Zelalem Regassa, "Determining the degree of drivers' responsibility for car accident: the case of Addis Ababa traffic office", Unpublished Master's Thesis, Addis Ababa University, 2009.

[10] Getnet M., "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 accidents: the case of Addis Ababa city", Unpublished Master's Thesis, Addis Ababa University, 2009.

[11] Tibebe Beshah and Shawndra Hill, "Mining Road Traffic Accident Data to Improve Safety: Role of Road-related Factors on Accident Severity in Ethiopia", 2010. Available at www.ai-d.org/pdfs/Beshah.pdf, Last accessed on March 13, 2013.

[12] Yang, W.T., Chen, H. C., and Brown, D. B., "Detecting Safer Driving Patterns by A Neural Network Approach", ANNIE '99 for the Proceedings of Smart Engineering System Design Neural Network, Evolutionary Programming, Complex Systems and Data Mining, Vol. 9, pp. 839-844, Nov. 1999.

[13] Sohn, S. Y. and Lee, S. H., "Data Fusion, Ensemble and Clustering to Improve the Classification Accuracy for the Severity of Road Traffic Accidents in Korea", Safety Science, Vol. 4, Issue1, pp. 1-14, February 2003.

Cloud Computing Security Framework for Banking Industry

Meskerem Alemu
emamenate@yahoo.com

Abrehet Mohammed Omer
Addis Ababa Science and Technology University, Ethiopia
abrehet@gmail.com

Abstract

Cloud computing is a prospering technology that most financial organizations are considering for adoption as a cost effective strategy for managing Information Technology (IT). However, financial organizations such as banks still consider the technology to be associated with many business risks that are not yet resolved. Such issues include security, privacy, legal, compliance and regulatory risks. As an initiative to address such risks, cloud security framework and bank enterprise framework have been proposed. However, the proposed framework focuses more on technical control and doesn't incorporate the overall administrative, legal and compliance control on cloud computing services. Further they are not also considered specific solutions for the bank industry compliance requirement and neglect some major bank information security issues. Due to lack of professionals and adequate frameworks in the area, the issue is getting scaled up to become a severe problem. The main objective of this paper is, therefore, to propose Cloud Computing Security Framework for the banking industry.

The study has been conducted on Banking Industry through systematic literature review on cloud computing standards, policy and best practices coupled with interview as methods of data collection. The survey result helps for identifying professionals thought on the subject and major pillars to propose new framework. Besides, the Sherwood Applied Business Security Architecture (SABSA) framework was used as a guide for designing the newly proposed cloud computing security framework for the banking industry focusing on architects view from six perspectives.

The proposed framework aggregates different temples: Risk Matrix Template, Control Domain Template, Compliance Matrix Template, and Security Strategy/Major, that help banks come up to solutions for measuring risk, compliance and setting suitable security major.

Keywords: Cloud computing; Banking industry; Metrics; Security; Threats; Vulnerability

1. Introduction

In order to satisfy customer need, banks use Information Technology (IT) services. However, traditional IT computing technology until now, has typically been a costly hurdle for financial institutions, particularly those in emerging markets where developing customized solutions or investing in advanced banking platforms has either been unfeasible or the result has seen too many failures, too many resources used and too much time wasted [1, 2].

Currently, cloud computing technology has brought the idea of storing and managing data on virtualized servers so that, applications, individuals and organizations around the world can have the ability to connect to data and computing resources anywhere and anytime. However, banks cannot afford the risk of a security breach since security of financial, personal data and mission-critical applications are paramount. Moreover, financial compliance regulations require that, data should not be intermixed with other data types, on shared servers or databases. Therefore, to move banks into cloud computing environment it is essential that security challenges in relation to regulatory policy, compliance and standards must be addressed primarily.

This research attempts to answer the research question:

What are the suitable security components to propose new Security Framework for the banking industry to adopt cloud computing services?

2. Related Work

Temenos Enterprise Framework Architecture (TEFA) (see Figure 1): The framework is focused in providing information to implement evolving product and service portfolios without disrupting banking operations [3]. In the framework, Security Management System (SMS) is incorporated as one component. The framework addresses one security domain issue, that is, identities and accesses management level. Since, cloud computing related to banking services is a broad concept other security related to administrative, technical and physical control should be addressed in broad manner.

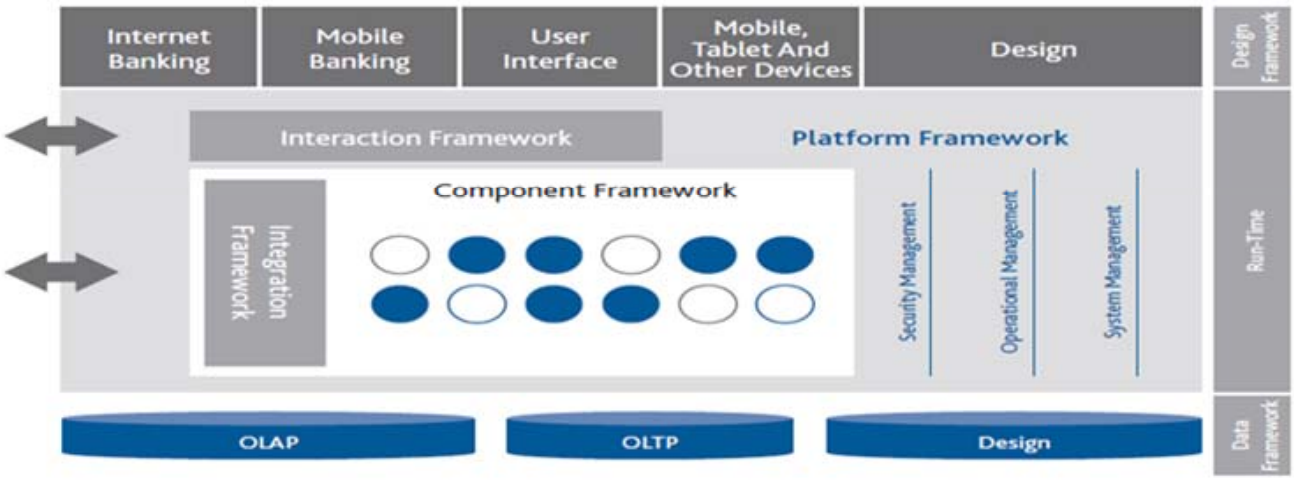

Figure 1: Temenos Enterprise Framework Architecture

IBM Security Overview on Cloud computing: in viewing of security in cloud computing, IBM proposed Security Framework (see Figure 2) [4]. IBM security framework is composed of five main components: people and identity, data and information, application and process, network, server, end point and physical infrastructure. IBM framework identifies main component in general. However, the framework is not designed basing cloud computing architecture and banking business security requirements. It doesn't address the security majors that should be considered at each level of cloud services model (i.e., SaaS, PaaS, IaaS) and deployment mode. In addition to this, regulation standards and compliances and the way to adopt cloud computing for banking services is not addressed in the framework [5].

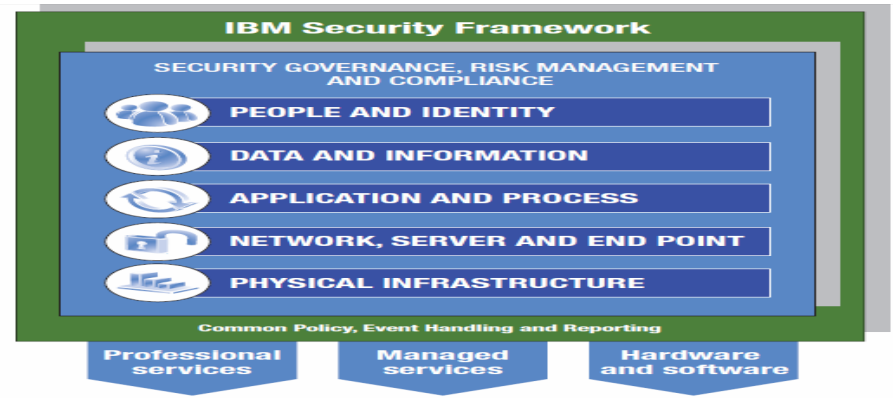

Figure 2: IBM Security Framework

In a paper "Security Issues of Banking Adopting the Application of Cloud Computing", the authors stated that, cloud technology makes possible to reuse IT resources for banks very efficiently. In order for banks to adopt this technology the paper states that two primary challenges, namely, security and regulatory compliance should be addressed, wherein financial institutions must select the right service, deployment, and operating models to address security and compliance concerns. The paper discusses security risk on cloud which includes, regulatory compliance, data location, data segregation, recovery investigation support and long term liability, leakage of data, database and server security for the system. The paper suggests that banks to use hybrid cloud since it would have private cloud for highly secured transaction [6].

In view of the advantages of cloud computing, the working group, Reserve Bank of India analyzes and recommends the support of non-financial services to ideally explore cloud computing as to gain more experience. The working group recommends that there is need of more research and development in the area of cloud governance and audit, cloud management and cloud securing technology for which the banking industry and the software industry could take the initiative so that regulatory authorities can make use of the same security framework standards [7, 8].

As different security standards and enterprise security architecture guidance provider indicates security framework shall provide much more to the business requirements than pure "security and control" [9, 10, 11, 12]. They propose enterprise security framework to address the entire three security control domain Technical, Administrative and Physical /environmental control.

Based on these facts we can say that previous works on security framework do not cover the overall security issues on cloud architecture and banking business operation.

Therefore, this paper addresses this gap by developing security framework considering banking business security requirements and cloud computing technical, physical and administrative control requirements.

3. The Proposed Framework

In order to come up with integrated security framework solution, through adopting enterprise security architecture tool of Sherwood Applied Business Security Architecture (SABSA) of architects view, we designed a cloud computing security framework for the banking industry (see Figure 3).

Cloud Model - (What and Who)

In the proposed framework of Figure 3, the system that needs to be secured in cloud system is represented with cloud architecture with main security item components such as Physical, Network, Compute, Storage, Application and Data. Cloud Services Model (SaaS, PaaS and IaaS) and Cloud Deployment Model (public, hybrid, community and private) are also integrated into the Cloud model as they are main parts of cloud architecture. For answering *who*, cloud actor is specified within cloud model. Specific actors are identified which are, bank (cloud consumer) and cloud services provider (CSP).

Security Model-(Why and Where)

Security model layers of the framework are defined with two security solutions package, Risk Matrix for answering *why* and Control Domain for answering *where*.

Risk assessment needs to occur before an enterprise enters into a cloud computing arrangement to help avoid surprises and minimize the costs of implementing and maintaining controls. More so, implementing too many controls may not be the best risk-mitigation approach, because the benefit from implementing controls should outweigh the cost. Therefore, other risk-mitigation measures such as transferring, avoiding or accepting the risk are worth considering as well. In order to address this issue through following ISO 3100 risk management guideline, we developed Risk Exposure Majoring Template (see Figure 4).

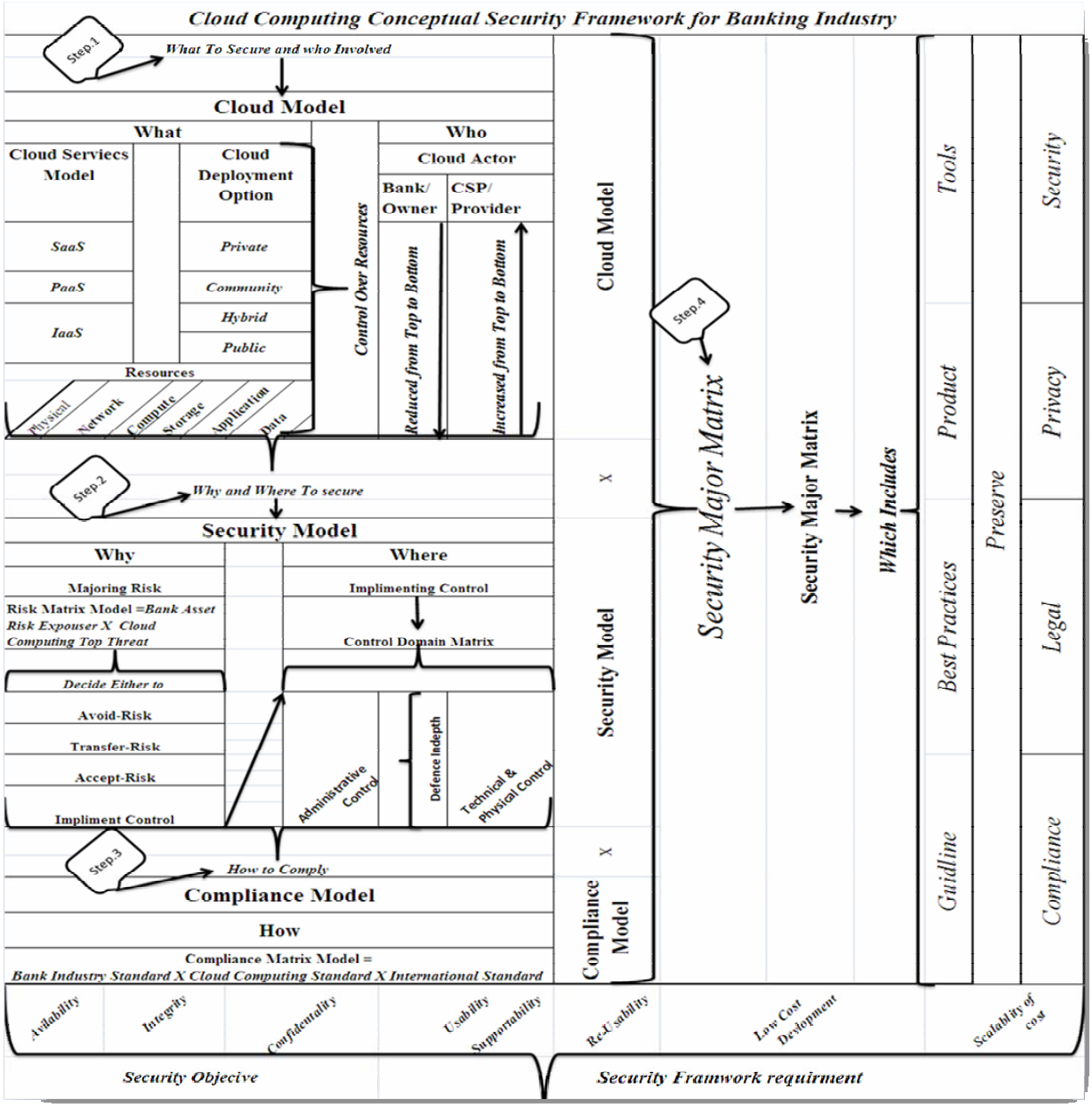

Figure 3: The Proposed Cloud computing Security Framework for Banking Industry

It is a tool for identifying banking business risk exposure through majoring risk impact and likelihood of risk exposure in deploying bank asset in each cloud deployment model.

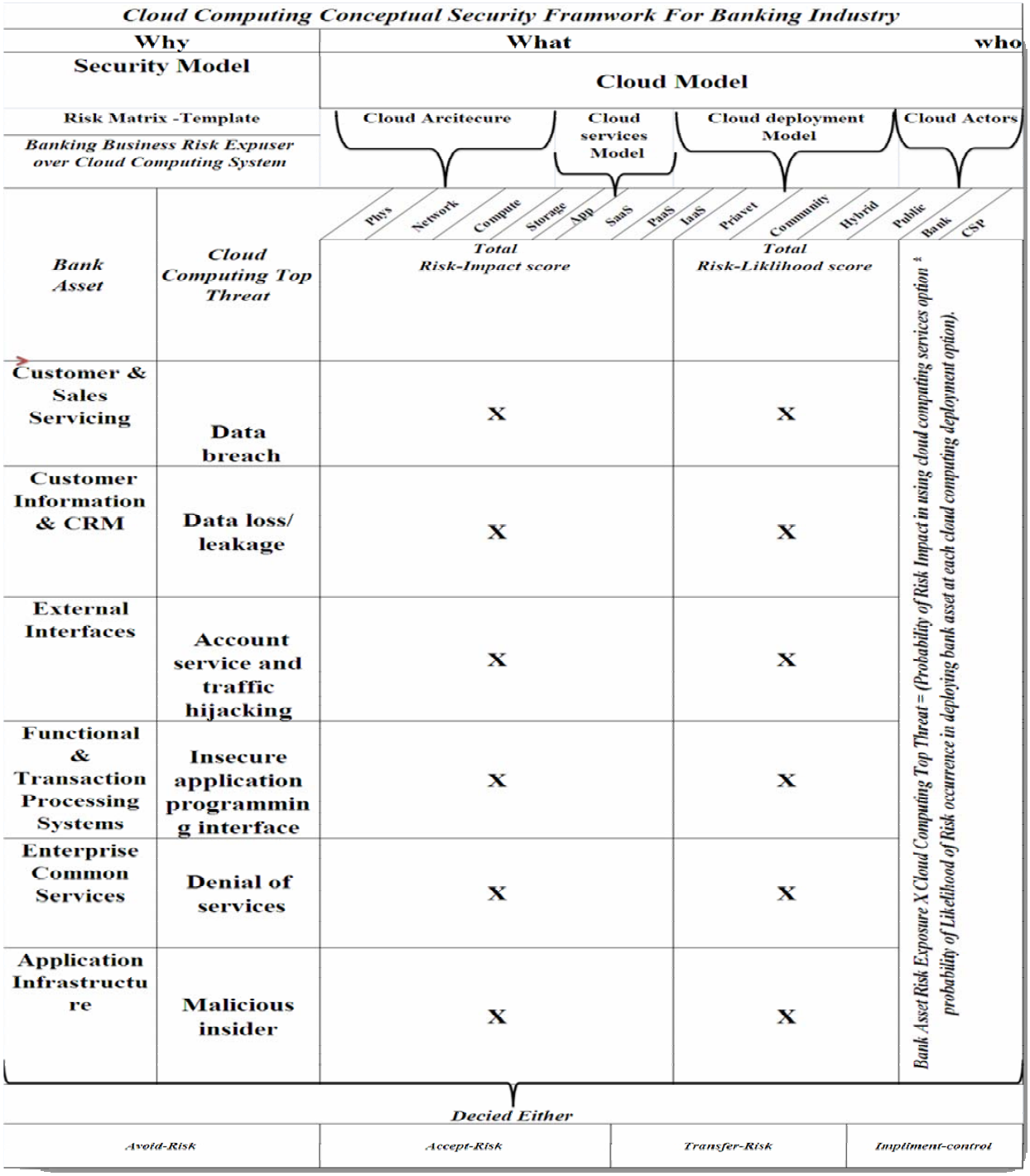

Figure 4: Risk Matrix template- Bank Business Risk Exposure

Compliance Model (How)

In the framework, in order to address legal and regulatory compliance issues, we designed Compliance Matrix Template (see Figure 5). It is developed basing banking industry standards (GLBA, BasleII/II, PCIDS, SOX), cloud computing control matrix and international regulatory best practices compliance requirement (ISO, COBIT, etc.). On the matrix, compliance requirement of those regulators is mapped to each specific control domain.

Cloud Computing Conceptual Security Framwork For Banking Industry					
Compliance Matrix-Template			How Compliance Model Compliance Matrix		
			Bank Industry Standards	Cloud Computing Standards	International Standards
			GLBA PCIDSS SOX BasleII/III	CCM-COBIT ,ISO/IEC 27001-2005, FedRAMP, PCI DSS	ISO HIPPA COBIT
Security Model	Administrativ-Control	Policy Document (Governance, Risk and Compliance)	x	x	x
		Legal	x	x	x
		Personal Security			
		Third party provider	x	x	x
		Business continuity and Resource provision			
	Technical - Control	Network, Host and VM security	x	x	x
		Appliaction security			
		Identity and access management	x	x	x
		Incident Management			
		Data secuity (Transit,storage,rest)	x	x	x
		Operational-managemer			
	Physical Control	Data center- Physical Security	x	x	x
		Data center- Enviromental Security			
		Data Center- Power And Network	x	x	x
		Data center- Human resourse			

Figure 5: Compliance Matrix Template

Based on the compliance matrix template, we designed cloud computing security strategy/major which includes specific tools, products, best practice and guidelines for the banking industry (see Table 1).

Table 1: Security Major Template

Cloud Computing Security Framework For Banking Industry	
Security Major Template	
Policy Document	Governance, Risk and Compliance
Legal	OECP, APEC
Personal Security	
Third Party Provider	
Business Continuity and Resource provision	Plan for managing resources, conduct impact analysis, check power & telecommunication infrastructure, back up-plan, disaster recovery-plan

Network, Host and VM security	*Wireless Network setting*: Perimeter Firewall, strong encryption for authentication & transmission. *Shared Network*: Bank security requirements, compliance with legislative, regulatory and contractual requirements, separation of production and non-production environments, preserve protection and isolation of sensitive data. *Network Control*: *Availability*: Check network architecture, e,g., Load balancing, Cluster architecture, Checkpoint restartor or robustness. *Integrity*: secured hashing algorithm e.g., SHA-512, Digital Certificates. Confidentiality: Strong password policy and access control e.g., 256-bit AES, SSL, SHA and TLS-1.1 *Virtual Machine Gust Hardening*: e.g., using firewall, web application, antivirus, file integrity monitoring & log. *Hypervisor Security*: Physical, operational security for hosting server.
Application security	Storage- Authentication, Digital signature/Hash, SAML, Audit logging, Web-services security get-way, and AAA.
Identity and access, management	RBAC, SSA token, OTP, SAML, XACML, SCLM
Incident Management	SLA, IODEF, RID, CEE
Data secuity (Transit,storage,rest)	DSLC, IDA, Encryption- Clint application, Network (SSL, VPN, SSH), Proxy, DAM, FAM, DLP-Tools, URL Filtering
Operational-Management	Documented operating procedures:-Capacity management, Change management, Exchange of information, System Acceptance.
Data center-Physical Security	Physical Security Perimeter, Resiliency - Equipment Location
Data center-Environmental Security	Smoke Detector & Fire Suppression System
Data Center- Power And Network	AC, Battery, UPS, Fire link detection, Automatic Fire extinguished, Generator
Data center- Human resource	Background Verification & Screening Agent

4. Conclusion and Future Work

The general objective of this research is to develop cloud computing security framework for the banking industry. Accordingly, in order to achieve this objective, detail assessment on cloud computing architecture, reference model, service, threat and attacks, policy, standards and guidelines were done. Similarly, assessment has also been made on bank industry regulatory security and compliance requirements. Based on these assessments to address the research problems we designed cloud computing security framework for the banking industry.

Our framework provides Risk Matrix Template for assessing and determining risk exposure of bank asset applications while moving to cloud deployment option. Integrated control domain component is proposed as a base for setting security major. Finally, for each defined control domain, security major/strategy (tools, products, guidelines and practices) is proposed.

For future work, we recommend automation of main security solutions, Risk Matrix template, Control Domain Template, Compliance matrix template for ease of use and updatability. For future research we also recommend setting maturity model for the proposed security framework.

Finally, banks are recommended to major their business risk exposure before moving their

application/asset/data to cloud deployment models (Public, Hybrid, Community, Private). We also recommend banks to set control major in compliance with regulatory requirements. On the other hand, to move banks to cloud computing more work is required form regulatory standard organization perspectives. This organization should provide and conduct more research and should provide updated security guidelines.

References

[1] David Bradshaw, Giuliana Folco, Gabriella Cattaneo, Marianne Kolding, "Quantitative Estimates of the Demand for Cloud Computing in Europe and The Likely Barriers to Up-take" IDC Analyze the Future, D4 Final report, Ver. 2.0, Brussels, Belgium, SMART 2011/0045, July 13, 2012.

[2] Ollivia La Barrer, "Bank System and Technology," Information week cloud, 2011, retrieved form, http://www.informationweek.com/cloud-computing/software/temenos-microsoft-bring-azure-clouds-to/231002099, Last access on August 15, 2013.

[3] H. Altalhi, Mohamed Sidek, Abdul Ghani, Othman, and Abdelkhaleq Al-Sheshtawil, "Enhancing the Security Framework Secure Cloud with the SWIFT Identity Management Framework", International Journal on Cloud Computing: Services and Architecture (IJCCSA), Vol. 2, No. 1, February 2012.

[4] IBM, "Security Over View Cloud Computing," Cloud Computing White Paper, United States of America, IBM Corporation, 2009.

[5] Federal Risk and Authorization Management Program, "FedRAMP," CONSOP Briefing, February 9, 2012.

[6] Sunita Rani and Ambrish Gangal, "Security Issues of Banking Adopting the Application of Cloud Computing," International Journal of Information Technology and Knowledge Management, Vol. 5, No. 2, July-December 2012.

[7] Working Group Report on Cloud Computing Option for Small Size Urban Cooperative Banks, Reserve Bank of India, retrieved from http://rbidocs.rbi.org.in/rdocs/PublicationReport/Pdfs/ RWGFUF031012.pdf, Last accessed on August 15, 2013.

[8] Working Group, "Information Security, Electronic Banking, Technology Risk Management and Cyber Fraud," Report and recommendation, Reserve bank of India, Mumbai, January 2011.

[9] John Sherwood, Andrew Clark, and David Lynas, "Enterprise Security Architecture" SABSA, White paper, 1995-2009 SABSA Limited.

[10] TOGOF, "The Open Group Architecture Framework," retrieved from, http://www.opengroup.org, Last accessed on August 14, 2013.

[11] SABAS, "The Sherwood Applied Business Security Architecture Framework" Available at http://www.sabsa-institute.org, Last accessed on March 15, 2013.

[12] ISO, "ISO/IEC 27002:2005", retrieved from, http://www.standardsconsultants.com/isoiec-270022005, Last accessed on August 14, 2013.

Adopting Cloud IaaS Architecture for Ethiopian e-Government Services

Teklu Itana
tekluitana@gmail.com

Abrehet Mohammed Omer
Addis Ababa Science and Technology University, Ethiopia
abrehet@gmail.com

Abstract

Government organizations delivering e-government services should look for better vehicles to decrease Information and Communication Technology budgets, and at the same time provide agile e-services to organizations and citizens. The paradigm of cloud computing services such as Infrastructure as a Service, Platform as a Service and Software as a Service offer means for governments to address issues of budget constraints and agility of service.

Infrastructure as a Service is one of the three main categories of cloud computing services. In a private cloud model, government organizations with similar business process can utilize resources of data center with relatively lower cost using a cloud Infrastructure as a Service. This paper investigates how to give efficient and cost effective e-Government services by adopting cloud Infrastructure as a Service architecture in a developing economy, Ethiopia, to gain a deeper insight into a research question: How can we provide cloud Infrastructure as a Service using few Data Centers in Ethiopia?

To answer the above question, exhaustive literature review on technologies that are the keystones to adopt and implement cloud Infrastructure as a Service architecture has been carried out. Assessments of some regional data centers, national data center and government offices that are connected to Ethiopian government private cloud, WoredaNet, have been done. The study specifically focused on those sectors that have automated their business process and have Management Information Systems application deployed in regional Data Center. Comparative analysis between the results of the survey and cloud Infrastructure as a Service architecture requirements was done.

We, therefore, adopted Cisco cloud Infrastructure as a Service architecture to enhance the current WoredaNet architecture for better and cost effective e-government services, particularly storage and computing power, from the regional Data Center.

We have analyzed the results of our testing that it can be scaled up to Wide Area Network like Ethiopian government private cloud, WoredaNet. We concluded that with limited number of servers we can solve storage and computing power problems of government organizations.

Keywords: Cloud Computing; Data Center; E-Government; Infrastructure as a Service; Virtualization

1. Introduction

The flow of information is essential for effective governance and managing the day-to-day business of government services for citizens. E-Government creates opportunities to reduce the costs of providing information and services to the public. It is the use of Information and Communication Technology that have the ability to transform relations with citizens, businesses, and different levels of government. Some of the e-government applications fall under the following categories [1].

- G2G (Government to Government): Administration, Inter-government enterprise, Control, monitor and distribution
- G2C (Government to Consumer): Registration/Land/Revenue Services, Hospital services, Agricultural services
- G2B (Government to Business): Tenders (e-tenders), Contract Management, Tax, etc.

E-government applications rely on Data Centers. Data Center is an essential corporate asset that connects all servers, applications and storage services. Therefore, Data Center is a key component that needs to be planned and managed carefully to meet the growing performance demands of users and applications [2].

Cloud computing is a promising technology model that is changing the way government organizations consume Information and Communications Technology (ICT), and how they deploy and deliver services to citizens and stakeholders. In addition, cloud architectures can benefit governments to reduce duplicate efforts and increase effective utilization of Data Center resources [3]. As it is described in [4], governments that adopt cloud computing strategy believe that "If the goal of e-Governance is to be achieved, it can do so by leveraging cloud infrastructure without having to purchase significant hardware, lowering both time and cost barriers in Data Centers". It is also indicated in [5] that E-Governance with cloud computing offers integration management with automated problem resolution, manages security end to end, and helps budget based on actual usage of data.

There are three main cloud services provided according to the demands of ICT customers. Firstly, Software as a Service (SaaS) provides access to complete applications as a service. Secondly, Platform as a Service (PaaS) provides a platform to develop other applications. Finally, Infrastructure as a Service (IaaS) provides an environment for deploying, running and managing virtual machines and storage services. The three cloud computing services can be viewed as a stack in Figure 1 [6].

Figure 1: Cloud Computing Service Layer

IaaS (Infrastructure as a Service) refers to computing infrastructure comprising of networking, hardware, virtualized operating systems and software servers offered as a service. At its core, Infrastructure as a Service is a way for organizations to get the hardware, storage, networking and other services they need to run their operations without worrying about buying, managing or maintaining the equipment [7].

1.1 Research Background

The Ethiopian Government has recognized the power of ICT in national development plan and had launched an e-government initiative in 2005. One of the e-government initiatives is establishing government cloud, called WoredaNet ICT Network, in which all public sectors in the country are to be connected to get e-Services from National Data Center.

Based on this initiative, there is 1 National Data Center and 9 Regional Data Centers in the country. The National Data Center is connected to the Regional Data Centers and to all Woredas that have WoredaNet ICT Network in the country. Currently, over 630 Woredas (districts) are connected; some government ministries and some regional bureaus that are connected to the WoredaNet are getting Internet and video conferencing services. The Ethiopian government cloud, WoredaNet, architecture is shown in Figure 2 [8].

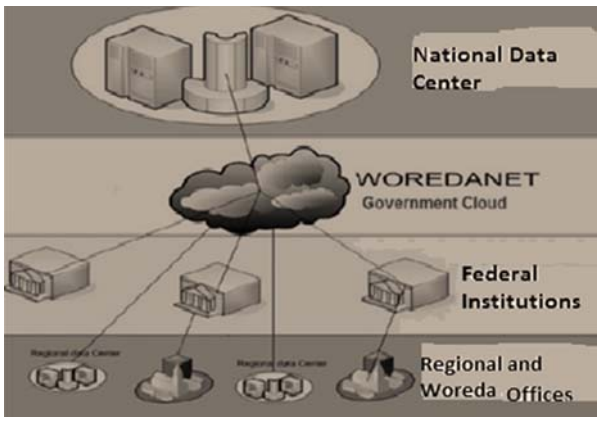

Figure 2: Ethiopian Government Cloud

1.2 Problems with Current e-Government Services

In Oromia region about 22 regional bureaus have conducted BPR (Business Process and Reengineering) study of their respective businesses and developed MIS applications since 2008, where each of them developed sector specific applications and databases. They require independent storage and computing power which is expensive to purchase for each of these regional bureaus. As the need of automating the manual business process increases within public sectors of the country, it requires huge amount of storage space and computing power. Therefore, the main purpose of this paper is to answer the research question: How can we provide Infrastructure as a Service using few Data Centers in Ethiopia?

To answer the above question, exhaustive literature review on cloud computing and some of its key benefits, particularly Infrastructure as a Service conceptual architecture, cloud IaaS architecture of different vendors and some of the parameters to adopt cloud Infrastructure as a Service architecture has been carried out. This study also identified technologies that are the keystones to adopt cloud Infrastructure as a Service architecture.

Assessments of some regional data centers, national data center and government bureaus that are connected WoredaNet have been done. As a result, primary data collection was done based on questionnaires developed to answer the research question and distributed to a total of 81 individuals (8 Federal IT experts, 7 regional IT heads and 66 regional IT Experts). About 64 respondents filled the questionnaires and based on their responses, comparative analysis between the results of the survey and cloud Infrastructure as Service architecture requirements was done.

The rest of the paper is organized as follows: Section 2 presents related work. The adopted cloud IaaS architecture is described in Section 3. Section 4 presents discussion. Section 5 concludes the work and points out future works.

2. Related Work

Shared hosting is the prominent hosting option for most businesses and government organizations in Ethiopia [9]. The principle of this hosting option comes from the fact that ethio-telecom is sharing the resources of a single server with many other users. However, in this hosting option there is a performance problem, and in addition resources of a data center such as storage and computing power are not provided as a service on demand.

Lowering costs is a major driver for adopting the cloud IaaS architecture. There are some core characteristics which describe what IaaS is. IaaS is generally accepted to comply with the following [7]:

- Resources are distributed as a service
- Allows for dynamic scaling
- Has a variable cost, utility pricing model
- Includes multiple users on a single piece of hardware

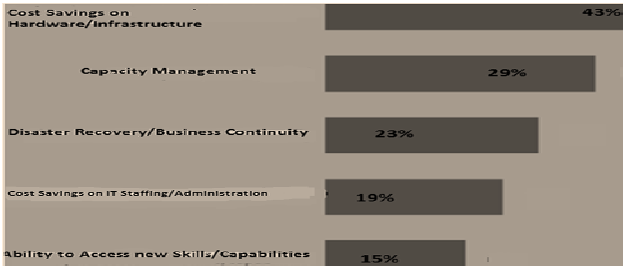

Figure 3: Top Five Motivations to Use IaaS

A survey conducted in [10] illustrates the top five motivations specified by respondents as reasons to use IaaS where cost savings is a key objective. The survey conducted is described in Figure 3.

IaaS provides simple provisioning of processing, storage, networks, and other fundamental computing resources over a network.

Table 1: Comparison of IaaS with other IT services options

Factor	*IaaS*	*Managed Services*	*Conventional Data Center*
Provisioning	On-demand, instant	On-demand, days to weeks	Weeks to months
Scaling	Instantly up or down	Weeks to months, up or down	Months, up only
Charges	No upfront, "pay as you consume"	Upfront fees, minimum term commitment, fixed monthly fees	Capital-equipment purchase or lease, maintenance charges, plus bandwidth utilization fees
Self-service?	Yes, via desktop console	No	Not automated

As it is compared in table 1 [10], IaaS service option is better provisioning option than other IT service options.

Cloud IaaS architecture designed by different vendors such as Cisco, IBM, Oracle and HP are discussed. Cisco is an enabler of the IaaS layer and provides specific services in the Software and Platform as a Service Layers. The Cisco cloud IaaS architecture [11] enables virtualized and on demand cloud services from a data center. IBM cloud IaaS architecture [12] also enables service based, scalable and elastic, shared and Internet based cloud services. Oracle Optimized solution for cloud IaaS [13] encompasses the compute node, storage infrastructure, network devices, virtualization software, operating systems, and management software to provide cloud services. HP Cloud System described [14] enables enterprises and service providers to build and manage services across private, public, and hybrid cloud environments with ease, reducing complexity and fostering business growth.

In this study, a thorough literature review of factors and opportunities to be considered for adopting Cloud IaaS architecture was conducted. Analyzing these factors and measuring their significance will help to choose which cloud IaaS architecture to adopt within government data center. The most powerful attribute of cloud IaaS architecture is the speed and simplicity with which pools of infrastructure resources can be fixed and flexibly configured and reconfigured to match infrastructure requirements. Therefore, factors to be considered to adopt cloud IaaS architecture are ranked to indicate their level of significance. Each factor is given low - for low significant factor, medium - for medium significant factor and high - for highly significant factor.

These factors are used for comparing and analyzing the different vendors' IaaS cloud architectures. A cloud IaaS architecture that considers factors that are highly significant were chosen and adopted in Oromia Regional government data center context.

Table 2: Factors and their level of significance

Factor	*Significance*
Bandwidth	High
Cost	High
Performance	High
Scalability	High
Security	Medium
management	Medium
Integration with other Services	Low
Standardization issue	Low
Ease of Use	Low

Based on factors and their level of significance in Table 2, for adopting cloud IaaS architecture in Oromia regional data Center context, Table 3 summarizes the comparison of the different vendors' cloud IaaS architectures.

Table 3: Comparison of vendors' cloud IaaS Architecture

Cloud IaaS Architecture	*Parameter for Adopting Cloud IaaS architecture in Oromia Regional Data Center*								
	Bandwidth Requirement	*Cost*	*Performance*	*Scalability*	*Security*	*Management*	*Integration*	*Standardization*	*Ease of Use*
IBM [12]	High	expensive	Greater latency	license	OK	Separate	OK	Vendor Lock in	complex
HP [15]	High	expensive	Greater latency	license	OK	Separate	OK	Vendor Lock in	complex
Oracle [13]	High	expensive	Greater latency	license	OK	Separate	OK	Vendor Lock in	complex
Cisco [11]	High	Less expensive	Lower latency	No license	OK	Unified	OK	Open	simple

Table 3 shows comparison of different cloud IaaS architectures based on factors to be considered before adopting the architecture. These factors are used for comparing and analyzing the different vendors' cloud IaaS architectures.

Different countries adopted cloud IaaS architecture. The Chinese University of Hong Kong transforms their IT infrastructure with Cisco Data Center .The University centralized and virtualized its information technology resources by deploying Cisco cloud IaaS architecture with its Unified Computing System [15]. Based on IBM cloud IaaS architecture, Vietnam Technology and Telecommunication (VNTT) offers cloud IaaS mainly the server or storage capacity and system capability for its clients, which are mainly for small and medium size enterprises. In collaboration with VMware Company, IBM has built a data center in South Africa, Johannesburg, to offer cloud services [16].

We had conducted literature review on the approaches that exist to adopt Cloud IaaS architecture. According to the comparison in Table 2, factors that are significant to adopt cloud IaaS architecture were clearly shown. Vendors' cloud IaaS architectures that considers these significant factors were compared. From the table, Cisco cloud IaaS architecture satisfies more than 66% of the factors that are significant to adopt cloud IaaS architecture. Therefore, It is preferable than other vendors' cloud IaaS architecture and adopted in Oromia regional data center.

3. The Proposed Solution

We have adopted Cisco cloud IaaS architecture based on the existing Ethiopian government private cloud, WoredaNet. In order to perform the adoption, the research first compares the different cloud IaaS architectures disused in the literature review with existing components in Oromia regional data center and identifies components of virtualization technology and defines their functionalities. Secondly, testing and demonstrations of Cisco cloud IaaS architecture in a small scale was carried out. Finally, by scaling up the adopted Cisco cloud IaaS architecture we propose the result to be implemented for Ethiopian e-government services.

3.1 Identifying Components of Cisco Cloud IaaS Architecture

Components in Cisco cloud IaaS architecture discussed in the literature review are identified and compared with Oromia Regional Data Center equipments to carry out the testing. The components are identified and compared as follows.

- *Application Software*: A client software ,XenCenter is to be installed on desktop/laptop computers.
- *Virtual Machine:* XenServer 6.1 host is to be installed on a physical server.
- *VMSwitch*: it is a built-in virtual switch which helps as a load balancer..
- *Storage and SAN*: it is a blade of servers or single server with SAN (Storage Area Network).

- *Compute, Access, Aggregation, Core, Peering*: are servers in rack, access switches, core switches, and routers respectively.
- *Backbone*: There is a WoredaNet Internet in National Data Center and Regional Data Centers.

Therefore, many of the components in Cisco cloud IaaS architecture are found in Oromia RDC.

3.2 Oromia Regional Data Center Services Scenario

Regional bureaus, zone and Woreda/district government offices are connected to regional data center and national data center via WoredaNet. These government offices are getting intranet services (e-mail, portal, MIS, Video Conference) from the Data Center through http://ip_address and then authenticated by their username and password.

3.3 Equipments that are found in Oromia Regional Data Center

1. Cisco Routers that are used to connect the data center to WoredaNet
2. Firewall to keep security
3. Cisco 6500 series, Core Switches are used to route packets among internal VLANs
4. Cisco 3560 series access switches
5. Video conference cameras and plasma displays
6. Server Farms are connected to SAN (Storage Area Network) through Veritas Net Backup Server
7. Veritas Net Backup Server is installed a software that manages all Severs and SAN arrays
   - It creates backup scheduling
   - Manage client computers
   - Manage SAN (Disk and Tape arrays)

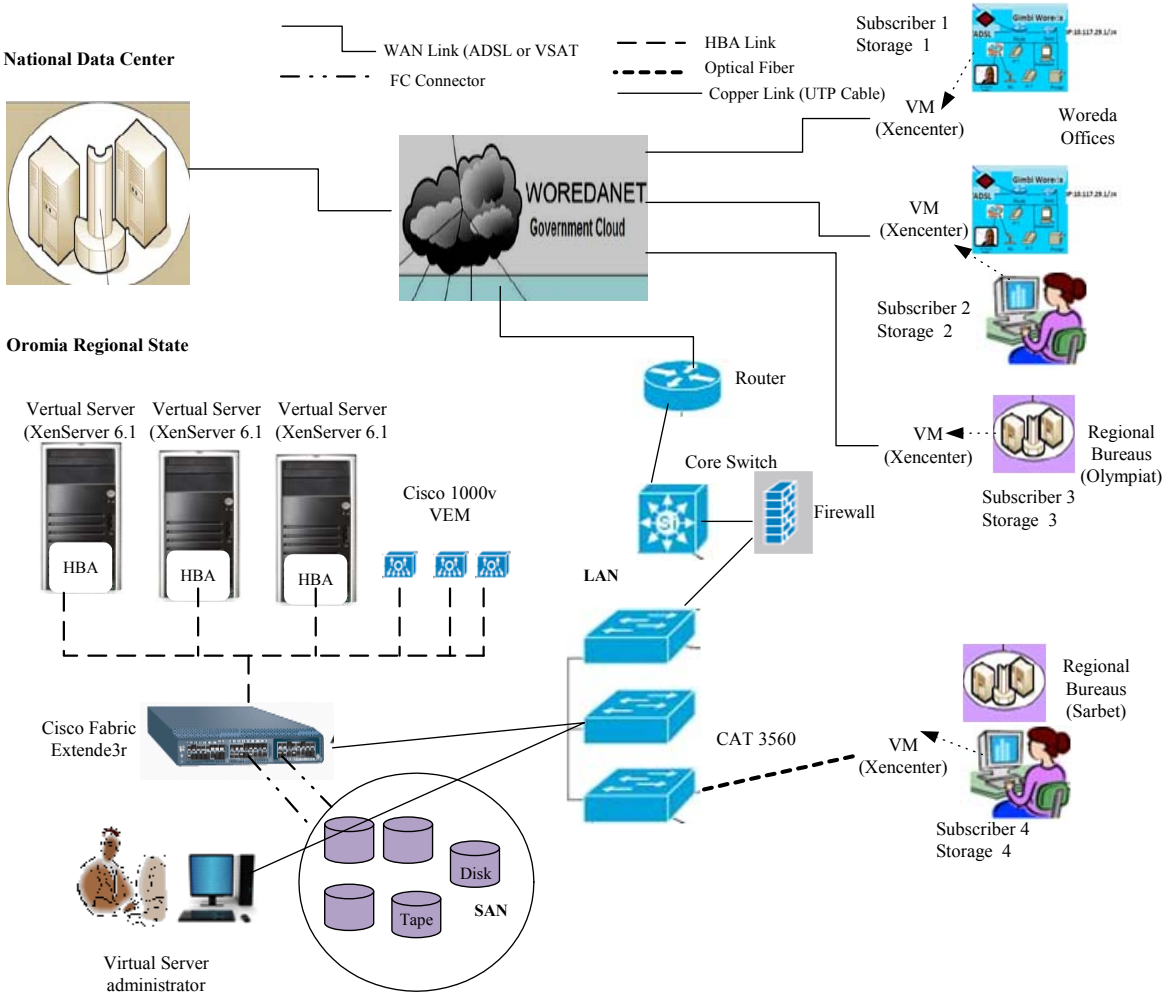

Figure 4: The Adopted Cisco Cloud IaaS Architecture

3.4 Description of the Adopted Architecture

Figure 4 shows an overview of the adopted Cisco cloud IaaS architecture and how the subscribers/users from regional bureaus, zones and Woreda sector offices could access server resources within Oromia Regional Data Center via WoredaNet. The number of servers is decreased from 7 to 3 in a new adopted architecture. The physical servers installed XenServer 6.1; it can run 500 virtual machines. This shows the 3 physical servers can run 1500 virtual machines which justify about 1500 users can connect and access the resources of the physical servers. Within six steps, the user can use a resource and then release it.

1. The user sends a request using the XenCenter.
2. User's verification using the built-in Active Directory feature in XenServer 6.1.
3. The user's request will be sent to the virtual infrastructure manager, XenServer 6.1.
4. The system will establish a connection between the requested service from the Cloud and the user.
5. The system synchronizes the service delivery between the user and the resource.
6. When the user is done, the system will terminate the session and disconnect the user from the cloud service.

Access to storage is done as follows [17]:

In the XenServer 6.1 Storage Repository, Virtual Disk Images are created as a storage abstraction of objects that can be presented to a Virtual Machine.

The Oromia ICT Development Agency director and two data center administrators validated the adopted Cisco Cloud IaaS architecture from its contribution to enhance the current data center services.

3.5 Contribution of the Adopted Cisco Cloud IaaS Architecture

The adopted Cisco Cloud IaaS architectures deliver a comprehensive solution for storage and computing power requirements of government organizations who have automated their business process. The key contributions of the adopted architecture include:

1. Cost Reduction: Few servers can serve for storage and computing power requirements.
2. Scalability: In case new server is added, more clients can be added to the data center.
3. Ease of Use: Using the management console, XenCenter, any network administrator with Microsoft Active Directory knowledge can manage resources of servers.
4. Performance: XenServer 6.1 has a feature of live migration of virtual machines between pools of servers.

4. Discussion

In this paper, the critical role of ICT in e-governance, limitations of the current ICT service delivery from regional data centers, and identification of better ICT service delivery technologies such as cloud IaaS architecture have been studied. After identifying the requirements and limitations, the research showed that Cloud IaaS architecture can overcome most of these limitations and fill the gap of storage requirements to improve e-government service delivery. Cloud IaaS architectures of different vendors and their components have been compared. From the different Cloud IaaS architectures, Cisco cloud IaaS architecture is chosen and adopted in Oromia regional data center. Cisco cloud IaaS architecture is chosen based on factors and parameters that are more significant to consider in adoption of cloud service architecture in the Ethiopian context. Open source virtualization tools, XenServer 6.1 and XenCenter have been used for the implementation and testing of the adopted Cloud IaaS architecture in small scale LAN environment.

5. Conclusion and Future Work

In this paper, we have adopted Cisco cloud IaaS architecture for Ethiopian e-government services. Through this work, we demonstrated the usefulness and effectiveness of services provided by Cisco IaaS cloud architecture. The testing of the adopted architecture in a small scale at one regional data center verified that the adopted Cisco cloud IaaS

architecture can scale up to large scale, at national level.

The study is limited to cloud IaaS architecture adoption in private cloud model. The study focused on government Data Centers that are already connected with WoredaNet. The adopted Cisco cloud IaaS architecture is a road map to adopt and implement other cloud computing services.

Future research should focus on adopting and implementing other cloud computing services beyond IaaS such as SaaS, PaaS and XaaS (Everything as a Service) in Ethiopian data centers. It should also focus on the impact of adopting cloud computing in developing economies such as Africa in general and Ethiopia in particular.

References

[1] Kuldeep Vats, Shravan Sharma, and Amit Rathee, "A Review of Cloud Computing and E-Governance", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 2, February 2012.

[2] Paul Stryer, "Global Knowledge Instructor, CCSI, CCNA, Understanding Data Centers and Cloud Computing", Expert Reference Series of White Papers, Global Knowledge Training Center, 2010.

[3] Fernando Macias and Greg Thomas, "Cloud Computing Advantages in the public Sector", white paper, 2011, retrieved from http://www.cisco.com/, Last accessed on October 22, 2012.

[4] Krishna Kant, "Data center evolution-state of the art, issues, and challenges", Journal of Computer Networks, Vol. 53, 2009, pp 2939-2965.

[5] Shahid Naseem," Cloud Computing and E-Governance", International Journal of Scientific and Engineering Research Vol. 3, Issue 8, August-2012.

[6] Ashish Rastogi, "A Model based Approach to Implement Cloud Computing in E-Governance", International Journal of Computer Application, Vol. 9, November 2010.

[7] CDW LLC., "Infrastructure as-a-Service Lowering IT costs is just one of many benefits Driving Organizations to IaaS", white paper, retrieved from http://www.CDW.com/, Last accessed on March 26, 2013.

[8] Mesfin Belachew, "Challenges and Opportunities in e-Government - Case of Ethiopia", in Proceedings of International Conference on e-Government, UNU-IIST, March 2010.

[9] "Internet and Data (VPN) Service", retrieved from: http://www.ethiotelecom.et/, Last accessed on January 20, 2013.

[10] TATA Communication, "Infrastructure as-a-Service Fulfilling the Promise of Cloud Computing", white paper, retrieved from http://www.tatacommunications.com/, Last accessed on March 26, 2013.

[11] Cisco Systems, "Cloud Computing -Data Center Strategy, Architecture and Solutions", white paper, 2009, retrieved from http://www.cisco.com/, Last accessed on November 20, 2012.

[12] IBM Global Technology Services, "Getting Cloud Computing Right", white paper, retrieved from http://www.ibm.com/cloud, Last accessed on March 23, 2013.

[13] Michael Wessler, "Enterprise Cloud Infrastructure", Oracle Special Edition, USA, 2012.

[14] Oracle Optimized Solutions, "Reduce Complexity and Accelerate Enterprise Cloud Infrastructure Deployments", white paper, retrieved from http://www.oracle.com/us/ solutions, Last accessed on March 25, 2013.

[15] Cisco Systems," How Today's Government, Education, and Healthcare Organizations Are Benefiting from Cloud Computing Environments", white paper, retrieved from http://www.cisco.com/en/US/solutions/cuhk_case_study.pdf, Last accessed on October 30, 2013.

[16] Nir Kshetri, "Cloud Computing in Developing Economies: Drivers, Effects and Policy Measures", PTC'10 Proceedings, retrieved from http://www.ptc.org/, Last accessed on March 25, 2013.

[17] Citrix XenServer 6.1, Installation Guide, retrieved from http://www.citrix.com/, Last accessed on June 10, 2013.

A Wireless Sensor Network Framework for Large-Scale Industrial Water Pollution Monitoring

Yohannes Derbew
yohanes.derbew@gmail.com

Mulugeta Libsie
Department of Computer Science, Addis Ababa University, Ethiopia
mulugeta.libsie@aau.edu.et

Abstract

Design and implementation of Wireless Sensor Networks (WSNs) is a challenging task because WSN research is usually application specific and each specific application requirement brings with it a different set of constraints and design objectives. Effective use of WSNs in large scale and long term monitoring of industrial water pollution requires satisfying two inherently contradicting requirements particular to this domain. On the one hand, it has to be time critical in order to provide early warning during compliance violations. On the other hand, it requires long-term data collection from a large number of distributed industrial sites for trend analysis to assist in decision making.

In this paper, we present the design of a comprehensive framework for using WSNs in industrial water pollution monitoring that aims to assist environment authorities in the decision making process of the battle against water pollution. The approach used to determine the suitability and effectiveness of our proposed framework is an experiment using a discrete-event simulator. The results show that our framework can provide in-network detection of compliance violations effectively without the need for powerful central processing stations thereby avoiding transmission of irrelevant data and consuming energy efficiently.

Keywords: Wireless Sensor Networks; Industrial water Pollution; Compliance Monitoring

1. Introduction

Water pollution is currently one of the major problems to the environment. With industrialization in major areas and urban cities growing, the water around them just keeps getting polluted. Recent research in Addis Ababa has classified Little and Great Akaki river basins as badly polluted to very badly polluted water [1].

There are a number of technologies developed to respond to the needs of effective monitoring to help prevent such pollution. Among these technologies, the most common option is the installation of industrial wastewater treatment plants that are capable of providing continuous and real-time monitoring services. However, these monitoring solutions require industrial users to install wastewater treatment plants that are very expensive and complex to operate.

This problem domain can benefit from WSNs. Compared with the present traditional monitoring techniques, enabling a monitoring system based on wireless sensor networks would present us with several advantages such as low cost, convenient monitoring arrangements, collection of a variety of parameters, high detection accuracy, high reliability of the monitoring results, etc.

Concerning that water is one of the pillars to sustainable economic development, a new, cost effective and efficient industrial water pollution monitoring solution using WSNs is proposed.

2. Background and Motivation

The Environmental Protection Authority of Ethiopia has standard procedures and policies to monitor and control the disposal of municipal and industrial wastes and by-products. Most of these policies and procedures are supported by manual practices that indirectly control the activities of plant

and industry users through document reviews and field visits. These practices are not efficient because they do not directly monitor the quantity and range of contamination released from polluting industries. As a result, industrial users are becoming highly ignorant on the control of the wastewater they dispose into the environment.

It is insufficient to rely on the present traditional monitoring techniques to satisfy the current monitoring requirements, which emphasizes the fact that water pollution monitoring must be continuous, dynamic, macro-scale, and swift. Hence, to protect the environment from threats of factory water pollution, a higher degree of effectiveness is required in the monitoring and control of industrial wastewater discharges than those obtained through traditional monitoring.

Most current environmental monitoring solutions based on WSNs are characterized by generating a large amount of data if continuously used over a long period of time, which limits their application to our specific scenario where we have thousands of industrial sites each being monitored based on applicable compliance standards regarding wastewater discharges. In order to alleviate this limitation, there should be a mechanism to minimize the amount of data generated from each monitored site while maintaining the required compliance information.

On the contrary, environment authorities need to get and keep long term compliance information about each monitored industry. This requires sending and storing complete information from each monitoring network.

3. Related Work

A number of research activities have been reported for water quality monitoring in the literature. In [2], the authors proposed a monitoring system to measure parameters such as pH, dissolved oxygen, conductivity, and temperature for aquiculture, river, and lakes monitoring, where the convenience of using these kinds of systems in terms of price, flexibility, and real time processing is explained. The work simply involves only a single deployment site with instrumentation components tailored to address a relatively small number of monitoring parameters.

Another effort related to water monitoring WSN is exposed in [3] with a study of a river watershed in North Carolina. The authors presented an end-to-end hardware and software structure to monitor water parameters through a large number of highly distributed sensor networks.

The work in [4] is a study that shows how a system for water quality monitoring can be set up using Zigbee WSNs. The study proposes the use of high power Zigbee based WSNs for low cost, ad hoc and easy installation and maintenance. However, the system is proposed for a vendor specific platform, Zigbee, and the authors did not mention if it is possible to use the proposed system for a wide range of WSN platforms.

In [5], an expert system called WPC-ES is presented for assisting departments of environmental management in their efforts to improve water quality in a city, applied in the Yellow River Basin of China. The system can analyze relationships between industrial water pollution and economic activities of industrial enterprises of a city.

While the majority of these WSN solutions are focused in natural environments, the framework shown in this paper is based on the requirements imposed by the government for industrial pollutions in wastewater discharges. It would not be possible to apply the protocols and algorithms proposed in these solutions for monitoring a large number, possibly thousands, of distributed industrial sites.

4. System Requirement

Design of a WSN framework for effective use in large scale and long term monitoring of industrial water pollution requires meeting a set of requirements specific to the application scenario. We have identified the following set of basic requirements:

Energy Efficiency: Since batteries are currently the only power source for wireless sensor nodes, power conservation is an important issue that can be optimized based on the intended use of the sensor network. In our specific application scenario, the trade-off between delay and energy overhead, optimizations of energy consumption per reported event, and network lifetime have been considered.

Quality of Service: QoS attributes in WSN highly depend on the nature of the intended application. The design of our proposed framework aims to satisfy two important QoS attributes that need to be considered for this usage scenario. The first consideration is the goal of *near real-time detection* which requires our monitoring solution to detect an event of compliance violations and report as early as possible. We refer to the second consideration as *pollution event detection/reporting probability* which determines the effectiveness of our monitoring solution in detecting and reporting all possible violations all the time.

Compliance Monitoring: a solution based on WSNs for monitoring industrial wastewater discharges should be able to monitor a given industry based on applicable standard to make sure that only wastewater that complies with the standard is discharged into the environment.

Minimum Generated Data: When using our proposed framework for large scale monitoring, there would be thousands of distributed sensor networks installed on many industries all over the country each producing its own pollution related data. Unless we introduce some efficient strategies to reduce the amount of data generated and transmitted from each distributed sensor network to the central database, storing, analyzing, and managing the sensor data becomes difficult.

5. Solution Overview

To satisfy the mentioned design requirements, our proposed solution involves the design of a WSN framework having four main components:

i. Sensor Deployment Scheme

Our sensor deployment scheme is proposed with the following goals in mind:

- The deployment should be easily performed around the existing infrastructure with little or no need for physical modification across any kind of industrial facility.
- In order to make the monitoring framework generic and countrywide, sensor deployment should not be site specific.
- In order to detect illegal discharges effectively, the sensor nodes should be placed on points that are representative of the spatial extent of the wastewater.

ii. Network Architecture and Topology Design

Our proposed architecture offers a hierarchical tree topology that organizes the sensor network into a multi-hop network that can adaptively control the monitoring nodes under various pollution conditions that can happen in the wastewater effluent.

The topology is constructed by the exchange of routing layer advertisement beacons among all nodes in the network. A child node attaches itself to a parent node only if both have similar sensing modality as shown in Figure 1.

Our proposed topology construction protocol optimizes the famous Collection Tree Protocol (CTP), the “de facto of data collection in WSNs” [6]. We tailored this protocol to the needs of our specific application because:

- CTP does not allow a two way communication of application layer data.
- CTP does not use sensor modality as a metrics for parent selection.
- Routing layer data in CTP is not visible to the application layer for intermediate nodes to perform in-network data aggregation.

Figure 1 shows a typical network topology where sensors with similar modality connect to the same branch of the topology tree.

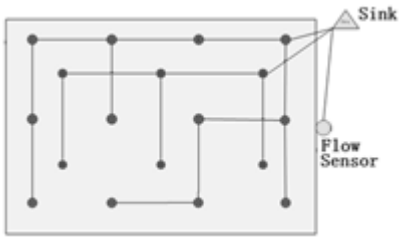

Figure 1: Sample network architecture and topology

iii. Context-aware Compliance Monitoring

Our proposed framework involves a rule-based node-centric context-aware model which reduces the amount of communication to the central server and improves energy efficiency. The following are the sources of context information that we consider:

- Conditions of pollutant parameters in the monitored wastewater: Pollutant concentration and Wastewater flow-rate
- Node role: (whether a node is a leaf or an intermediate node)

Our model is based on the Event-Control-Action (ECA) pattern presented in [7] for rule based context aware applications.

Figure 2 shows the high-level architecture of our proposed rule based node centric context-aware model.

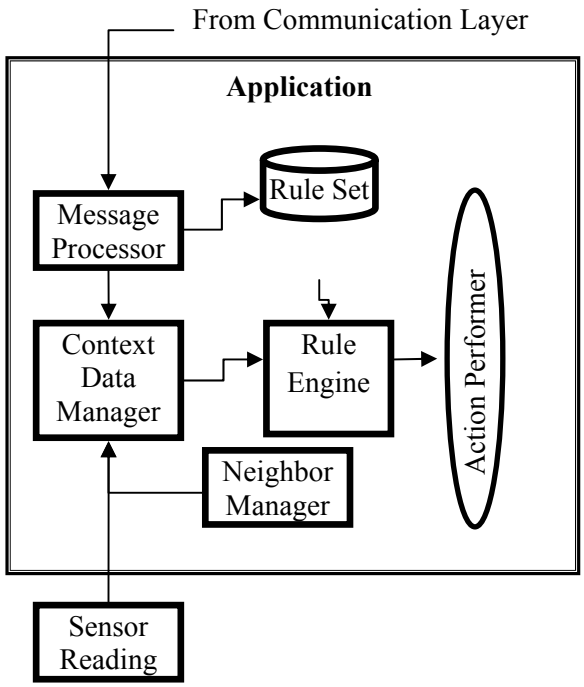

Figure 2: Architecture of the proposed node centric context aware compliance monitoring model

Message Processor identifies the type of message received from the communication layer and forwards the message to either the *Context Data Manager* which performs the task of capturing context information (about the wastewater and node role) from other nodes or locally from *Sensor Reading* as well as from the *Neighbor Manager* and provides temporary storage of this context information. The *Rule Engine* runs all the rules retrieved from the Rule Set to evaluate the context information coming from the local sensors as well as the context information gained from neighboring nodes. *Action Performer* is responsible for performing the actions triggered by the rule engine.

To cope with the limited local memory in the sensor nodes, the structure of the rules should be defined in order to express the compliance rules in a compact way as shown in Figure 3.

ID	Sensor Type	Condition	Min	Max	Action
(4)	(6)	(3)	(16)	(16)	(3)

Figure 3: Proposed rule format

The purpose of each field in the proposed rule format is discussed below:

ID (4 bits): a unique number that identifies each rule in the rule set.

Sensor Type (6 bits): is used to determine which sensor will be used in this rule. Each potential pollution parameter in the wastewater requires a specific type of monitoring sensor.

Condition (3 bits): the logical condition represents the current state of the wastewater such as current pollutant concentration, current role of a node, etc.

Min (16 bits), Max (16 bits): The values in these columns represent the limit values for each pollutant parameter as stated in the corresponding compliance standard.

Action (3 bits): Represent the operations of the rule that determine the reactions of the nodes in case a condition is true.

In order to reduce the computation at the sensor nodes, the contexts have to be described at the user

side by defining a set of conditional rules which need to be applied for those contexts. After that, these rules will be translated into a form that the sensor node can understand using the proposed format.

For example, the ph of wastewater needs to be kept in the range of [5, 8] before discharging. Any value of the monitored wastewater ph out of this range has to be reported. For this scenario, the context rule can be set as follows:

IF ph OUT_OF [6, 8] THEN SEND READING

Here, the *ph* is represented by a value in the *Sensor Type* field; *OUT_OF* is represented by the *Condition* field; [6, 8] is represented by the *Min* and *Max* fields; and *SEND READING* is represented by the *ACTION* field.

iv. Environment aware Intra-Network Communication Protocol

In our proposed framework, the communication between the monitoring nodes and the sink node is environment aware that is based on the conditions of the wastewater effluent and builds upon the node-centric and rule-based context aware model that we proposed. It consists of four different phases: Initialization Phase, No-Pollution Phase, Pollution-Alert Phase, and No-Flow Phase.

A state diagram showing how the monitoring nodes and the sink node in the network transit between these phases based on environmental conditions in the wastewater effluent is given in Figures 4 and 5.

As can be seen from the state transition diagrams, each node switches from one phase to another based on the environmental conditions in the monitored wastewater. Each phase has a specific objective to achieve. The INITIALIZATION phase handles the exchange of control messages and context rules during network setup. The NO-POLLUTION and NO-FLOW Phases are designed to reduce network traffic and minimize energy consumption when pollutant concentration of the wastewater is in the acceptable range or when the flow rate of wastewater discharge is zero. The objective of the POLLUTION-ALERT phase is to ensure that our framework satisfies the near real-time detection goal.

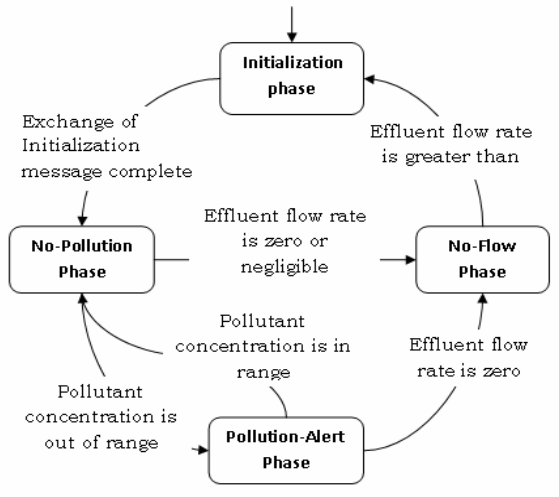

Figure 4: State Transition Diagram of a monitoring node

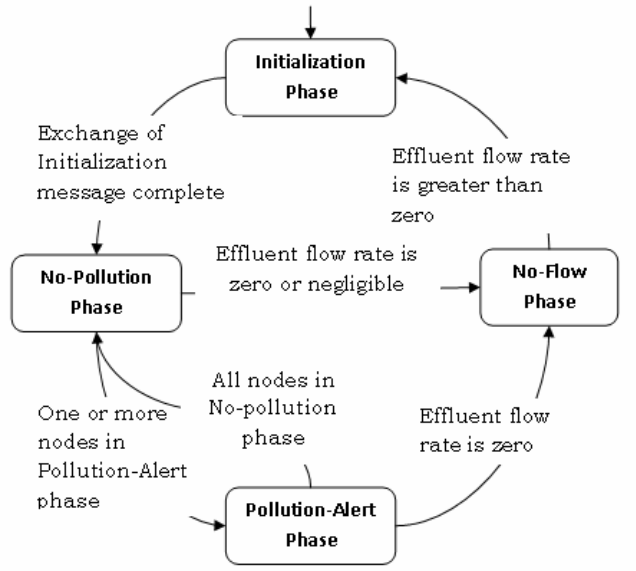

Figure 5: State Transition Diagram of the sink node

Figure 6 shows the format of the report message the monitoring nodes use to transmit non-complying pollutant concentration to the sink node during the POLLUTION-ALERT phase.

Node Count	Pollutant Type	Current Value

Figure 6: Format of the report message

The *Node Count* represents the number of nodes that have transmitted this report message so far. *Pollutant Type* is the index of the pollutant whose current value is being transmitted and *Current Value* represents the current average concentration for

intermediate nodes and the current abnormal reading for the origin node.

Each intermediate node computes the average concentration of the pollutant using the aggregation function given below and sends out a single value to the next hop.

$$C_{avg} = \frac{\sum_{n=1}^{R}(T_n C_n) + C_{slf}}{N}$$, where

- C_{avg} = average pollutant concentration determined by the current node
- N = sum of the *Node Count* values of all report messages since last sample plus the current node
- R = total number of report messages received since last sample
- T_n = Node *Count value* of report message n
- C_n = value of the *Current Value* of report message n
- C_{slf}= concentration sensed by the current node

In the same manner, the sink calculates overall pollutant concentration in the wastewater periodically based on the report it receives from the monitoring nodes and sends a report message to interested parties only if the current concentration is out of the permitted range. The sink node uses the following aggregation function for this purpose:

$$C_{ocl} = \frac{1}{S}\sum_{n=1}^{N}(T_n C_n)$$, where

- C_{ovl} = overall pollutant concentration in the reservoir
- S = total number of sensor nodes in the network
- N = sum of all *Node Count* values in all received report messages
- T_n = *Node Count* in report message n
- C_n = *Current Value* in report message n

The third quantity the sink should determine during the POLLUTION-ALERT phase is the hourly *load* of the pollutant contained in the nonconforming wastewater effluent that is being discharged from the monitored factory. A pollutant *load* is the mass or weight of pollutant transported in a specified unit of time from pollutant source to the environment.

The following function is used to estimate the hourly pollutant load of the discharged wastewater effluent:

$$Load = K\sum_{n=1}^{N} C(\tau_n)Q(\tau_n)$$, where

- *Load* = hourly load of the pollutant under consideration
- K = a constant for converting units
- $C(\tau_n)$ = concentration of the pollutant during sampling period n
- $Q(\tau_n)$ = effluent discharge flow rate during the sampling period n
- N = number of samples taken in one hour

6. Experimental Results and Discussion

In order to prove the framework's applicability and efficiency, a model of the proposed framework is implemented using Castalia Simulator [8] that builds on the OMNeT++ simulation framework. The developed WSN model consists of 18 nodes that are organized into two groups. The first group contains 12 nodes (nodes 1 - 12) that monitor temperature. The remaining six nodes (nodes 13 -18) are used to monitor conductivity. Node 0 is the sink node in our model just as Castalia takes node 0 as a sink node by default. The deployment pattern is uniform among similar sensors and the deployment area is 50 x 60 meter.

We used Castalia's *Customizable Physical Process* model and made necessary changes to represent the variations in our monitored environment, which is *Temperature* or *Conductivity* of the wastewater. We also developed a separate application and routing layer messages formatted specifically to support our proposed network topology and communication protocols.

We performed simulations using different configurations to evaluate the various components of

our framework and different design choices that we considered while deciding on a framework component. For each of the configurations, the simulation run is 100 minutes.

Scenario 1: Number of packets generated vs. node-centric context awareness compliance monitoring

In order to evaluate the effect of our context-aware model on the number of packets generated, we configured node 12 without context awareness (*Node12NoneContext*) while the rest of the nodes are configured with context aware operation (*ContextApplication*). The sink node is configured to collect data from each sensor node. The context rule used for this demonstration is a simple one. Each node transmits its sensor reading if the current reading is greater than a predefined report threshold. The number of packets received by the sink node from each sensor node, for each case is as shown in Figure 7.

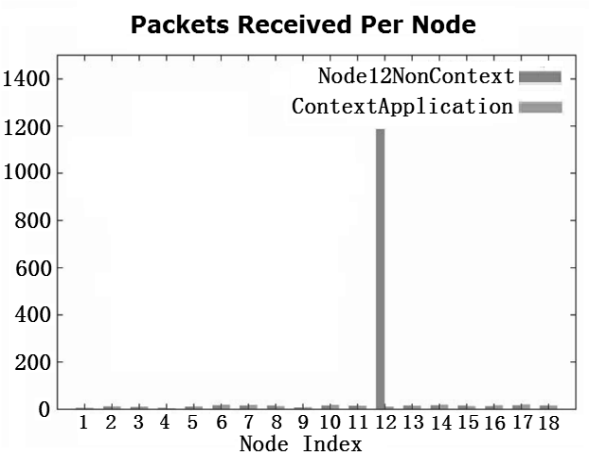

Figure 7: Number of packets received per node

Note that the sink received the largest number of packets from node 12 (which is run without our proposed context model) than from any other node during the simulation run.

In order to get better insight into the effect of using context aware sensor nodes for compliance monitoring, we performed a second simulation where all nodes, including node 12, are first run in context mode and then in non-context. The result of this simulation run is given in Figure 8. The graph shows the total number of packets generated by the whole network for the two cases.

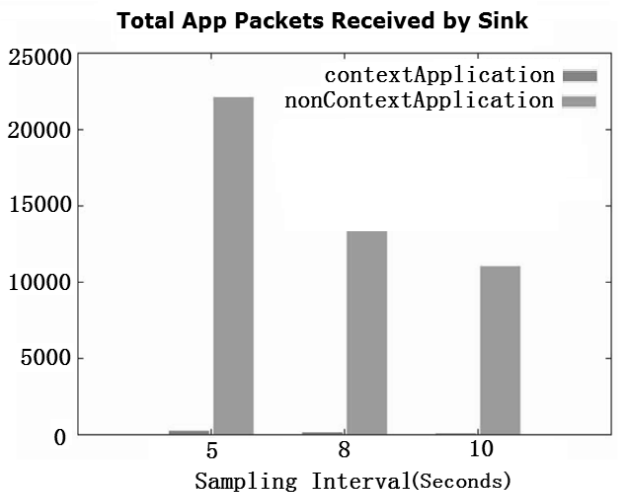

Figure 8: Effect of node centric context aware compliance monitoring on total network data traffic

Note from Figure 8 that it is also possible to reduce data traffic using larger sampling interval values. However, using longer sampling interval time increases the chance of missing events of hazardous pollutant discharge. This implies that our proposed compliance-monitoring model can reduce the amount of generated data even at higher sampling frequencies.

*Scenario 2: E*nergy consumption vs. node-centric context awareness compliance monitoring

As discussed before, by making sensor nodes aware of environmental context it is possible to reduce energy consumption of the sensor nodes and hence improve overall network lifetime. To show this, all nodes are first run in context mode and then in non-context mode using the same set of sampling rate values. The result of this simulation is presented in Figure 9, which shows the average energy consumption of the network in the two cases.

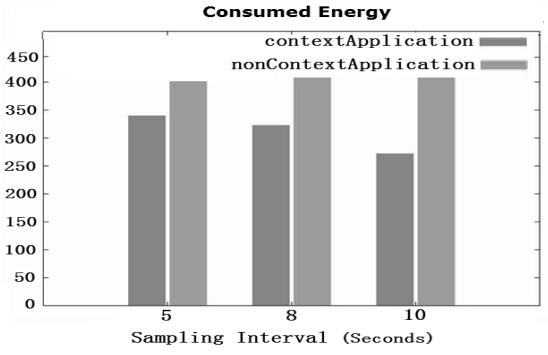

Figure 9: Effect of node-centric context aware compliance monitoring on energy consumption

Note from Figure 9 that the network consumes the maximum possible energy in the case of non-context operation even though we vary the sampling interval. However, in the case of context operation, the network can make considerable energy saving even for small sampling interval values.

Scenario 3: Number of packets collected vs. in-network data aggregation

Here, we evaluated our in-network data aggregation protocol based on its effect on the reduction of packets received by the sink node. To achieve this, we run all the sensor nodes alternately with the in-network aggregation protocol we developed and then without aggregation. The result of this simulation run is as shown in Figure 10.

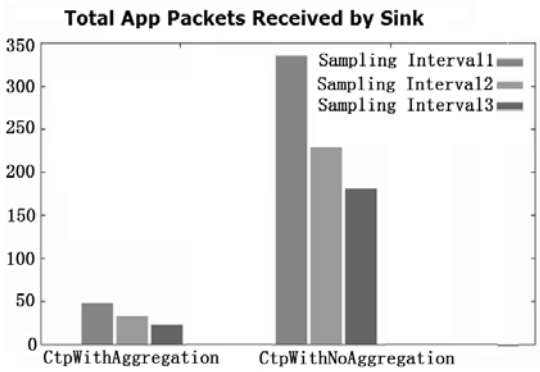

Figure 10: Effect of the proposed in-network aggregation protocol on network data traffic

The graph shows the number of packets received at the sink node during the simulation of 100 minutes both in the case of data aggregation and without it. Our proposed in-network aggregation protocol greatly reduces the amount of data the sink has to receive.

7. Conclusion

In this paper WSNs for use in large-scale industrial water pollution monitoring is presented. The outcome of the study is a novel WSN framework that incorporates four components each with a specific set of design objectives. The components of the proposed framework, along with their associated algorithms and protocols, are designed and optimized in order to satisfy the application specific requirements. The most important requirements that this work considered are the large number of distributed industrial facilities to be monitored, the possibility of future expansion and the need for industry specific compliance monitoring. Minimizing the amount of generated data traffic, energy efficiency, ability to monitor different industrial sites based on applicable compliance standards, and early detection are considered as design goals and treated in this work.

References

[1] G. Gebre and D. Van Rooijen, "Urban Water Pollution and Irrigated Vegetable Farming in Addis Ababa", 34th WEDC International Conference, Addis Ababa, Ethiopia, 2009.

[2] H.-B. Xia, P. Jiang, and K.-H. Wu, "Design of Water Environment Data Monitoring Node Based on Zigbee Technology", in Proceedings of CiSE '09, Wuhan, China, December 2009.

[3] G. W. Eidson, S. T. Esswein, and J. B. Gemmilletal, "The South Carolina Digital watershed: End-To-End Support for Real-Time Management of Water Resources", in Proceedings of the 4th International Symposium on Innovations and Real-time Applications of Distributed Sensor Networks (IRADSN '09), Vol. 2010, May 2009.

[4] Z. Rasin and M. Rizal Abdullah, "Water Quality Monitoring System Using Zigbee", International Journal of Engineering and Technology IJET Vol. 9, No. 10, 2010.

[5] Cheng H., Yang Z., and Chan C. W, "An Expert System For Decision Support of Municipal Water Pollution Control", Engineering Applications of Artificial Intelligence, Vol. 16, No. 2, 2003.

[6] O. Gnawali and R. Fonseca, "Collection Tree Protocol", ACM, Berkley CA, 2009.

[7] L. Maria Daniele, "Towards a Rule-based Approach for Context-Aware Applications", Unpublished Master's Thesis, University of Cagliari, Italy, 2006.

[8] Athanassios Boulis, "Castalia: a Simulator for Wireless Sensor Networks and Body Area Networks", User's Manual, Version 3.2, 2011.

HiLCoE

School of Computer Science and Technology

HiLCoE, one of the first higher education institutions in Ethiopia, was established in July 1997. It is a specialized Computer Science and Technology College that adheres to computing and quality of education. It is now widely recognized that HiLCoE prepares students to become industry and academic leaders who can apply technology and computer science principles across a wide variety of fields.

HiLCoE, since its inception has a vision of being Center of Excellence in ICT Education, Research and Development (ER&D). To realize this vision, in addition to its role in academia, it offers cutting-edge consultancy in the areas of Information Systems Security, Enterprise Architecture, Strategic and Solution Architecture, Service Engineering, Software Development, Project Management, Distributed Systems and Mobile Applications and many more.

Opportunities are tremendous with attractive reward for every effort you put to know and implement the evergreen field of Computer Science and Technology. Your dreams become realized by choice to join this pioneer and **Center of Excellence** in ICT ER&D in parallel with cutting-edge consultancy.

A commitment to excellence, coupled with continuous improvement, will result in HiLCoE being recognized as innovative, dynamic, and exciting community to learn, teach, and work with.

HiLCoE offers:

- **BSc in Computer Science** and

Two postgraduate programmes:

- **MSc in Computer Science** and
- **MSc in Software Engineering**

HiLCoE Computer Systems Engineering, from its industry link, offers **specialized tailored training and consultancy** in:

- Enterprise Architecture,
- Strategic IT Architecture,
- Information Systems Security,
- Project Management,
- Systems Development, and
- Mobile Application and Computer Network Design and Implementation.

HiLCoE paves a fabulous highway that makes you capable to be partner and contributor to the development arena.