

Final Report



Masters of Science in Computer Science  
University of Nairobi  
School of computing and informatics

# **Framework for Process Mining in Semi-structured Information Systems**

By

Mariam Mohammed Njoka

P58/9011/2006

08 Apr 2011

Supervisor: Andrew Mwaura

Submitted in Partial Fulfilment of the Requirements for the Degree of Masters of  
Science in Computer Science

University of NAIROBI Library



0478781 8

Bd 364683

Cho  
Afr  
TN  
275  
N56

## Declaration

This research project entitled "**Framework for Process Mining in Semi-structured Information Systems**" is my original work and has not submitted to any other university.

Sign:  ..... Date: 26/08/2011 .....

Mariam Mohammed Njoka  
P58/9011/2006

This research project entitled "**Framework for Process Mining in Semi-structured Information Systems**" has been submitted for examination with approval as the university supervisor.

Sign:  ..... Date: 26/8/2011 .....

  
Andrew Mwaura  
Supervisor

# Abstract

There are many application systems in the developing world that do not adhere to the systematic software development procedures and are not process-aware. Many of these systems do not meet their objective of efficiency & effectiveness. However, these applications can be analyzed and improved using information derived from event logs.

We discuss what process mining is, how to extract & prepare event logs from systems and the different tools that are used to process the data and give meaningful information, how to analyse the results and eventually introduce a framework that can be used to go through process mining.

We begin by following a framework developed by R.S. Mans, M.H. Schonenberg, M. Song, W.M.P. van der Aalst, and P.J.M. Bakker. However, the special circumstances of semi-structured systems are looked into, leading to changes in the framework. The proposed framework considers three main analysis methods to be performed on the event logs; *control flow analysis*, *performance analysis*, and *role analysis*. Data collection and preparation are given prominence due to the nature of discovery that is done on semi-structured systems. These systems in many cases will not have had previous statistical information of this crucial information. The process mining plugins help in getting this information. This research can be used to assist many organizations in cleaning up their systems and coming up with optimal processes which meet their objectives.

## Acknowledgements

I extend my heartfelt gratitude to my family especially my late mother Shakilla Juma, who continuously without fail impressed upon me on the importance of education, my son Ahmed Yussuf Wanjugi, who endured a busy mum with work and school and was very patient, he is the best son ever. I thank my sisters Mwini, Fatma and Umami and my brothers Ramadhan and Abdulhalim for their constant encouragement. Last but not least my father, Mohammed Njoka, who stands proud of his precious children.

In addition, I would like to give my thanks to my friends who encouraged me throughout my studies especially when I was feeling overwhelmed with the amount of work both in school & job.

I would also like to thank my classmates in the Msc Computer Science (May 2007), whose pressure kept me up and encouraged me.

I would wish to acknowledge the tireless guidance that I got from my supervisor, Andrew Mwaura and the contributions of the entire staff of the School of Computing and Informatics, University of Nairobi.

Last but most important, I give my utmost gratitude to Allah, who has brought me so far and enabled me to undertake this project.



# Table of Contents

Declaration .....	2
Abstract .....	3
Acknowledgements .....	4
Table of Contents .....	5
List of Tables.....	6
List of Figures.....	7
List of Abbreviations .....	8
Introduction .....	9
Background.....	9
Project Justification.....	9
Project Objectives .....	10
Project Methodology.....	10
Project Scope .....	10
Thesis statement .....	11
Literature Review .....	12
Process Mining .....	12
The ProM Framework Tool .....	16
MXML.....	18
Data Collection and Preparation .....	22
Data Collection .....	22
Event logs data cleanup .....	25
Conversion of the event logs to MXML .....	26
Mining and Analysis.....	31
The Proposed Framework.....	40
Implications of Research .....	43
Conclusion.....	44
Bibliography .....	45
Appendices .....	46
Appendix 1: How to configure an ODBC connection to a Microsoft Access Database on your computer.....	46
Appendix 2: Log summary of the event log .....	49

# List of Tables

**Table 1:** Example of a log file

**Table 2:** Identifying cases

**Table 3:** Dashboard Statistics

**Table 4:** Distribution of cases

# List of Figures

**Figure 1:** Process mining framework

**Figure 2:** Structure of an MXML file

**Figure 3:** Data before cleanup

**Figure 4:** Data after cleanup

**Figure 5:** Audit\_Trail\_Entries Table

**Figure 6:** Process mining table ERD

**Figure 7:** ProM Import Tool

**Figure 8:** Opening a Log file

**Figure 9:** Log Dashboard

**Figure 10:** Process Model 1

**Figure 11:** Process Model 2

**Figure 12:** Screenshot of the window to create a new data source

**Figure 13:** Compliant case following the most frequent path

**Figure 14:** Compliant case with the path covered by some of the cases

**Figure 15:** Non-compliant case with 25 events, depicting abnormality

**Figure 16:** Routing probabilities

**Figure 17:** LTL checker

**Figure 18:** Performance on different patterns

**Figure 19:** The proposed framework

**Figure 20:** Screenshot of the window to create a new data source

**Figure 21:** Screenshot of the window to set up the name of the database that has to be provided after jdbc:odbc: in the field DbHostUrl in the MS Access database plugin

**Figure 22:** Screenshot of the window to set up the database location

**Figure 23:** Screenshot of the window to set up the "username" and "password" provided to MS Access database plugin

**Figure 24:** Log Summary

## List of Abbreviations

**MXML** – M eXtensible Markup Language

**XML** – eXtensible Markup Language

**BI** – Business Intelligence

**BPR** – Business Process Reengineering

**WFM** – Workflow Management

**EAI** – Enterprise Application Integration

**ERP** – Enterprise Resource Planning

**WS** – Web Services

# List of Figures

**Figure 1:** Process mining framework

**Figure 2:** Structure of an MXML file

**Figure 3:** Data before cleanup

**Figure 4:** Data after cleanup

**Figure 5:** Audit\_Trail\_Entries Table

**Figure 6:** Process mining table ERD

**Figure 7:** ProM Import Tool

**Figure 8:** Opening a Log file

**Figure 9:** Log Dashboard

**Figure 10:** Process Model 1

**Figure 11:** Process Model 2

**Figure 12:** Screenshot of the window to create a new data source

**Figure 13:** Compliant case following the most frequent path

**Figure 14:** Compliant case with the path covered by some of the cases

**Figure 15:** Non-compliant case with 25 events, depicting abnormality

**Figure 16:** Routing probabilities

**Figure 17:** LTL checker

**Figure 18:** Performance on different patterns

**Figure 19:** The proposed framework

**Figure 20:** Screenshot of the window to create a new data source

**Figure 21:** Screenshot of the window to set up the name of the database that has to be provided after jdbc:odbc: in the field DbHostUrl in the MS Access database plugin

**Figure 22:** Screenshot of the window to set up the database location

**Figure 23:** Screenshot of the window to set up the "username" and "password" provided to MS Access database plugin

**Figure 24:** Log Summary

# List of Abbreviations

MXML – M eXtensible Markup Language

XML – eXtensible Markup Language

BI – Business Intelligence

BPR – Business Process Reengineering

WFM – Workflow Management

EAI – Enterprise Application Integration

ERP – Enterprise Resource Planning

WS – Web Services

# Introduction

## Background

There are very many application systems in the developing world that do not adhere to the systematic software development procedures and are not process aware. These systems aim to achieve efficiency within organizations. The fact is many do not meet this objective due to the fact that the developers do not come up with process models before development and they lack a measure of the processes from start to end to know if they are actually efficient enough. Human activities are also not taken into account when calculating how much time will be saved by using a particular system versus what existed before.

However, this does not mean that all is lost; these applications can be analyzed with a view of improving them. This will also go a long way in uncovering and measuring the discrepancies between prescriptive process models and actual process executions.

Process mining involves extracting information from event logs to capture the business process as it is being executed. Process mining aims at improving this by providing techniques and tools for discovering process, control, data, organizational, and social structures from event logs. The framework for process mining in semi-structured systems will be used more frequently in the third world given that many application systems do not take into account how efficient the processes are. Tools that can assist in analyzing these processes will go a long way in assisting organizations do their business process reengineering [BPR] and business intelligence [BI] more effectively.

Technologies such as workflow management (WFM), enterprise application integration (EAI), enterprise resource planning (ERP), and web services (WS) typically focus on the realization of IT support rather than monitoring the operational business processes. These systems have been created based on process models; however it is common that there are certain events or conditions that may not have been anticipated therefore there is no way they were captured in the model.

The tool that we chose for this research is the ProM Framework Tool.

## Project Justification

Many application systems in the developing countries, do not adhere to the systematic software development procedures and are not process aware. These systems aim to achieve efficiency within organizations. The fact is many do not meet this objective due to the fact that the developers do not measure the processes from start to end to know if they are actually efficient enough. Human activities are also not taken into account when calculating how much time will be saved by using a particular system versus what existed before.

However, these applications can be analyzed by following the framework which will assist the person analysing in finally coming up with recommendations which realise the benefits when implemented. This will also go a long way in uncovering and measuring the discrepancies between prescriptive process models and actual process executions.

## **Project Objectives**

In Kenya as in many third world countries there are so many organizations especially the government that are semi-computerized or in the process of getting computerized. Looking at the semi manual way of doing things there is a lot in terms of resources both time & labor that can be saved by merely computerizing, however, the question remains whether the computerized processes would save these resources optimally. Business process re-engineering has come in to enforce maximum efficiency in processes but there is a gap after implementation.

This gap is caused by the unforeseen events that happen as the process is actually executed. We will show how these organizations can analyze and evaluate their processes through actual study of the event logs after implementation and eventually streamline them. The main objectives of this research are;

1. To review literature on existing techniques for process mining
2. To put the topic of process mining into context, discuss the main issues around process mining, the different tools and methodologies of process mining and their application in semi structured systems in Kenya
3. Demonstrate using a process mining tool & event logs from a live semi-structured system, the data collection, data preparation, & finally analysis of processes from the system.
4. To propose a framework for process mining in semi-structured systems

## **Project Methodology**

The methodology used is as follows;

1. Literature review
2. Data collection & cleanup of event logs from a live application system
3. Analysis of the processes using a process mining tool
4. Creation of a process mining framework which shows the procedures followed and main entities to be taken into account.

## **Project Scope**

We cover the context of process mining, it discusses in detail how it is achieved using demonstrations from a system and a process mining tool called ProM framework, then we discuss the most challenging problems and propose a framework for process mining for a semi structured application system.



## Thesis statement

We present a framework for process mining of semi-structured systems from data collection, cleaning, conversion, mining & analysis of the event logs using the Prom Framework Tool among other tools. It is meant to guide others in going through the process as there are huge numbers of similar systems in use currently.

# Literature Review

## Process Mining

Process mining starts by gathering information about processes as they take place as W.M.P. van der Aalst and A.J.M.M. Weijters rightfully say in their research “Process mining – A Research Agenda”. Any transactional system would give this information in one form or another. Many information systems have a timestamp on events as they happen and usually these events are the first source of information for this process. The basic idea of process mining is to extract knowledge from event logs recorded by an information system. Example of a log is on Table 1. It contains five cases, each case one or more tasks depending on how far along the process the case is.

case 1	task A
case 2	task A
case 3	task A
case 3	task B
case 1	task B
case 1	task C
case 2	task C
case 4	task A
case 2	task B
case 2	task D
case 5	task E
case 4	task C
case 1	task D
case 3	task C
case 3	task D
case 4	task B
case 5	task F
case 4	task D

**Table 1: Example of a log file**

Until recently, the information in these event logs was rarely used to analyze the underlying processes. Process mining aims at improving this by providing techniques and tools for discovering process, control, data, organizational, and social structures from event logs. Fuelled by the omnipresence of event logs in transactional information systems (cf. WFM, ERP, CRM, SCM, and B2B systems), process mining has become a vivid research area. The process mining group at Eindhoven University of Technology in Netherlands have been specifically involved in process mining research and have many publications and tools that they have come up with.

Many vendors are now pushing technologies such as Business Process Analysis (BPA) and Business Activity Monitoring (BAM). These systems typically aim at basic performance indicators such as cycle time and frequencies. The goal of researchers on process mining is to allow for more advanced concepts where knowledge is extracted from logs and causalities can be discovered.

There are many workflow tools that have come up over time however there is one that is interesting enough because of the kind of information and analysis it provides. The ProM Framework is an extensible framework that supports a wide variety of process mining techniques and many researchers e.g. W.M.P. van der Aalst and H.T. de Beer and B.F. van Dongen in their research "Process Mining and Verification of Properties: An Approach based on Temporal Logic" have used it especially the mining plug-in, which implements a mining algorithm that constructs a Petri net based on an event log, the export plug-in, which implement a "save as" functionality for objects e.g. graphs etc.

Process mining should not be confused with Business Intelligence (BI). Business intelligence is a broad category of applications and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions. BI applications include the activities of decision support systems, query and reporting, online analytical processing (OLAP), statistical analysis, forecasting, and data mining. While process mining digs deeper at the event logs that are produced by enterprise systems in order to come up with meaningful information on process performance, audit etc.

BI is a term that was used as early as September, 1996, when a Gartner Group report said: By 2000, Information Democracy will emerge in forward-thinking enterprises, with Business Intelligence information and applications available broadly to employees, consultants, customers, suppliers, and the public. The key to thriving in a competitive marketplace is staying ahead of the competition. Making sound business decisions based on accurate and current information takes more than intuition. Data analysis, reporting, and query tools can help business users wade through a sea of data to synthesize valuable information from it - today these tools collectively fall into a category called "Business Intelligence".

Today's Business Intelligence (BI) tools used in many industries, like Cognos, Business Objects, or SAP BI, typically look at aggregate data seen from an external perspective (frequencies, averages, utilization, service levels, etc.), as R.S. Mans, M.H. Schonenberg, M. Song, W.M.P. van der Aalst, and P.J.M. Bakker indicate in their research "Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital". These BI tools focus on performance indicators e.g. the number of knee operations, the length of waiting lists, and the success rate of surgery. Process mining looks "inside the process" at different abstraction levels. So, in the context of a hospital, unlike BI tools, the concern is mainly with the care paths followed by individual patients and whether certain procedures are followed or not. Process mining would also focus on process performance e.g. time taken in each

activity, what process paths were frequently followed, what can be improved from this to make the process easier on patients.

This research topic has recently gained a lot of interest as information systems come of age and computerization is not seen as the only answer, more and more benefits are being sought out of these systems.

An interesting class of information systems that produce event logs are the so-called *Process-Aware Information Systems* (PAISs). Examples are ERP systems (e.g. SAP), case handling systems (e.g. FLOWer), CRM systems (e.g. Microsoft Dynamics CRM), middleware (e.g., IBM's WebSphere), hospital information systems (e.g., Chipsoft), etc. These systems provide very detailed information about the activities that have been executed.

Process mining addresses the problem that most "process/system owners" have limited information about what is actually happening. In practice, there is often a significant gap between what is prescribed or supposed to happen, and what *actually* happens. Only a concise assessment of reality, which process mining strives to deliver, can help in verifying process models, and ultimately be used in system or process redesign efforts.

*The idea of process mining is to discover, monitor and improve real processes (i.e., not assumed processes) by extracting knowledge from event logs.* We consider three basic types of process mining; *discovery*, *conformance*, and *extension*.

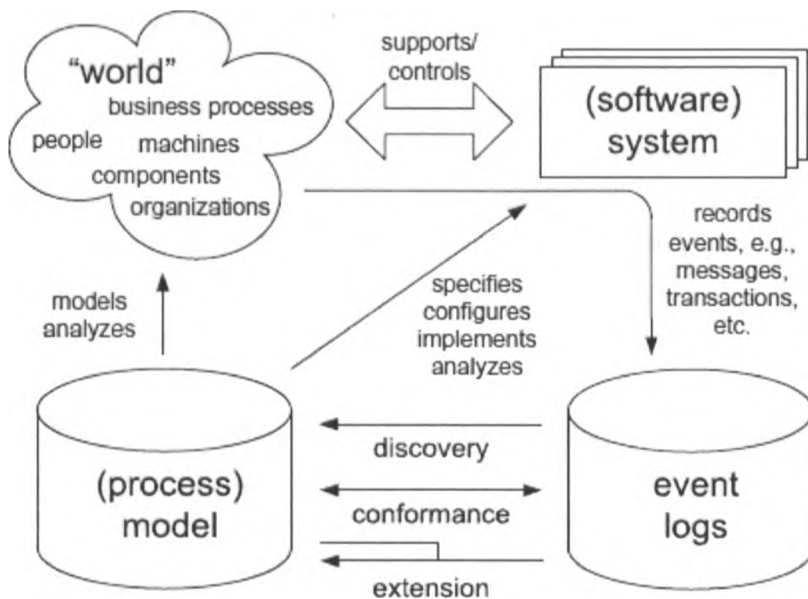
#### **Framework reviewed**

The framework to be developed in this study will be based on the framework developed by R.S. Mans, M.H. Schonenberg, M. Song, W.M.P. van der Aalst, and P.J.M. Bakker in their research "Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital". They considered three basic types of process mining: *discovery*, *conformance*, and *extension*.

**Discovery:** Traditionally, process mining has been focusing on *discovery*, i.e., deriving information about the original process model, the organizational context, and execution properties from enactment logs. An example of a technique addressing the control flow perspective constructs a Petri net model describing the behaviour observed in the event log. It is important to mention that there is no a-priori model, i.e., based on an event log some model is constructed. However, process mining is not limited to process models (i.e., control flow) and recent process mining techniques are more and more focusing on other perspectives, e.g., the organizational perspective, performance perspective or the data perspective. For example, there are approaches to extract social networks from event logs and analyze them using social network analysis. This allows organizations to monitor how people, groups, or software/system components are working together. Also, there are approaches to visualize performance related information, e.g. there are approaches, which graphically shows the bottlenecks and all kinds of performance indicators, e.g., average/variance of the total flow time or the time spent between two activities.

**Conformance:** This is an a-priori model. This model is used to check if reality conforms to the model. Conformance checking may be used to detect deviations, to locate and explain these deviations, and to measure the severity of these deviations.

**Extension:** This is an a-priori model. This model is extended with a new aspect or perspective, i.e., the goal is not to check conformance but to enrich the model with the data in the event log. An example is the extension of a process model with performance data, i.e., some a-priori process model is used on which bottlenecks are projected. At this point in time with mature tools such as the ProM framework, featuring an extensive set of analysis techniques which can be applied to real-life logs while supporting the whole spectrum depicted in the figure below..



**Figure 1: Process mining framework**

### Critique of the framework

R.S. Mans, M.H. Schonenberg, M. Song, W.M.P. van der Aalst, and P.J.M. Bakker' model, captures the process and the relationships from the business requirements and reality to the system software which does the actual logging as processes are done through it, to the event logs which are then mined and finally to the process model, which again feeds into the world realities and business requirements.

However, in many semi-structured systems there is most often than not a lack of process model, so it is expected that once the process mining process is through then a process model will be produced and fed into the system.

Another handicap with semi-structured systems is the data cleanliness, this is a huge aspect as very often there will be data which does not conform perhaps it was testing data or otherwise. Many systems are debugged and correctly on live environments. Thus cleanup and preparation of data should be considered as an important part of the process.

Finally, the best way to analyse is to look at the logs with a view of getting the control flow analysis, performance analysis and role analysis. With this kind of information, then all manner of decisions touching on how to control the work flow, how to improve the organizational/role allocation and how to improve the performance of the processes themselves can be done.

One of the tools to support process mining is the process mining framework ProM. It is plugin-based to support new areas and techniques. In the last decade, process mining evolved from control flow discovery to a broad area of research to get all kinds of information from a log, which resulted in more than 250 different plugins within ProM. This shows that many different techniques exist to apply process mining. Since there are so many, it is not clear anymore when to use which plugin. Although many case studies, have been performed, the main problem with these case studies is that they were all done case-by-case on the insights and knowledge of the researcher performing the case study. I.e. it is hard to make process mining a repeatable service as said in a research titled "Process Diagnostics: a Method Based on Process Mining" by Melike Bozkaya, Joost Gabriels & Jan Martijn van der Werf LaQuSo.

## The ProM Framework Tool

**ProM** is an **extensible** framework that supports a wide variety of process mining techniques in the form of plug-ins. It is **platform independent** as it is implemented in Java. **ProM** framework is issued under an **open source** license, namely the Common Public License (CPL), and researchers and developers are invited to **contribute** in the form of new plug-ins. It was developed by the Process Mining Group, Eindhoven Technical University.

ProM is a generic framework for implementing process mining tools in a standard environment. The ProM framework receives as input logs in the Mining XML (MXML) format. Currently, this framework has plug-ins for process mining, analysis, monitoring and conversion.

ProM is available as binary distribution files for the Windows, Mac OS X and UNIX platforms, and as source code under the terms of the CPL license. It requires a present installation of the Java Runtime Environment, version 1.5/5.0 or greater (Version 5.0 is recommended for Windows, Linux, and Mac OS X).

The following are the different types of plugins that are found in ProM;

**Mining plug-ins**, such as:

1. Plugins supporting control-flow mining techniques (such as the Alpha algorithm, Genetic mining, Multi-phase mining, etc.)

2. Plugins analysing the organizational perspective (such as the Social Network miner, the Staff Assignment miner, etc.)
3. Plugins dealing with the data perspective (such as the Decision miner, etc.)
4. Plugins for mining less-structured, flexible processes (such as the Fuzzy Miner)
5. Elaborate data visualization plugins (such as the Cloud Chamber Miner)
6. (and many more)

**Analysis plug-ins** dealing with:

1. The verification of process models (e.g., Woflan analysis)
2. Verification of Linear Temporal Logic (LTL) formulas on a log
3. Checking the conformance between a given process model and a log
4. Performance analysis (Basic statistical analysis, and Performance Analysis with a given process model)

**Export plug-ins**, which implement “save as” functionality for some objects (such as graphs). For example, there are plug-ins to save EPCs, Petrinets, spreadsheets, etc.

**Import plug-ins**, which implement an “open” functionality for exported objects, e.g., load instance-EPCs from ARIS PPM.

**Conversion plug-ins**, which implement conversions between different data formats, e.g. EPCs to Petri nets.

Finally, ProM sports a large array of **log filters**, which are a valuable tool for cleaning logs from undesired, or unimportant, artefacts.

There are other tools in the market that are used in the process mining field, examples are below though this are offered commercially;

1. Futura Reflect, a Process Mining and Process Intelligence suite developed by Futura Technology
2. Interstage Automated Process Discovery, a Process Mining service offered by Fujitsu, Ltd. as part of the Interstage Integration Middleware Suite.
3. BPM|one, offering both basic process mining functionality as well as a more comprehensive process mining module as part of the Pallas Athena BPM|one software suite.
4. Nitro is a tool by Fluxicon for easily converting CSV and XLS event logs for ProM.
5. ARIS Process Performance Manage, a Process Mining and Process Intelligence Tool offered by Software AG as part of the Process Intelligence Solution.

The following reasons led me to choose ProM;

1. It is an open source software
2. It has been used extensively for research, with several researchers adding to its functionality.
3. It is rich in functionalities and different mining & analysis plugins.
4. It is widely seen as a leader in process mining tools.

## MXML

MXML is an XML-based user interface markup language first introduced by Macromedia in March 2004. Application developers use MXML in combination with ActionScript to develop Rich Internet applications. This is a vendor-independent format to store event logs. One MXML file can store information about multiple processes. Per process, events related to particular process instances (cases) are stored. Each event refers to an activity.

### *Data model*

MXML has three kinds of objects: the documents, the nodes and the iterators. A document has mainly a root node, which holds all the top-level document nodes. Even if only one root -tag- node is allowed in a valid XML file, it is possible to have more non-tag nodes, like comments, processing instructions, and directives. In a document object, MXML also stores eventual errors and the line where an error has been encountered. The document is also used to carry around "formatting style" requests in input or output.

A node is the minimal unit of information; MXML distinguishes among 6 kinds of nodes:

1. Document nodes: they are just "transparent" holder for top-level nodes. The root node in a document is always a document type node.
2. Tag nodes: they are the nodes used to store information in the XML file: an example could be `<item> data </item>`
3. Data nodes: They are the data held by an item. An item can have more than a data node child, as it can, as in this example:



```

<bibliography>
  <book isbn="12345678" publisher = "pb1">
    <author>
      [language = English]
        <firstname> Manolis <firstname>
        <lastname> Gergatsoulis <lastname> [ ]
      [language = Greek]
        <firstname> Μ Ο & <firstname>
        <lastname> " Ο & <lastname> [ ]
    <author>
    <author>
      [language = English]
        <firstname> Panos <firstname>
        <lastname> Rondogiannis <lastname> [ ]
      [language = Greek]
        <firstname> Ο& <firstname>
        <lastname> Ρ Ο Ο & <lastname> [ ]
    <author>
    <title>
      [language = English]
        Multidimensional Programming Languages [ ]
      [language = Greek]
        Ο " & ! " & Ο Ο [ ]
    <title>
    <price currency="USD">
      [period = discount client = regular]
        100 [ ]
      [period = normal client = regular]
        120 [ ]
      [period in {discount, normal} client = special]
        100 [ ]
    <price>
    <year> 1999 <year>
  <book>
  ...other book elements ...

  <publisher id = "pb1">
    [language = English] NCSR Demokritos [ ]
    [language = Greek] ΕΚΕ Ε Ο ο& [ ]
  <publisher>
  ...other publisher elements ...

</bibliography>

```

4. Comment nodes: they are comments put by the document designers to state something meaningful about what the document (or document part) are meant for. They assume the form of a string inside a block that begins with `<!--` and ends with `-->`: `<!-- This is an XML comment -->`

5. Processing instructions: they are blocks that are passed as they are to 3d party processors; php escape (`<?php ... code ... ?>`) is an example. Also the `<?xml` declaration is treated by XML as a processing instruction; this could change in next releases of MXML.
6. Directives: they are nodes inside a `<! ... >` block. The DOCTYPE and all the DTD declarations, along with the internal and external entity declarations are directives. MXML does not parse them in any way, and handles them untranslated to the application.

Iterators are meant to access easily the document nodes structure, allowing partitioning the tree in subtrees.

### *Nodes in-depth*

Each node has a name, a set of attributes (each of which is a couple of string, the name and the value), a data and the links to the siblings nodes. Some node types can have some of this values not set. In example, both processing instructions and directives has a name (the string immediately following the opening tag) and a data (the content of the tag), but they have not an attribute set. Data and comments notes have only data, while tag nodes have pretty everything. Document nodes are empty, "transparent". To simplify the work of programmers having to scan configuration oriented XML documents, if a tag node has only one data node among its children, its data element will be merged with the data element of that child, and the data node will be removed thus "flattening" the structure. Is it also possible to create a new tag node with its data element set to a string. If, later on, there is the necessity to add a new data node to the item, it does not matter if it has the data element merged: on output, both the data element and the entire child data node will be correctly written. The data element is always written after all the children nodes.

Navigation in the node tree is guaranteed by four node attributes, pointing respectively to the next node in the same level of the hierarchy (and with the same parent), to the previous one, to the first node of the node children, and finally to the parent. So, to traverse the whole tree, one has to start from the root node and then descend recursively in the first child node: then all the "brothers" of that node are scanned up to the last node in the tree.

### *Utilities for node management*

The API is discussed in detail in the documentation; what is necessary to attention, is that some utilities are provided to access/modify the content of the attribute list of the nodes, and to retrieve the depth and, eventually, the path of a node. The depth of a node is its distance from the root node, counted in steps necessary to reach it or his first brother. The path of a node is the list of the name of all its ancestors, plus the node name, separated by a "/" character: `/main/item` is a path indicating that the node named `item` can be found immediately below the node named `"main"`, which is at top-level. Node paths are not unique, moreover, only a node having a name, and whose parents are all named, can have a path.

### Naming conventions

MXML naming conventions are rigid ways in which each function, structure and data is called. Each function in the library begins with "mxml\_", the name of the object that the function refers to and the operation on that object. In example, the function to create a new node is "mxml\_node\_new()". Mxml function namings are not necessarily restricted to structures, but could also refer to abstract objects, like "mxml\_path\_\*" (which operates on strings representing node paths), or "mxml\_attribute\_\*" (which handle the attribute lists for nodes). Every symbol defined in .h file begins with "MXML\_". So, a document is typed as an MXML\_DOCUMENT, and a node is a structure named MXML\_NODE.

### Structure of an MXML file

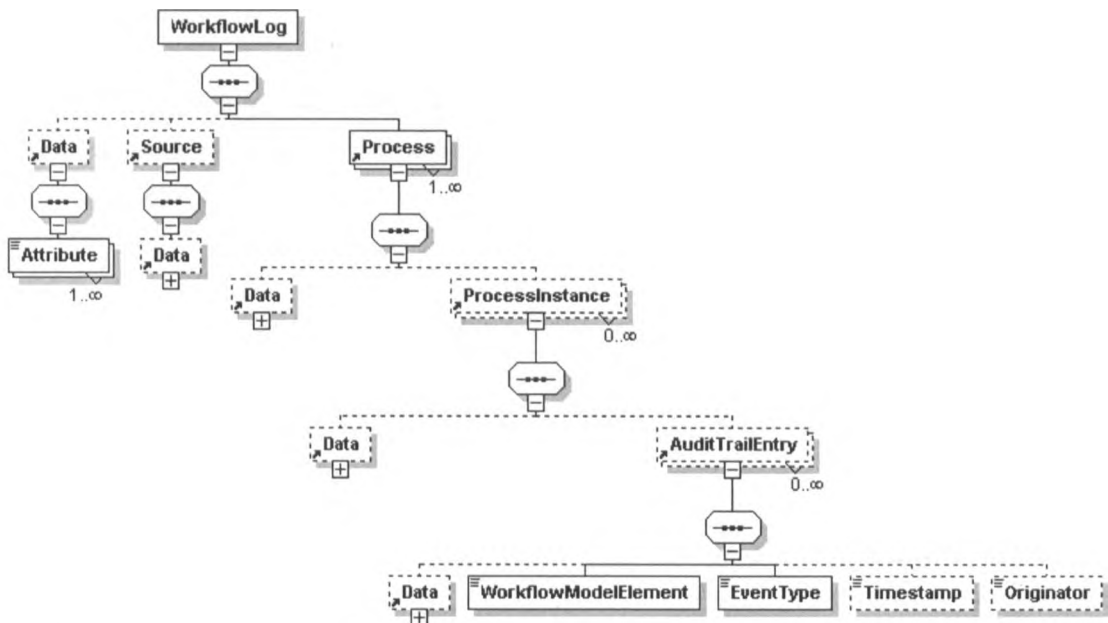


Figure 2: Structure of an MXML file

# Data Collection and Preparation

As the research is data-heavy a lot of emphasis was placed on data collection & preparation for its analysis.

Below is the detailed description of the procedures and techniques that were used in the data collection, conversion and eventual analysis. As the main objective of the paper is to come up with a framework the procedures below should be given extra attention as they give the information required to carry out this exercise given any system.

## Data Collection

The first thing is to understand the makeup of the system's database and especially the files or tables that contain the workflow information & audit trails.

The system currently under study has a web-based front-end, and a back-end of Microsoft SQL database. It is a procurement system currently in use in a professional services firm. It has several modules; however the one in use currently is the LPO approval process. Therefore the process that can be derived from it is as follows,

1. Raising the LPO,
2. Submitting the LPO for endorsement to the first partner
3. Endorsement of the LPO
4. Request for approval to the second partner
5. Approval of the LPO

The events that signify that each task has completed are as follows;

1. Submitted for Endorsement
2. Endorsed
3. Approved

With "Submitted for Endorsement" is the event that signifies the start of the process and "Approved" signifies the end of the process. After the approval the LPO is printed and given to the supplier.

The system contains several workflow tables, however after thoroughly understanding what each table does we picked up on the following three tables which were of interest to the research;

1. workflow\_history
2. workflow\_document
3. workflow\_definition

*work\_history* – this is the main audit trail table, it contains the tasks, timestamp & originator id

Its definition is as follows;

```
[workflow_history_id] [int] IDENTITY(1,1) NOT NULL, PK
[document_id] [int] NOT NULL, FK
[workflow_definition_id] [int] NULL, FK
[workflow_date] [datetime] NULL,
[current_status_description] [varchar](50) NULL,
[workflow_action] [varchar](50) NULL,
[staff_id] [varchar](50) NULL,
[comments] [varchar](1000) NULL
```

It has all the required parameters for analysis except for the eventtype(start,complete) which we had to add for purposes of the process instance beginning & end being recognized automatically. The other parameters are; process instance id, audit trail id, timestamp, originator, instance action.

*workflow\_documents* – This is the documents table, when a new document is raised, its record is stored here e.g when it was created, the person who initiated it etc. when it undergoes tasks such as requests for approvals, endorsement & approvals, this information is saved in the *workflow\_history* table but the main connecting factor is the document id. This table was used as basis for the cases (process instances) because of this tying factor.

Its definition is as follows;

```
[workflow_document_id] [int] IDENTITY(1,1) NOT NULL, PK
[item_category_id] [int] NULL, FK
[workflow_document_source_id] [int] NULL, FK
[workflow_job_code] [varchar](100) NULL,
[workflow_document_initiator] [varchar](255) NULL,
[workflow_document_date_created] [datetime] NULL CONSTRAINT [DF_WorkFlow_Documents_date_created]
DEFAULT (getdate()),
[workflow_document_date_due] [datetime] NULL,
[workflow_document_current_state_id] [int] NULL,
[workflow_document_priority] [int] NULL,
[workflow_job_codew] [varchar](100) NULL
```

E.g. In the image below the document\_id field is the one that contains the unique id for the audit trail record; the document\_id is similar for all the tasks executed for that document.

1	workflow_history_id	document_id	workflow_definition_id	workflow_date	workflow_action	staff_id
2	564	78	41	3/8/2010 7:19	Declined	KE000359
3	563	78	42	3/8/2010 7:18	Submitted for Endorsement	KE000359
4	565	78	42	3/8/2010 7:23	Submitted for Endorsement	KE000359
5	566	78	43	3/8/2010 7:23	Endorsed	KE000359
6	616	78	44	3/8/2010 10:56	Approved	KE000229
7	567	79	41	3/8/2010 7:39	Submitted for Endorsement	KE000359

**Table 2: Identifying cases**

Therefore this is the table to be used to get the process instances. The parameters required for analysis from this table were only two; document\_id and description. However, since there was no description available from the table thus we used the document\_id as description.

*workflow\_definitions* – this table contains the definitions of the processes

Its definition is as follows;

```
[workflow_definition_id] [int] IDENTITY(1,1) NOT NULL, PK
[workflow_id] [int] NULL, FK
[item_category_id] [int] NOT NULL, FK
[state_name] [varchar](50) NOT NULL,
[state_order] [int] NOT NULL CONSTRAINT [DF_Item_Category_WorkFlow_state_order] DEFAULT ((0)),
[state_enabled] [int] NOT NULL CONSTRAINT [DF_Item_Category_WorkFlow_state_enabled] DEFAULT ((1)),
[state_owner] [varchar](50) NULL,
[state_is_final_approval] [int] NOT NULL CONSTRAINT
[DF_Item_Category_WorkFlow_state_is_final_approval] DEFAULT ((0)),
[go_back_state_id] [int] NULL,
[go_next_state_id] [int] NULL,
[go_back_advice_description] [varchar](50) NULL,
[go_next_advice_description] [varchar](50) NULL,
[current_status_description] [varchar](50) NOT NULL,
[go_back_document_result_status] [varchar](50) NULL,
[go_next_document_result_status] [varchar](50) NULL,
```

E.g. when an approval is declined it is indicated within the same process definition, a new definition is not created for that event, therefore the field relating to declines and deletions has to be added for that purpose.

After analysis of the database and how the tables relate to each other we went to data cleanup. We exported the data to Ms Excel for the cleanup.

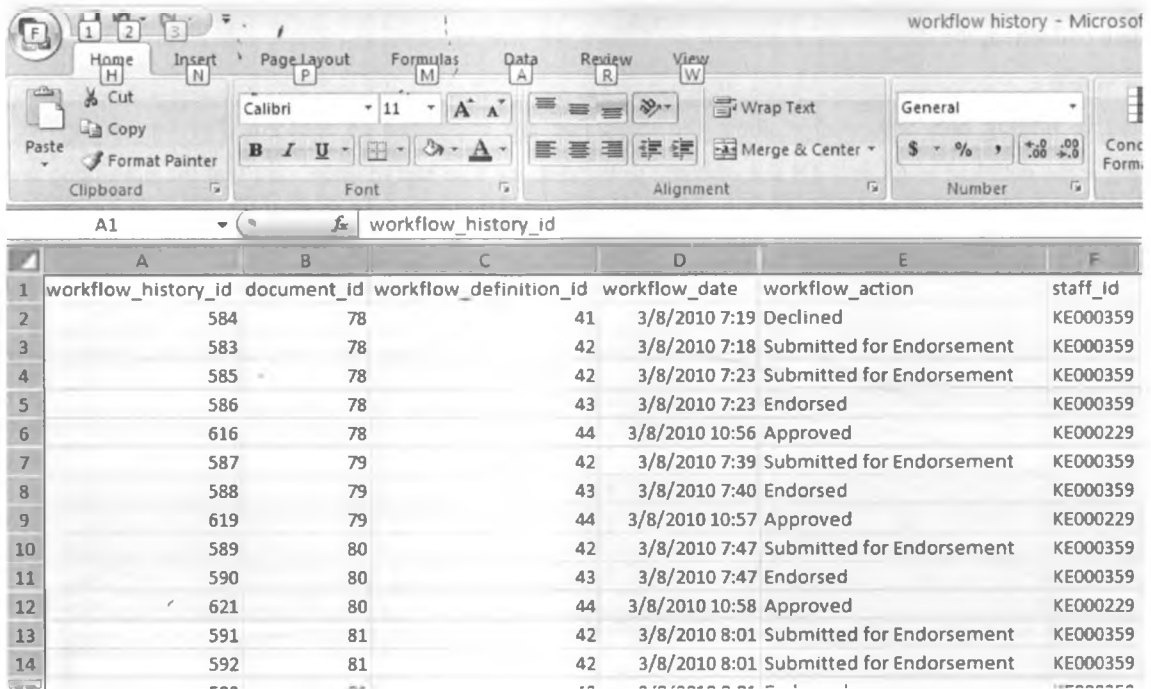
## Event logs data cleanup

After data collection, the data has to be cleaned up. Before applying any mining technique to an event log, you may want to remove unnecessary information from the log before you start the mining. For instance, you may be interested in mining only information about the cases that are completed or conform. For the system we are analysing some data had to go as it did not conform to the rest, we also had to make sure that the timestamp was in the correct format. We used Excel to do most of the cleanup as we could easily isolate what we did not want. The cleaning step is usually a projection of the data to consider only the data you are interested in.

The next step was to remove the columns that were unnecessary for mining ensuring that the data does not lose integrity. Below is an image of the workflow\_history data before and after the cleanup.

ID	Timestamp	Status	Description
583.78.42.2010-03-08 07:18:55.560	2010-03-08 07:18:55.560	Awaiting Endorsement, Submitted for Endorsement, KE000359	Access control systems on Rwanda office. This is per your quotation Proj/kgj/KCB/08-1
581.78.42.2010-03-08 07:18:55.560	2010-03-08 07:18:55.560	Awaiting Endorsement, Submitted for Endorsement, KE000359	Access control systems on Rwanda office. This is per your quotation Proj/kgj/KCB/08-1
584.78.41.2010-03-08 07:19:55.210	2010-03-08 07:19:55.210	Awaiting Initiator Amendment Declined, KE000359	
585.78.42.2010-03-08 07:23:19.850	2010-03-08 07:23:19.850	Awaiting Endorsement, Submitted for Endorsement, KE000359	Ok now connected
586.78.43.2010-03-08 07:23:35.827	2010-03-08 07:23:35.827	Awaiting Approval, Endorsed, KE000359	
587.78.42.2010-03-08 07:39:31.810	2010-03-08 07:39:31.810	Awaiting Endorsement, Submitted for Endorsement, KE000359	Fire alarms systems accessories and installation. Payment terms is 50% on order and balance on commissioning
588.79.43.2010-03-08 07:40:04.983	2010-03-08 07:40:04.983	Awaiting Approval, Endorsed, KE000359	This is in order as per quote Proj/kgj/KCB/10-1
589.80.42.2010-03-08 07:47:35.130	2010-03-08 07:47:35.130	Awaiting Endorsement, Submitted for Endorsement, KE000359	CCTV systems as per your quote Proj/kgj/KCB/010-1. Payment terms 50% with order, balance on commissioning
590.80.43.2010-03-08 07:47:44.970	2010-03-08 07:47:44.970	Awaiting Approval, Endorsed, KE000359	This is in order
591.81.42.2010-03-08 08:01:05.913	2010-03-08 08:01:05.913	Awaiting Endorsement, Submitted for Endorsement, KE000359	Office works on Rwanda office
592.81.42.2010-03-08 08:01:05.913	2010-03-08 08:01:05.913	Awaiting Endorsement, Submitted for Endorsement, KE000359	Office works on Rwanda office
593.81.43.2010-03-08 08:01:05.913	2010-03-08 08:01:05.913	Awaiting Approval, Endorsed, KE000359	This is ok
594.83.42.2010-03-08 08:13:24.057	2010-03-08 08:13:24.057	Awaiting Endorsement, Submitted for Endorsement, KE000359	Quotation on supplying and fixing on cabinets, floor and doors
595.83.43.2010-03-08 08:13:29.240	2010-03-08 08:13:29.240	Awaiting Approval, Endorsed, KE000359	
596.84.42.2010-03-08 08:34:13.047	2010-03-08 08:34:13.047	Awaiting Endorsement, Submitted for Endorsement, KE000359	Power cabling and access control, UPS and PABX installation
597.84.43.2010-03-08 08:34:24.467	2010-03-08 08:34:24.467	Awaiting Approval, Endorsed, KE000359	This is ok
598.85.42.2010-03-08 08:37:54.383	2010-03-08 08:37:54.383	Awaiting Endorsement, Submitted for Endorsement, KE000359	Access flooring material as per proforma 16-02-2009-001
599.85.43.2010-03-08 08:38:01.873	2010-03-08 08:38:01.873	Awaiting Approval, Endorsed, KE000359	This is ok
600.86.42.2010-03-08 08:41:33.140	2010-03-08 08:41:33.140	Awaiting Endorsement, Submitted for Endorsement, KE000359	
601.86.43.2010-03-08 08:41:37.680	2010-03-08 08:41:37.680	Awaiting Approval, Endorsed, KE000359	
602.87.42.2010-03-08 08:48:37.470	2010-03-08 08:48:37.470	Awaiting Endorsement, Submitted for Endorsement, KE000359	Tea area cabinets as per discussion with our Chao Sio

Figure 3: Data before cleanup



	A	B	C	D	E	F
1	workflow_history_id	document_id	workflow_definition_id	workflow_date	workflow_action	staff_id
2		584	78	41	3/8/2010 7:19 Declined	KE000359
3		583	78	42	3/8/2010 7:18 Submitted for Endorsement	KE000359
4		585	78	42	3/8/2010 7:23 Submitted for Endorsement	KE000359
5		586	78	43	3/8/2010 7:23 Endorsed	KE000359
6		616	78	44	3/8/2010 10:56 Approved	KE000229
7		587	79	42	3/8/2010 7:39 Submitted for Endorsement	KE000359
8		588	79	43	3/8/2010 7:40 Endorsed	KE000359
9		619	79	44	3/8/2010 10:57 Approved	KE000229
10		589	80	42	3/8/2010 7:47 Submitted for Endorsement	KE000359
11		590	80	43	3/8/2010 7:47 Endorsed	KE000359
12		621	80	44	3/8/2010 10:58 Approved	KE000229
13		591	81	42	3/8/2010 8:01 Submitted for Endorsement	KE000359
14		592	81	42	3/8/2010 8:01 Submitted for Endorsement	KE000359

Figure 4: Data after cleanup

## Conversion of the event logs to MXML

For data to be uploaded in ProM, it has to be in MXML format. MXML is a **pure C** library that is meant to help developers implementing XML file interpretation in their projects. The compact design is easy to put it in any project, for it is very small, an average program will grow from 15 to 30 kb when it is included.

There were several solution to convert the data, however we chose to use Ms Access database as an in between since it there was an available plugin that could change this data to MXML. With Ms Access the data has to be prepared in a manner that can be easily converted to MXML. Note that MXML is node-based thus the data had to be in a manner that it can be picked as nodes and child nodes within it.

To convert data in a Microsoft Access database about cases and tasks that have been executed to the ProM MXML format, *four* tables have to be defined which have a similar structure to that of the fields in the MXML format. The elements in the MXML format that can contain information about cases and tasks that have been executed are respectively the *Process Instance* element and the *Audit Trail Entry* element. Furthermore, both the *Process Instance* element and the *Audit Trail Entry* element can have *Data* as sub element which can contain respectively additional information about process instances and audit trail entries. An image of the MXML format can be found is above.



Therefore, the first table is *Process\_Instances* which needs to be filled with the identifier of a certain process instance (field *PI-ID*) and, if available, its accompanying description (field *description*). Furthermore, it is important to note that the *PI-ID* field has to be a primary key in the table. The second table *Data\_Attributes\_Process\_Instances* needs to be filled with additional information about each process instance, the so called data attributes. Therefore, this table contains the fields *PI-ID*, *Name* and *Value*. The *PI-ID* field is needed to identify to which process instance each data attribute belongs (actually, this is a foreign key for the *PI-ID* field in table *Process\_Instances*). Also, each process instance can have zero or more data attributes that belong to it. Furthermore, the field *name* represents the name of the data attribute and the field *value* represents the value of the data attribute.

The third table *Audit\_Trail\_Entries* needs to be filled with data about tasks that have been performed during the execution of a process instance. Not surprisingly, this table contains fields with name *WFMEIt* (for the name of the task), *EventType* (the task event type, e.g. start, complete), *Timestamp* (the time in which the task changed its state), and *Originator* (the person or system that caused the change in the task state). However, we also have the columns *PI-ID* and *ATE-ID*. The field *ATE-ID*, which is a unique identifier for each audit trail entry (so, this has to be a primary key in the table). The reason for introducing this field is because additional information about each audit trail entry that is relevant could exist, but does not fit in the other fields of table *Audit\_Trail\_Entries*. For this additional information, we have table *Data\_Attributes\_c* which is set up in a similar way as table *Data\_Attributes\_Process\_Instances*. Another reason for introducing the *ATE-ID* field is because it is needed in table *Data\_Attributes\_Audit\_Trail\_Entries* to be able to identify to which audit trail entry each data attribute belongs to.

After creating the above tables, we populated them as follows;

*Process\_Instances (PI-ID, Description)*: we put *document\_id* for both fields as we lacked the description

*Audit\_Trail\_Entries (ATE-ID,PI-ID,WFMEIt,EventType,Timestamp,Originator)*: we put the necessary data from the *workflow\_history* table which contained this information.

A populated *audit\_trail\_entries* table in Ms Access

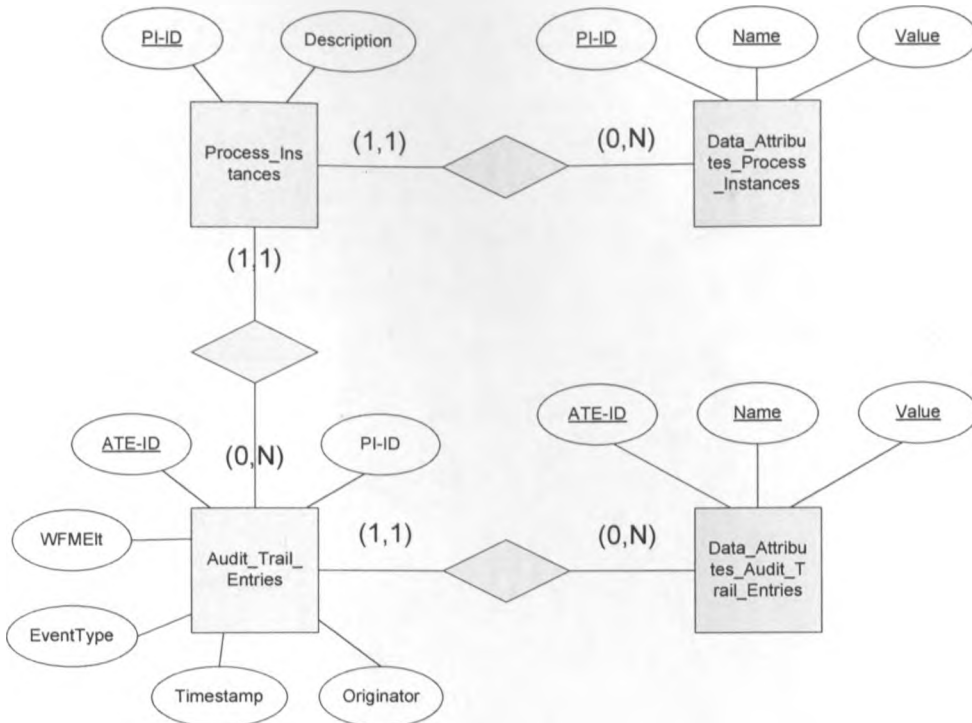
ATE-ID	PP-ID	WFMEIT	EventType	Timestamp	Originator
584	78	Submitted for Endorsement	start	3/8/2010 7:58:00 AM	KE000359
585	78	Declined		3/8/2010 7:59:00 AM	KE000359
585	78	Submitted for Endorsement	start	3/8/2010 7:59:00 AM	KE000359
586	78	Submitted for Endorsement	start	3/8/2010 7:59:00 AM	KE000359
586	78	Endorsed		3/8/2010 7:59:00 AM	KE000359
587	79	Submitted for Endorsement	start	3/8/2010 7:59:00 AM	KE000359
588	79	Endorsed		3/8/2010 7:59:00 AM	KE000359
589	80	Submitted for Endorsement	start	3/8/2010 7:59:00 AM	KE000359
590	80	Endorsed		3/8/2010 7:59:00 AM	KE000359
591	81	Submitted for Endorsement	start	3/8/2010 8:00:00 AM	KE000359
592	81	Submitted for Endorsement	start	3/8/2010 8:00:00 AM	KE000359
593	81	Endorsed		3/8/2010 8:00:00 AM	KE000359
594	83	Submitted for Endorsement	start	3/8/2010 8:13:00 AM	KE000359
595	83	Endorsed		3/8/2010 8:13:00 AM	KE000359
596	84	Submitted for Endorsement	start	3/8/2010 8:34:00 AM	KE000359
597	84	Endorsed		3/8/2010 8:34:00 AM	KE000359
598	85	Submitted for Endorsement	start	3/8/2010 8:37:00 AM	KE000359
599	85	Endorsed		3/8/2010 8:38:00 AM	KE000359
600	86	Submitted for Endorsement	start	3/8/2010 8:41:00 AM	KE000359
601	86	Endorsed		3/8/2010 8:41:00 AM	KE000359
602	87	Submitted for Endorsement	start	3/8/2010 8:48:00 AM	KE000359
603	87	Endorsed		3/8/2010 8:48:00 AM	KE000359
604	88	Submitted for Endorsement	start	3/8/2010 8:56:00 AM	KE000359
605	88	Endorsed		3/8/2010 8:56:00 AM	KE000359
606	89	Submitted for Endorsement	start	3/8/2010 8:59:00 AM	KE000359
607	89	Endorsed		3/8/2010 8:59:00 AM	KE000359
608	90	Submitted for Endorsement	start	3/8/2010 9:09:00 AM	KE000359
609	90	Endorsed		3/8/2010 9:09:00 AM	KE000359
610	91	Submitted for Endorsement	start	3/8/2010 9:12:00 AM	KE000359
611	91	Endorsed		3/8/2010 9:12:00 AM	KE000359
612	90	Approved	completed	3/8/2010 10:51:00 AM	KE000329
613	91	Approved	completed	3/8/2010 10:53:00 AM	KE000229

Figure 5: Audit\_Trail\_Entries Table

The next step was to search for tools that can do the conversion to MXML. There exists an open source program known as ProMImport, it is a tool that was developed to convert logs to the acceptable formats that ProM uses. We used it to convert the MsAccess database that we had created to MXML.

For *Data\_Attributes\_Process\_Instances*, *Data\_Attributes\_Audit\_Trail\_Entries* tables we did not require any additional information so we left them without any data. As can be seen from the image below, they do not necessarily have to have records in them. It depends on the system that is being analysed.

Figure 6 is the ERD of the process mining tables;



**Figure 6: Process mining table ERD**

The database is now ready for upload.

The next step was to ensure that an ODBC connection for the access database has to be setup on the computer. See appendix 1.

***The ProM Import Framework***

The ProM Import contains several plugins e.g. Apache2, general CSV file, PeopleSoft, which can be used to convert data of these formats to MXML format, however with Ms Access the plugin we used is the MS Access Database plugin that comes with the latest version 7.0. We entered the information as can be seen from the diagram below then run the program.

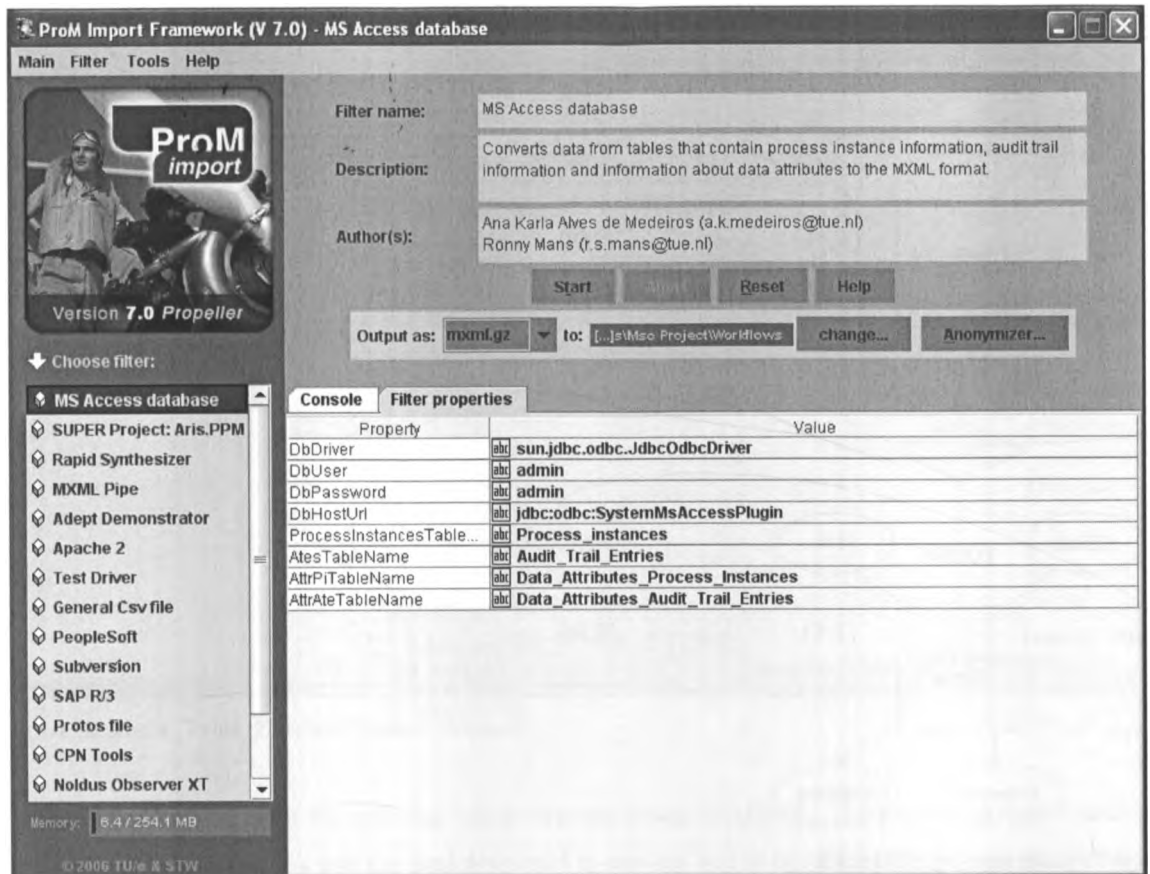


Figure 7: ProM Import Framework (V 7.0)

The MXML file was now ready for analysis.

I used ProM [5.2.] to do the process mining. It is an open source system which many researchers have worked on and is considered to be one of the best in the field. See the image below on loading the MXML log to ProM.

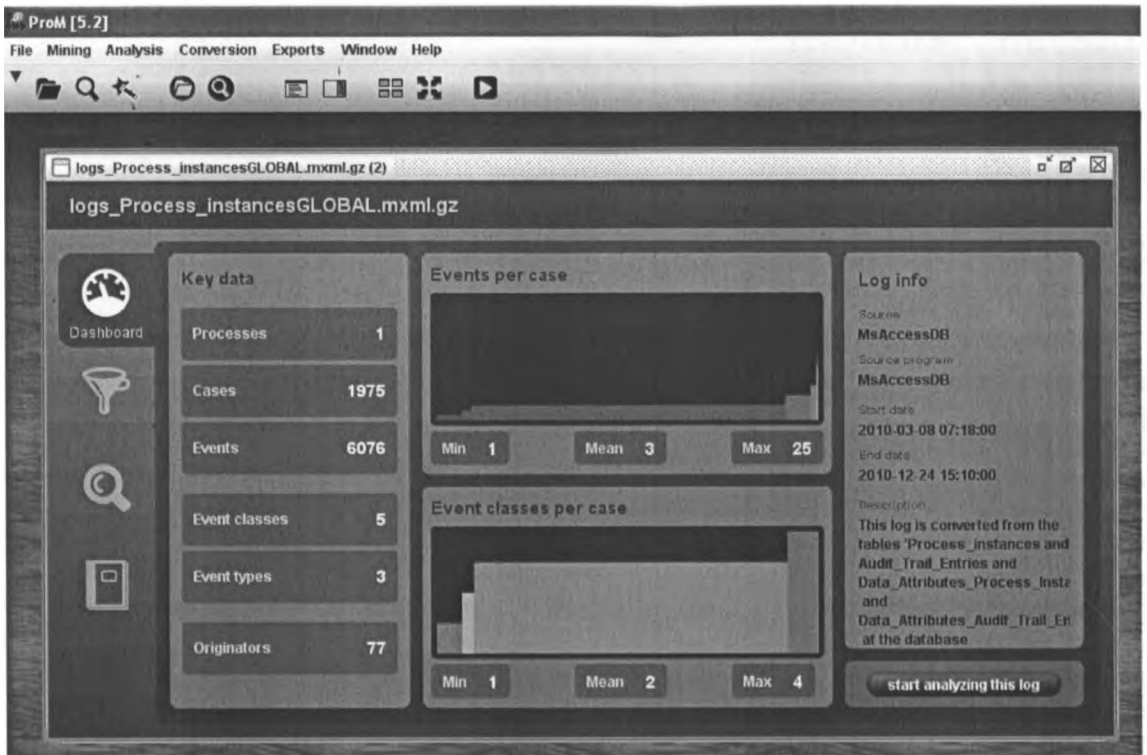
## Mining and Analysis

In this section we present some results obtained from analysis of the e-procurement logs. We concentrate on the information that can be derived using ProM.



Figure 8: Opening a Log file

Immediately after loading ProM does an initial filtering of data that quickly captures the information in the file and arranges it according to processes, cases and events. The MXML file has to have been created from properly prepared data otherwise ProM will read it incorrectly. See the image below for the dashboard that comes up automatically after the initial loading.



**Figure 9: Log Dashboard**

We can quickly deduce the following from the dashboard above;

No. of Processes	1
No. of Cases	1975
No. of Event Types	3
No. of Originators	77
Total no. of events	6076
Minimum no. of events per case	1
Maximum no. of events per case	25
Average no. of events per case	3
No. of Event Classes	5
Minimum no. of event classes	1
Maximum no. of events classes	4
Average no. of event classes	2
Absolute Start date of the log	2010-03-08 07:18
Absolute End date of the log	2010-12-24 15:10

**Table 3: Dashboard Statistics**

More analysis will reveal more information on these events. See appendix 2, which contains the log summary of the event log that we have uploaded into the ProM Framework. It is as a result of a default log filter that identifies the cases, whether they are complete or not and provides summaries of the entry logs.

Given the system I'm mining we can answer the following questions after mining;

1. What is the most frequent path for every process model?

This is the model that is followed most frequently by many cases. Using the Fuzzy Miner plug-in ProM has assisted in quickly coming up with the model and the numbers in each step indicate the frequency of the cases that pass through this route.

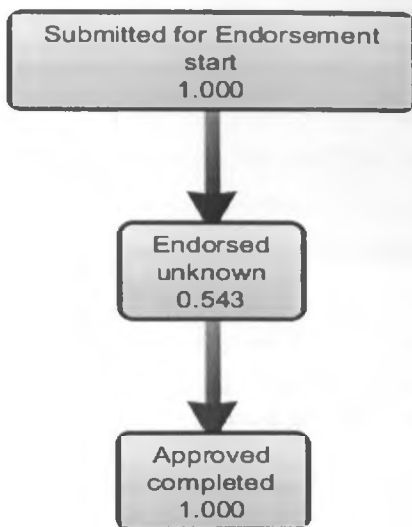


Figure 10: Process Model 1

Other models as mined are as follows; the one below shows many different phases many processes may follow especially if there at any one point a certain request is declined.

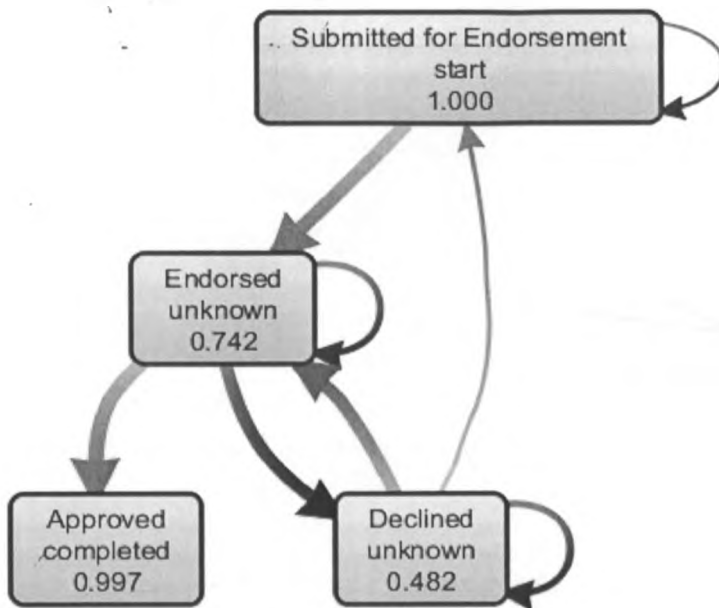


Figure 11: Process Model 2

2. How is the distribution of all cases over the different paths through the process?

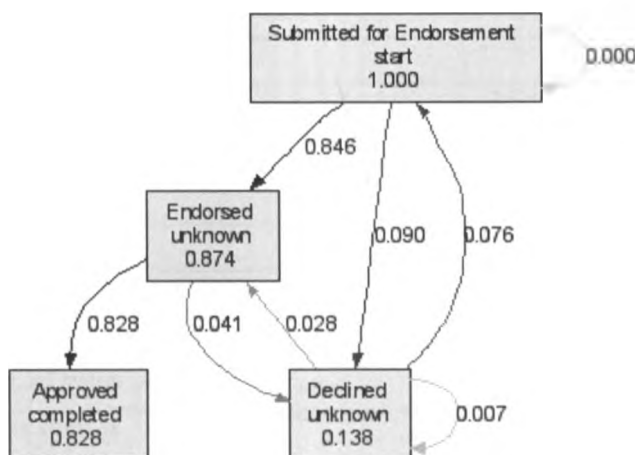


Figure 12: Distribution of Cases obtained from the frequency abstraction miner



The above diagram shows the following

From Activity	To Activity	Frequency
Submitted for Endorsement	Submitted for Endorsement	0%
Submitted for Endorsement	Endorsed	84.6%
Submitted for Endorsement	Declined	9%
Declined	Submitted for Endorsement	7.6%
Endorsed	Approved	82.8%
Endorsed	Declined	4.1%
Declined	Endorsed	2.8%
Declined	Declined	0.7%

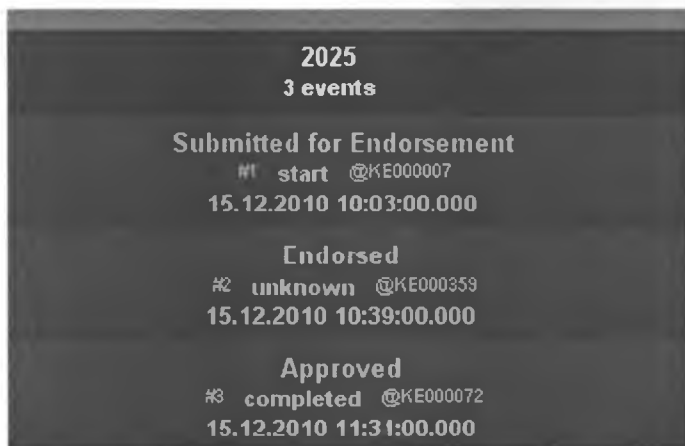
**Table 4: Distribution of cases**

- How compliant are the cases (i.e. process instances) with the deployed process models? Where are the problems? How frequent is the (non-) compliance?

The process model set out in the system as explained earlier is as follows;

- Raising the LPO
- Submitting the LPO for endorsement to the first partner, the partner may decline at this point and it goes back to the initiator of the LPO, the initiator may submit it again or it ends at this point
- Endorsement of the LPO
- Request for approval to the second partner, the partner may decline and the LPO goes back to the endorser. It may be endorsed again or it ends at this point
- Approval of the LPO

Most of the cases are compliant; they follow the most frequent path of 3 tasks in the process, i.e. submission for endorsement, endorsement, approval, as in the diagram below;



**Figure 13: compliant case following the most frequent path**

Other cases whereby there is decline follow the process as in the example below;

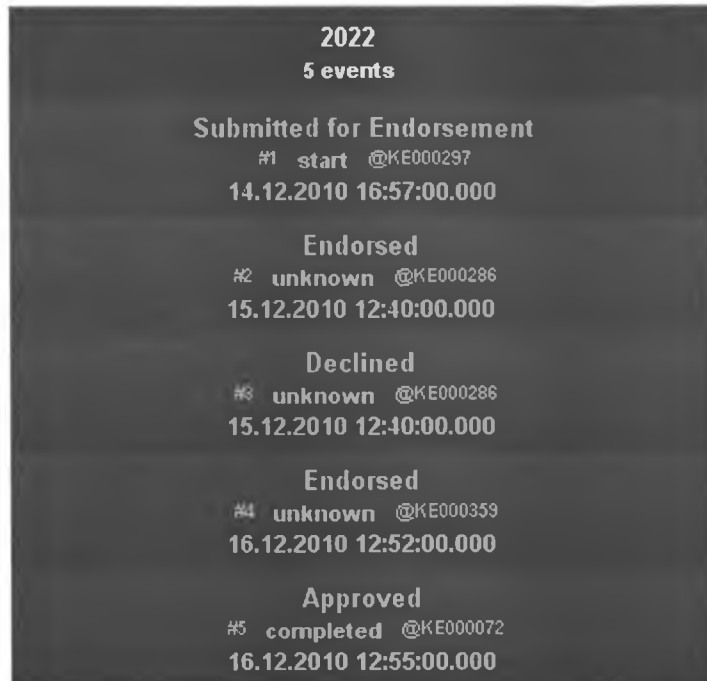
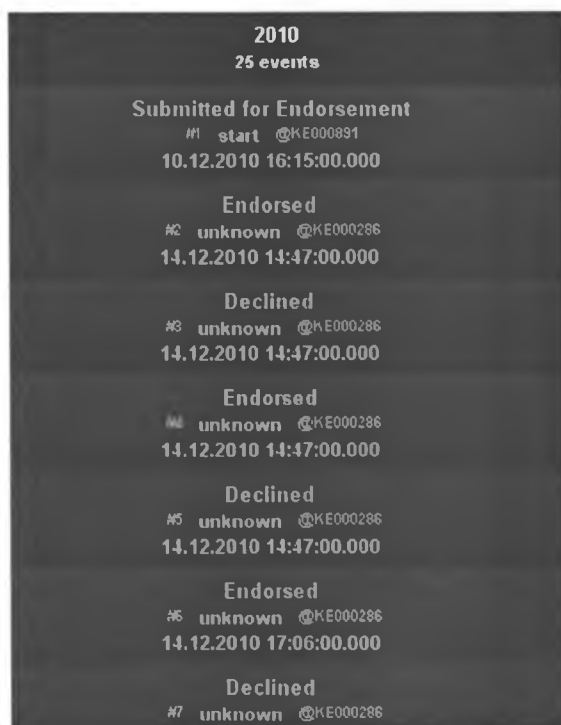


Figure 14: Compliant case with the path covered by some of the cases

However, there are a few exceptions. As in case no. 2018, there were 9 events. This is unusual as there were several tasks for endorsing and declining by the same partner. This cannot be the case since if a partner endorses the LPO goes to another partner for an approval. Another exception noted was case no. 2010, which had 25 events of the same nature as the one above. See the diagram below;

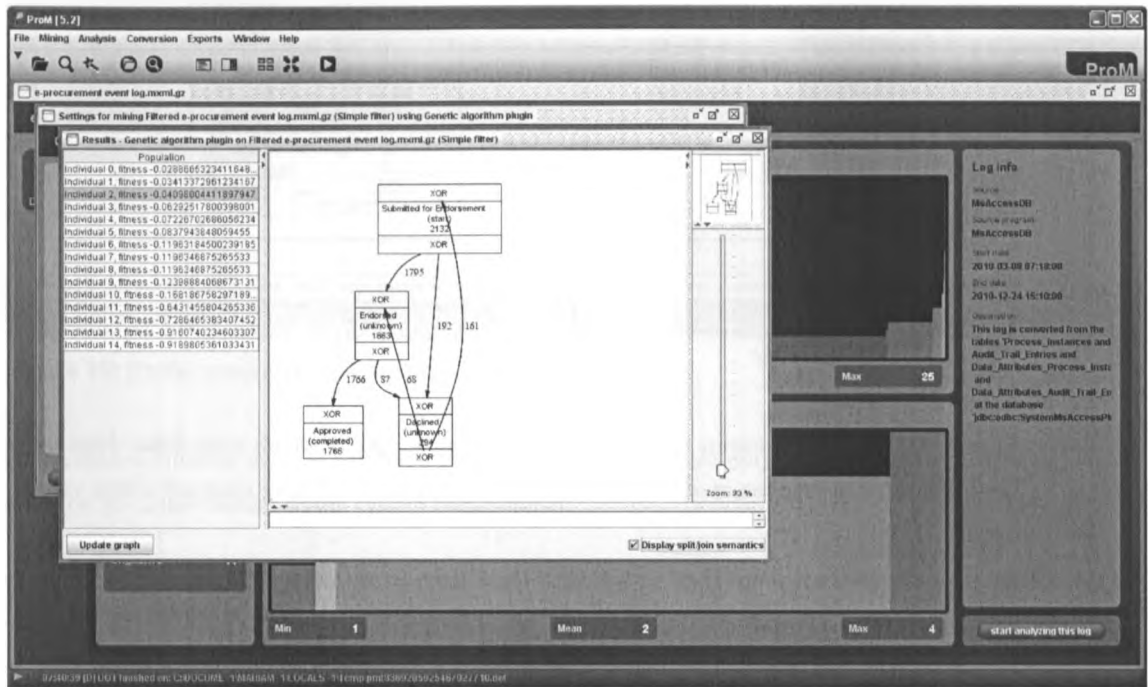


**Figure 15: Non-compliant case with 25 events, depicting abnormality**

Abnormalities are an indication of a bug, fraud etc. The frequency of the exceptions was however not much in this log.

4. What are the routing probabilities for each split task (XOR or OR split/join points)?

The diagram below indicates the routing probabilities of the events in the process. E.g. at the start stage an event can either be endorsed or declined.



**Figure 16: Routing probabilities**

5. Are the events following laid down policies

Organizations have different procedures which must be followed, e.g. a person cannot approve his own request. In the diagram below, the plug-in that assists in checking whether the policies have been followed is shown, it is known as LTL checker.

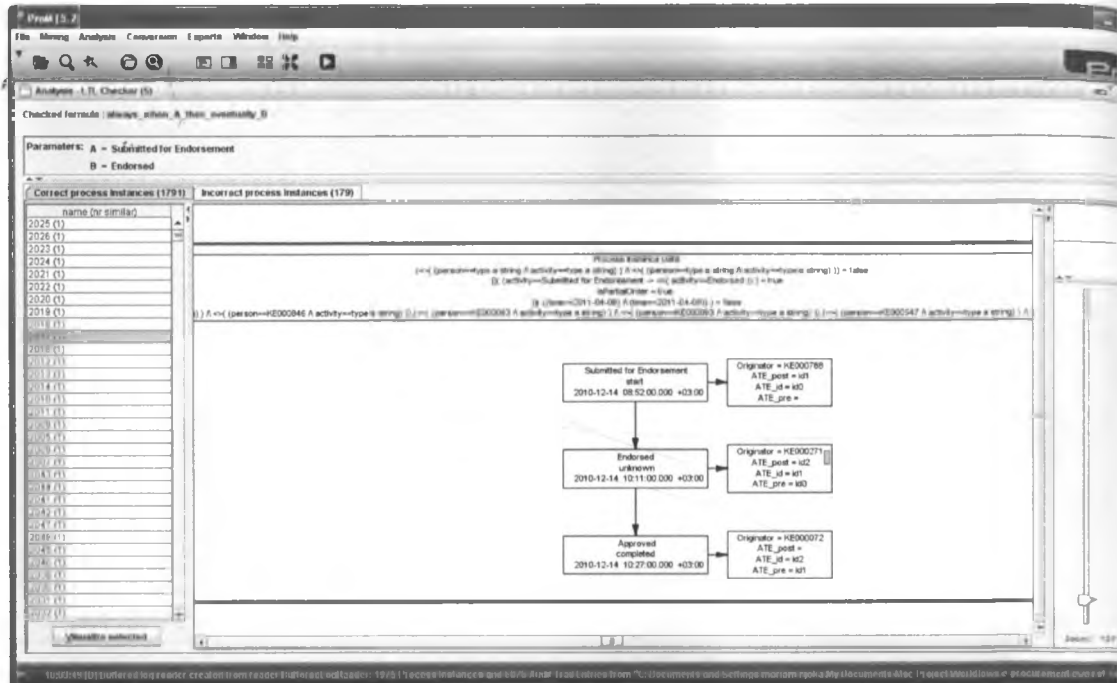


Figure 17: LTL checker

In this case, given two events that follow one another, it checks which cases have these events follow one another. 1791 have returned a true value while 179 of the cases have returned a false value.

- Which paths take too much time on average? How many cases follow these routings? What are the critical sub-paths for these paths?

Using the performance sequence diagram, 19 patterns came up and upon filtering off those patterns that time in minutes less than zero and those that had only one case, we were left with 15 patterns. This shows that the cases can follow any of the 15 patterns, the requirement here is to evaluate what patterns are acceptable and get rid of the unacceptable ones.

E.g. in the diagram below pattern one takes an average of 59 hours and follows the most frequent path shown in figure 10.

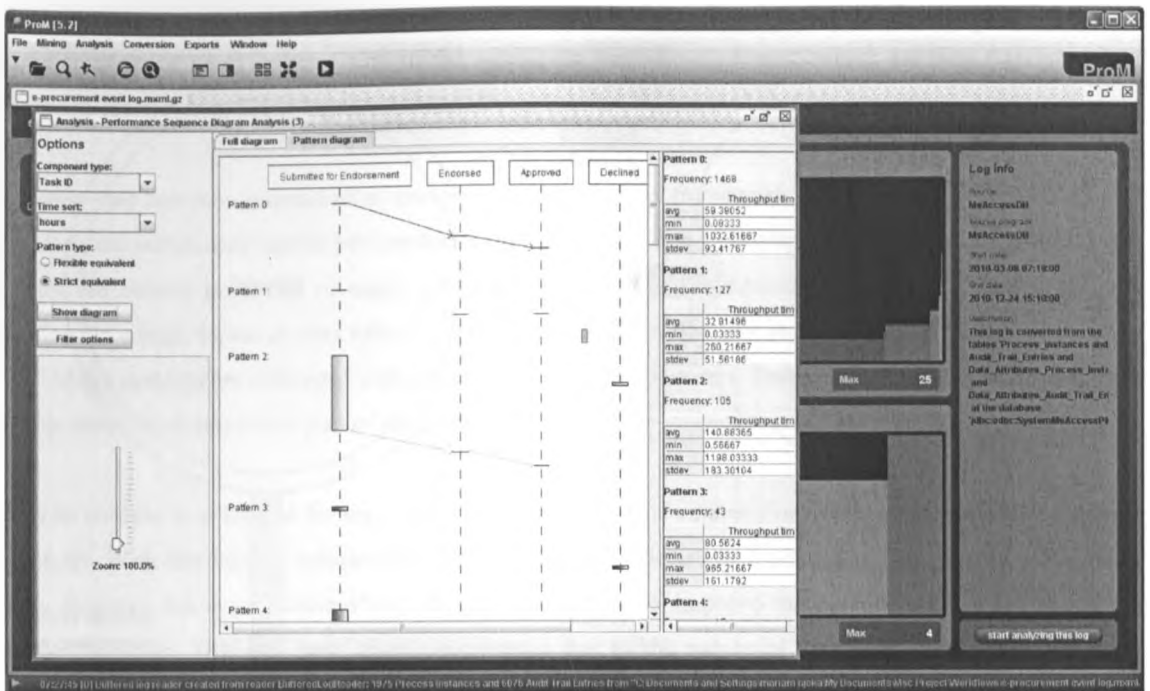


Figure 18: Performance on different patterns

Given the information derived using the above mining and analysis plug-ins, then recommendations can be made on the improvement of the system being mined.

### The Proposed Framework

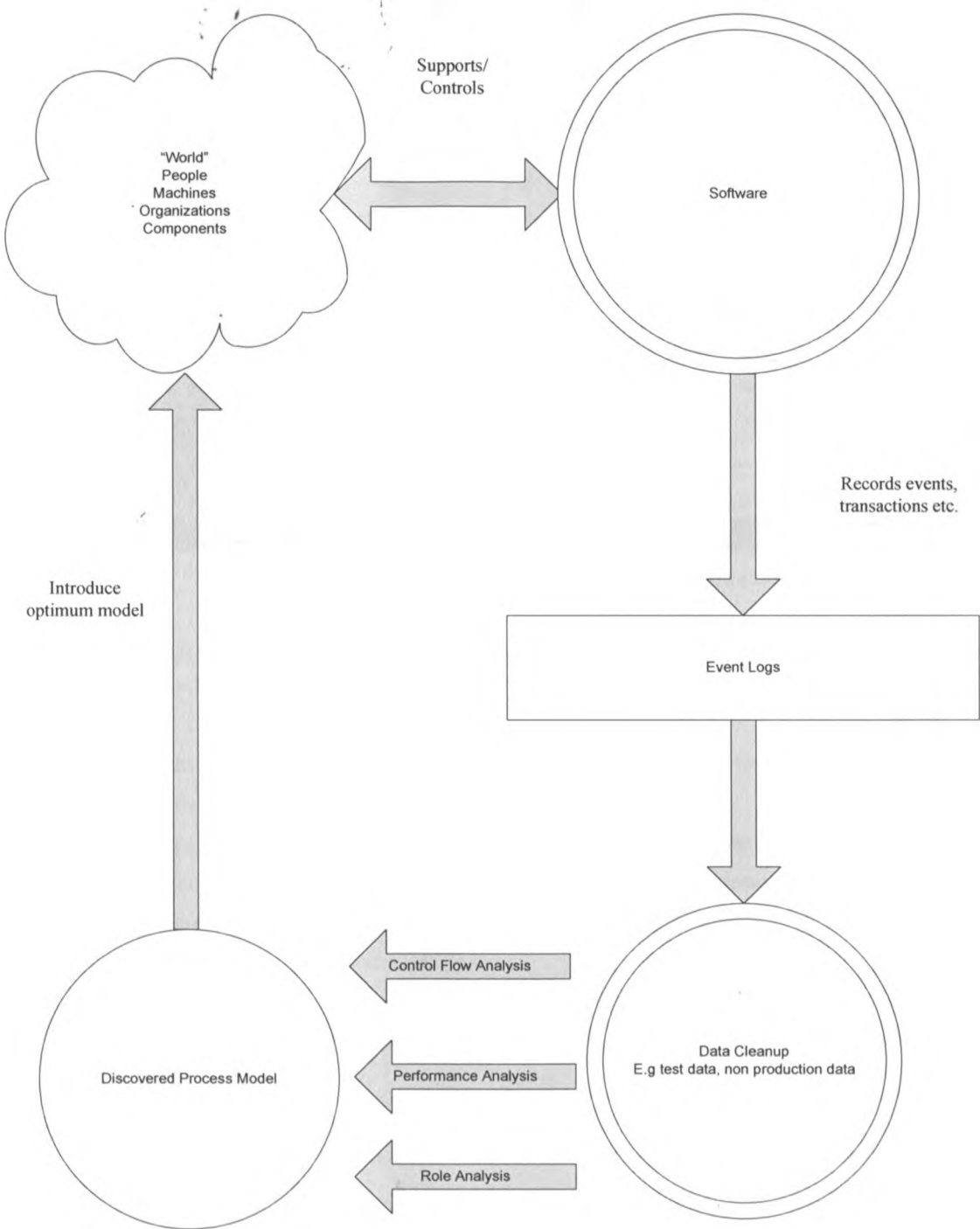


Figure 19: The proposed framework

The proposed framework takes in the vision of the world, the business processes, which are then converted into systems. Through the systems, event logs are created. Event logs are then mined and the next step is conducted.

The data cleanup step has been introduced as compared to the previous framework. Given that most semi structured systems contain data which may not be just production data, e.g. test data. Cleanup in this case becomes prominent to ensure only the correct processes or cases are picked for analysis. This handicap in semi-structured systems makes data cleanup a huge aspect as very often there will be data which does not conform perhaps it was testing data or otherwise. Many systems are debugged and corrected on live environments. Thus cleanup and preparation of data should be considered as an important part of the process.

The best way to analyse is to look at the logs with a view of getting the *control flow analysis*, *performance analysis* and *role analysis*. With this kind of information, then all manner of decisions touching on how to control the work flow, how to improve the organizational/role allocation and how to improve the performance of the processes themselves can be done. This has been given prominence due to the nature of discovery that is done on semi structured systems. These systems in many cases will not have had previous statistical information of these crucial data. Considering there may be no documented process models to follow, or even performance tests done on the system previously and recorded then these analysis methods will give us the information.

#### **Control flow analysis**

The data is analyzed to discover all the different process patterns from the event logs. These are then analyzed to get the most optimum one. This will give us the different paths that the processes in the system follow.

E.g.

Submit for endorsement – Decline – Submit for endorsement – Endorsed – Approved

Submit for endorsement – Endorsed – Approved

All these paths can then be analyzed in terms of frequency, correctness among others.

#### **Performance analysis**

Given the timestamp available in logs, information such as average time for the process, average times per pattern, minimum, maximum etc are discovered. In performance analysis all the aspects of what could affect the success of the process in analyzed. Things to be considered are like the length of the path a process follows, the time it takes to complete, the average for the processes.

With this kind of statistics, simulations can be done and the most optimum model derived.

### **Role analysis**

With the tools that discover the relationships between the people carrying out the activities, detailed role analysis can be carried out and streamlined to improve the process. E.g. Roles being carried out by several people may show some inefficiencies thus reductions can be done etc.

In addition, workloads and the flow from one person to another can be analyzed, with this knowledge, organizations are then able to tell the bottlenecks and steps are taken to clear them.

The result of the analysis is the discovered process model which will be incorporated into the organization either through a soft process change or business process reengineering given the weight. The software will have to be changed accordingly to reflect the new business process and the cycle continues.



## Implications of Research

A process model can have positive or negative effects on an organization. If properly designed then efficiencies are increased otherwise there are many organizations that have introduced computerized systems and are yet to feel the benefits due to the poor structure of the process model on that system.

Given that many semi-structured systems did not follow a structured way when they were developed, then its most likely they will be having a negative effect on the organization they are running in.

This research can be used to assist many organizations in cleaning up their systems and coming up with optimal processes to run their operations.

# Conclusion

In this research, the concentration has been on the process of process mining and how applicable it can be using the context of a procurement system. It takes you through the process of data collection and its preparation thereof. It also describes how to convert the data to the MXML format acceptable in the ProM framework. The research takes you through the process of mining and analysis and the result is the discovery of the optimum process mining framework by going through control flow analysis, performance analysis and role analysis using the tools available in the ProM Framework. These tools assisted in production different types of information such as model patterns, frequencies etc.

The literature review gives a good understanding of process mining, the ProM framework and MXML technologies. A good understanding of the technologies is required to implement process mining. Further research can be carried out as it is a relatively new area. Future work can focus on developing more tools that can be used to analyze entry logs and come up with meaningful information.

This research has demonstrated the implementation of process mining on a system that is already in use, this has led to coming up with a framework that can be used to analyze and eventually improve these systems. The framework introduced was initially derived from the framework developed by R.S. Mans, M.H. Schonenberg, M. Song, W.M.P. van der Aalst, and P.J.M. Bakker. The framework looks at the special circumstance of semi-structured systems which are common in the developing world. It is needed urgently in streamlining processes within the application systems developed locally due to their lack of structure during development.

By following this framework, systems developers will be able to analyze the impact of their systems after they have gone into production and therefore come up with better process models and improvements. It will assist in structuring systems that have already started being used.

## Bibliography

1. Peter Van Den Brand. (2004). *Architecture of the process mining framework*
2. B.F. van Dongen, A.K.A. de Medeiros, H.M.W. Verbeek, A.J.M.M. Weijters and W.M.P. van der Aalst. (2005). *The ProM framework: A new era in process mining tool support*, 26th International Conference on Applications and Theory of Petri Nets, G. Ciardo and P. Darondeau, LNCS 3536, pages 444-454, 2005
3. <http://prom.win.tue.nl/tools/prom/> by the Process Mining Group, Eindhoven Technical University.
4. W.M.P. van der Aalst and A.J.M.M. Weijters. *Process Mining: A Research Agenda*. Department of Technology Management, Eindhoven University of Technology, P.O. Box 513, NL-5600 MB, Eindhoven, The Netherlands.
5. W.M.P. van der Aalst and H.T. de Beer and B.F. van Dongen. *Process Mining and Verification of Properties: An Approach based on Temporal Logic*. Department of Technology Management, Eindhoven University of Technology, P. O. Box 513, NL-5600 MB, Eindhoven, the Netherlands.
6. Gregory A. Hansen. (1997). *Automating Business Process Reengineering, Second Edition*. Prentice hall PTR
7. W.M.P. van der Aalst and A.J.M.M. Weijters. *Process Mining: A Research Agenda*. Department of Technology Management, Eindhoven University of Technology, P.O. Box 513, NL-5600 MB, Eindhoven, The Netherlands.
8. W.M.P. van der Aalst and H.T. de Beer and B.F. van Dongen. *Process Mining and Verification of Properties: an Approach based on Temporal Logic*. Department of Technology Management, Eindhoven University of Technology, P. O. Box 513, NL-5600 MB, Eindhoven, The Netherlands.
9. Gregory A. Hansen. (1997) *Automating Business Process Reengineering, Second Edition*. Prentice hall PTR
10. Melike Bozkaya, Joost Gabriels, Jan Martijn van der Werf LaQuSo, *Process Diagnostics: a Method Based on Process Mining* Laboratory for Quality Software
11. Yannis Stavarakas 1, Manolis Gergatsoulis 1, and Panos Rondogiannis. *Multidimensional XML*. Institute of Informatics & Telecommunications
12. R.S. Mans<sup>1</sup>, M.H. Schonenberg<sup>1</sup>, M. Song<sup>1</sup>, W.M.P. van der Aalst<sup>1</sup>, and P.J.M. Bakker<sup>2</sup>. *Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital*.
13. Luca Rossetti (2006) What is Business Intelligence <http://searchdatamanagement.techtarget.com/definition/business-intelligence>

## Appendices

### Appendix 1: How to configure an ODBC connection to a Microsoft Access Database on your computer

The procedure to load set up an ODBC connection to a Microsoft Access database located at your computer is:

1. Open the "Control Panel" window by clicking on "Start/Setting/Control Panel".
2. Double-click "Administrative Tools".
3. Double-click "Data Sources (ODBC)".
4. Select the tab "System DSN".
5. Click on the button "Add..." of the "System DSN" tab. You should get a window like the one in Figure 12.

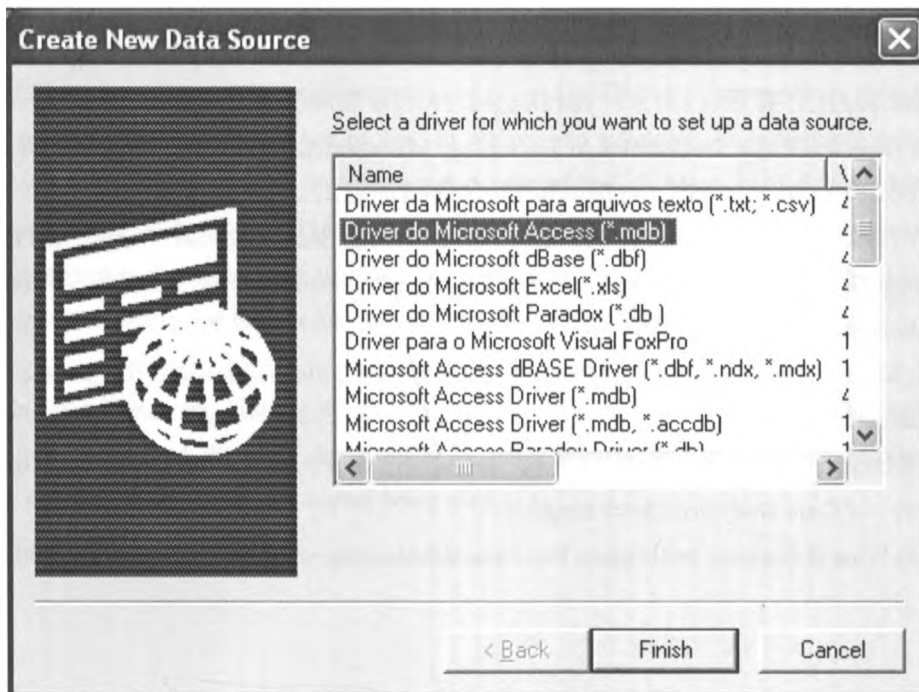


Figure 20: Screenshot of the window to create a new data source

6. Select "Driver to Microsoft Access (\*.mdb)" and click the button "Finish".
7. The next window is like the one in Figure 13. This window allows you to define the Provide the "Data Source Name", click on the button "Select..." to inform where the Microsoft Access database is located (see Figure 14), and on the button "Advanced..." to set up the "Username" and "Password" to be able to access this database. See Figure 15.

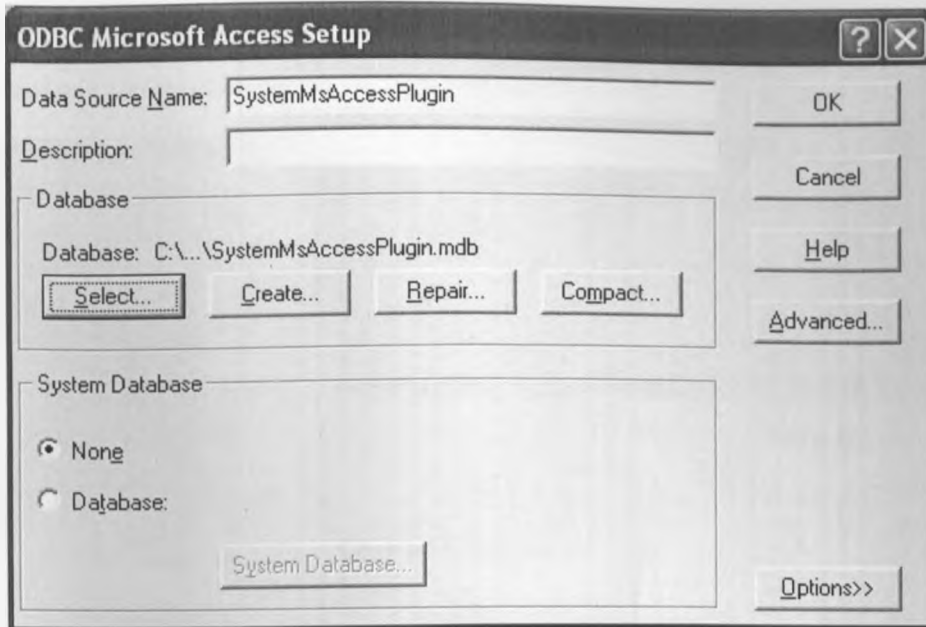


Figure 21: Screenshot of the window to set up the name of the database that has to be provided after jdbc:odbc: in the field DbHostUrl in the MS Access database plugin

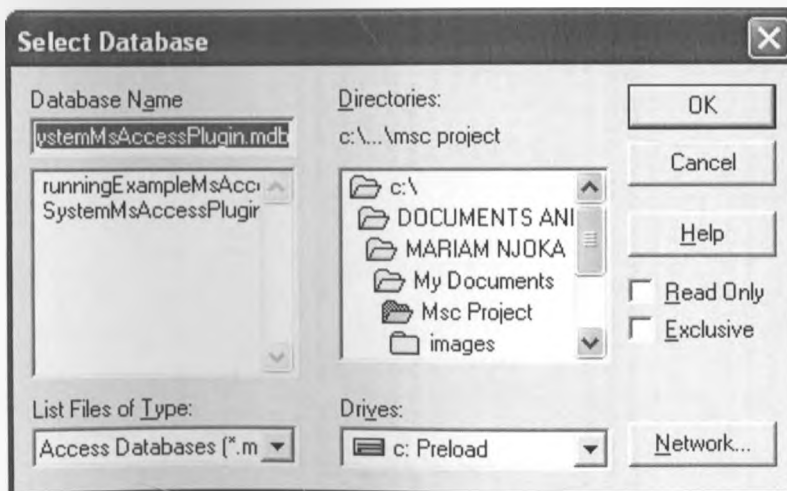
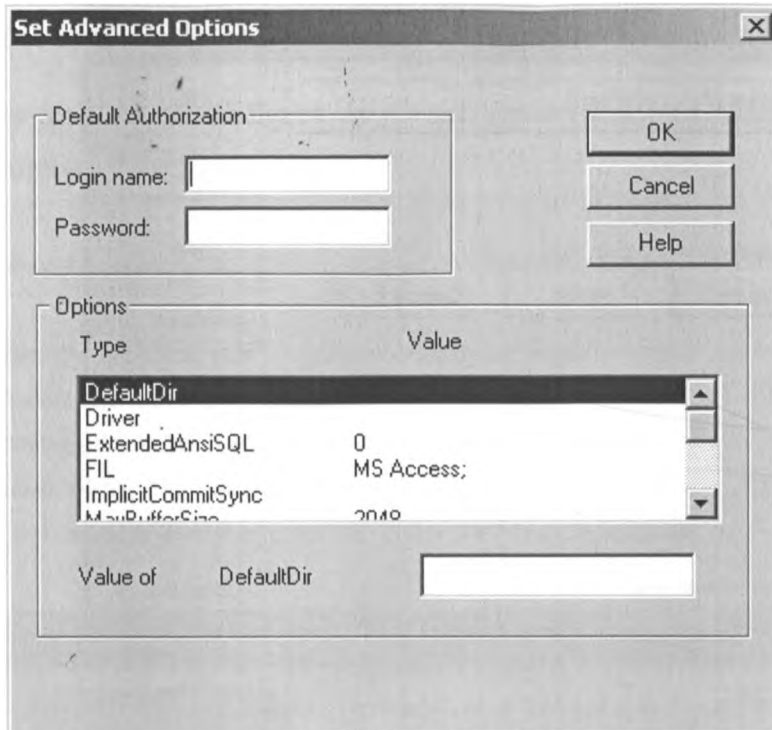


Figure 22: Screenshot of the window to set up the database location



**Figure 23:** Screenshot of the window to set up the "username" and "password" provided to MS Access database plugin

## Appendix 2: Log summary of the event log

The log has been truncated in some parts to avoid it being too long.

### Log Summary

Number of processes: 1

Total number of process instances: 1975

Total number of audit trail entries: 6076

Name: logs\_Process\_instancesGLOBAL.mxml.gz

Description: This log is converted from the tables 'Process instances and Audit\_Trail\_Entries and Data\_Attributes\_Process\_Instances and Data\_Attributes\_Audit\_Trail\_Entries' at the database 'jdbc:odbc:SystemMsAccessPlugin'

Attribute name	Value
os.version	5.1
os.arch	x86
user.name	Mariam Njoka
mxml.creator	MXMLib ( <a href="http://promimport.sf.net/">http://promimport.sf.net/</a> )
java.version	1.6.0_17
mxml.version	1.1
java.vendor	Sun Microsystems Inc.
os.name	Windows XP
app.name	ProM Import Framework
app.version	7.0 (Propeller)

### Source

Name: MsAccessDB

Description:

Attribute name	Value
program	MsAccessDB

## Process Instances

Number of process instances entries: 1975

Process Instance	Occurrences (absolute)	Occurrences (relative)
100, 1000, 1001, 1002, 1003, 1004, 1005, 1006, 1007, 1008, 1009, 101, 1010, 1011, 1012, 1013, 1014, 1015, 1016, 1017, 1018, 1019, 102, 1020, 1021, 1022, 1023, 1024, 1025, 1026, 1027, 1028, 1029, 103, 1030, 1031, 1032, 1033, 1034, 1035, 1036, 1037, 1038, 1039, 104, 1040, 1041, 1042, 1043, 1044, 1045, 1046, 1047, 1048, 1049, 105, 1050, 1051, 1052, 1053, 1054, 1055, 1056, 1057, 1058, 1059, 106, 1060, 1061, 1062, 1063, 1064, 2044, 2045, 2046, 2047, 2048, 2049, 205, 2050, 2051, 2052, 2053, 2054, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304	1	0.051%

## Log events

Number of audit trail entries: 5

Model element	Event type	Occurrences (absolute)	Occurrences (relative)
Submitted for Endorsement	start	2132	35.089%
Endorsed	unknown	1868	30.744%
Approved	completed	1770	29.131%
Declined	unknown	298	4.905%
Submitted for Itinerary Advice	start	8	0.132%



## Starting Log Events

Number of audit trail entries: 2

Model element	Event type	Occurrences (absolute)	Occurrences (relative)
Submitted for Endorsement	start	1970	99.747%
Submitted for Itinerary Advice	start	5	0.253%

## Ending Log Events

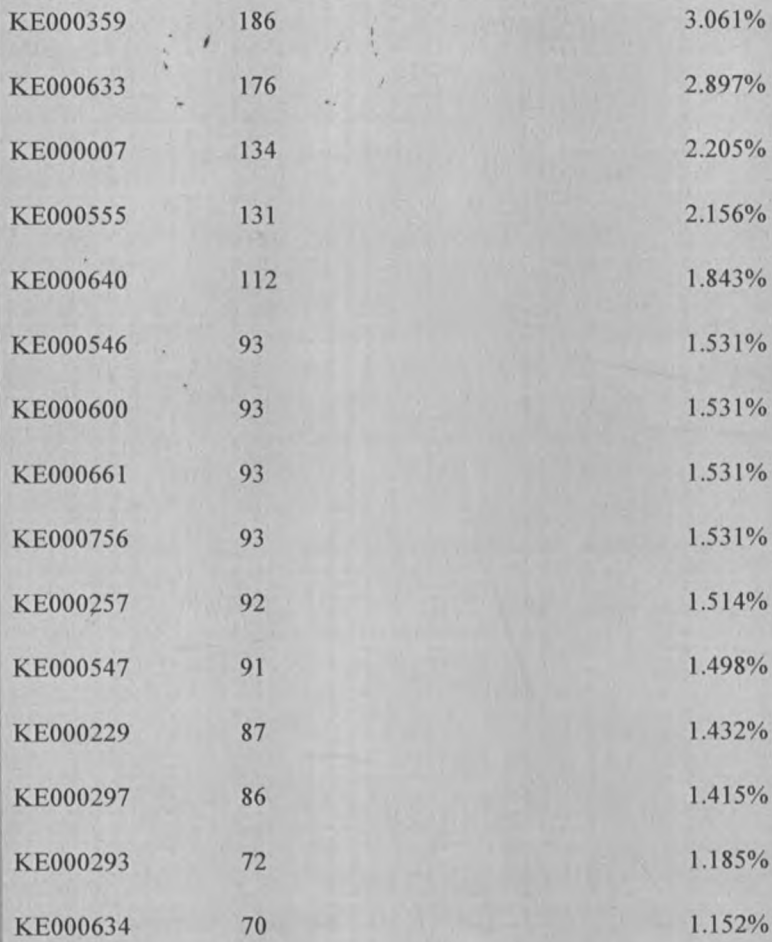
Number of audit trail entries: 5

Model element	Event type	Occurrences (absolute)	Occurrences (relative)
Approved	completed	1770	89.62%
Submitted for Endorsement	start	135	6.835%
Declined	unknown	59	2.987%
Endorsed	unknown	10	0.506%
Submitted for Itinerary Advice	start	1	0.051%

## Originators

Number of originators: 77

Originator	Occurrences (absolute)	Occurrences (relative)
KE000072	1275	20.984%
KE000258	847	13.94%
KE000485	344	5.662%
KE000204	262	4.312%
KE000286	252	4.147%
KE000093	212	3.489%



KE000359	186	3.061%
KE000633	176	2.897%
KE000007	134	2.205%
KE000555	131	2.156%
KE000640	112	1.843%
KE000546	93	1.531%
KE000600	93	1.531%
KE000661	93	1.531%
KE000756	93	1.531%
KE000257	92	1.514%
KE000547	91	1.498%
KE000229	87	1.432%
KE000297	86	1.415%
KE000293	72	1.185%
KE000634	70	1.152%

**Figure 24: Log Summary**

Final Report



Masters of Science in Computer Science  
University of Nairobi  
School of computing and informatics

# **Framework for Process Mining in Semi-structured Information Systems**

By

Mariam Mohammed Njoka

P58/9011/2006

08 Apr 2011

Supervisor: Andrew Mwaura

Submitted in Partial Fulfilment of the Requirements for the Degree of Masters of  
Science in Computer Science

University of NAIROBI Library



0478781 8

Ed 364683

Cho  
Afr  
TN  
275  
N56

## Declaration

This research project entitled “**Framework for Process Mining in Semi-structured Information Systems**” is my original work and has not submitted to any other university.

Sign:  ..... Date: 26/08/2011 .....

Mariam Mohammed Njoka  
P58/9011/2006

This research project entitled “**Framework for Process Mining in Semi-structured Information Systems**” has been submitted for examination with approval as the university supervisor.

Sign:  ..... Date: 26/8/2011 .....

  
Andrew Mwaura  
Supervisor

# Abstract

There are many application systems in the developing world that do not adhere to the systematic software development procedures and are not process-aware. Many of these systems do not meet their objective of efficiency & effectiveness. However, these applications can be analyzed and improved using information derived from event logs.

We discuss what process mining is, how to extract & prepare event logs from systems and the different tools that are used to process the data and give meaningful information, how to analyse the results and eventually introduce a framework that can be used to go through process mining.

We begin by following a framework developed by R.S. Mans, M.H. Schonenberg, M. Song, W.M.P. van der Aalst, and P.J.M. Bakker. However, the special circumstances of semi-structured systems are looked into, leading to changes in the framework. The proposed framework considers three main analysis methods to be performed on the event logs; *control flow analysis*, *performance analysis*, and *role analysis*. Data collection and preparation are given prominence due to the nature of discovery that is done on semi-structured systems. These systems in many cases will not have had previous statistical information of this crucial information. The process mining plugins help in getting this information. This research can be used to assist many organizations in cleaning up their systems and coming up with optimal processes which meet their objectives.

## Acknowledgements

I extend my heartfelt gratitude to my family especially my late mother Shakilla Juma, who continuously without fail impressed upon me on the importance of education, my son Ahmed Yussuf Wanjugi, who endured a busy mum with work and school and was very patient, he is the best son ever. I thank my sisters Mwini, Fatma and Umami and my brothers Ramadhan and Abdulhalim for their constant encouragement. Last but not least my father, Mohammed Njoka, who stands proud of his precious children.

In addition, I would like to give my thanks to my friends who encouraged me throughout my studies especially when I was feeling overwhelmed with the amount of work both in school & job.

I would also like to thank my classmates in the Msc Computer Science (May 2007), whose pressure kept me up and encouraged me.

I would wish to acknowledge the tireless guidance that I got from my supervisor, Andrew Mwaura and the contributions of the entire staff of the School of Computing and Informatics, University of Nairobi.

Last but most important, I give my utmost gratitude to Allah, who has brought me so far and enabled me to undertake this project.

# Table of Contents

Declaration .....	2
Abstract .....	3
Acknowledgements .....	4
Table of Contents .....	5
List of Tables.....	6
List of Figures.....	7
List of Abbreviations .....	8
Introduction .....	9
Background.....	9
Project Justification.....	9
Project Objectives .....	10
Project Methodology.....	10
Project Scope .....	10
Thesis statement .....	11
Literature Review .....	12
Process Mining .....	12
The ProM Framework Tool .....	16
MXML.....	18
Data Collection and Preparation .....	22
Data Collection .....	22
Event logs data cleanup .....	25
Conversion of the event logs to MXML .....	26
Mining and Analysis.....	31
The Proposed Framework.....	40
Implications of Research .....	43
Conclusion.....	44
Bibliography .....	45
Appendices .....	46
Appendix 1: How to configure an ODBC connection to a Microsoft Access Database on your computer.....	46
Appendix 2: Log summary of the event log .....	49

# List of Tables,

**Table 1:** Example of a log file

**Table 2:** Identifying cases

**Table 3:** Dashboard Statistics

**Table 4:** Distribution of cases



# List of Figures

**Figure 1:** Process mining framework

**Figure 2:** Structure of an MXML file

**Figure 3:** Data before cleanup

**Figure 4:** Data after cleanup

**Figure 5:** Audit\_Trail\_Entries Table

**Figure 6:** Process mining table ERD

**Figure 7:** ProM Import Tool

**Figure 8:** Opening a Log file

**Figure 9:** Log Dashboard

**Figure 10:** Process Model 1

**Figure 11:** Process Model 2

**Figure 12:** Screenshot of the window to create a new data source

**Figure 13:** Compliant case following the most frequent path

**Figure 14:** Compliant case with the path covered by some of the cases

**Figure 15:** Non-compliant case with 25 events, depicting abnormality

**Figure 16:** Routing probabilities

**Figure 17:** LTL checker

**Figure 18:** Performance on different patterns

**Figure 19:** The proposed framework

**Figure 20:** Screenshot of the window to create a new data source

**Figure 21:** Screenshot of the window to set up the name of the database that has to be provided after jdbc:odbc: in the field DbHostUrl in the MS Access database plugin

**Figure 22:** Screenshot of the window to set up the database location

**Figure 23:** Screenshot of the window to set up the "username" and "password" provided to MS Access database plugin

**Figure 24:** Log Summary

# List of Abbreviations

**MXML – M eXtensible Markup Language**

**XML – eXtensible Markup Language**

**BI – Business Intelligence**

**BPR – Business Process Reengineering**

**WFM – Workflow Management**

**EAI – Enterprise Application Integration**

**ERP – Enterprise Resource Planning**

**WS – Web Services**

# List of Figures

**Figure 1:** Process mining framework

**Figure 2:** Structure of an MXML file

**Figure 3:** Data before cleanup

**Figure 4:** Data after cleanup

**Figure 5:** Audit\_Trail\_Entries Table

**Figure 6:** Process mining table ERD

**Figure 7:** ProM Import Tool

**Figure 8:** Opening a Log file

**Figure 9:** Log Dashboard

**Figure 10:** Process Model 1

**Figure 11:** Process Model 2

**Figure 12:** Screenshot of the window to create a new data source

**Figure 13:** Compliant case following the most frequent path

**Figure 14:** Compliant case with the path covered by some of the cases

**Figure 15:** Non-compliant case with 25 events, depicting abnormality

**Figure 16:** Routing probabilities

**Figure 17:** LTL checker

**Figure 18:** Performance on different patterns

**Figure 19:** The proposed framework

**Figure 20:** Screenshot of the window to create a new data source

**Figure 21:** Screenshot of the window to set up the name of the database that has to be provided after jdbc:odbc: in the field DbHostUrl in the MS Access database plugin

**Figure 22:** Screenshot of the window to set up the database location

**Figure 23:** Screenshot of the window to set up the "username" and "password" provided to MS Access database plugin

**Figure 24:** Log Summary

# List of Abbreviations

**MXML** – M eXtensible Markup Language

**XML** – eXtensible Markup Language

**BI** – Business Intelligence

**BPR** – Business Process Reengineering

**WFM** – Workflow Management

**EAI** – Enterprise Application Integration

**ERP** – Enterprise Resource Planning

**WS** – Web Services

# Introduction

## Background

There are very many application systems in the developing world that do not adhere to the systematic software development procedures and are not process aware. These systems aim to achieve efficiency within organizations. The fact is many do not meet this objective due to the fact that the developers do not come up with process models before development and they lack a measure of the processes from start to end to know if they are actually efficient enough. Human activities are also not taken into account when calculating how much time will be saved by using a particular system versus what existed before.

However, this does not mean that all is lost; these applications can be analyzed with a view of improving them. This will also go a long way in uncovering and measuring the discrepancies between prescriptive process models and actual process executions.

Process mining involves extracting information from event logs to capture the business process as it is being executed. Process mining aims at improving this by providing techniques and tools for discovering process, control, data, organizational, and social structures from event logs. The framework for process mining in semi-structured systems will be used more frequently in the third world given that many application systems do not take into account how efficient the processes are. Tools that can assist in analyzing these processes will go a long way in assisting organizations do their business process reengineering [BPR] and business intelligence [BI] more effectively.

Technologies such as workflow management (WFM), enterprise application integration (EAI), enterprise resource planning (ERP), and web services (WS) typically focus on the realization of IT support rather than monitoring the operational business processes. These systems have been created based on process models; however it is common that there are certain events or conditions that may not have been anticipated therefore there is no way they were captured in the model.

The tool that we chose for this research is the ProM Framework Tool.

## Project Justification

Many application systems in the developing countries, do not adhere to the systematic software development procedures and are not process aware. These systems aim to achieve efficiency within organizations. The fact is many do not meet this objective due to the fact that the developers do not measure the processes from start to end to know if they are actually efficient enough. Human activities are also not taken into account when calculating how much time will be saved by using a particular system versus what existed before.

However, these applications can be analyzed by following the framework which will assist the person analysing in finally coming up with recommendations which realise the benefits when implemented. This will also go a long way in uncovering and measuring the discrepancies between prescriptive process models and actual process executions.

## Project Objectives

In Kenya as in many third world countries there are so many organizations especially the government that are semi-computerized or in the process of getting computerized. Looking at the semi manual way of doing things there is a lot in terms of resources both time & labor that can be saved by merely computerizing, however, the question remains whether the computerized processes would save these resources optimally. Business process re-engineering has come in to enforce maximum efficiency in processes but there is a gap after implementation.

This gap is caused by the unforeseen events that happen as the process is actually executed. We will show how these organizations can analyze and evaluate their processes through actual study of the event logs after implementation and eventually streamline them. The main objectives of this research are;

1. To review literature on existing techniques for process mining
2. To put the topic of process mining into context, discuss the main issues around process mining, the different tools and methodologies of process mining and their application in semi structured systems in Kenya
3. Demonstrate using a process mining tool & event logs from a live semi-structured system, the data collection, data preparation, & finally analysis of processes from the system.
4. To propose a framework for process mining in semi-structured systems

## Project Methodology

The methodology used is as follows;

1. Literature review
2. Data collection & cleanup of event logs from a live application system
3. Analysis of the processes using a process mining tool
4. Creation of a process mining framework which shows the procedures followed and main entities to be taken into account.

## Project Scope

We cover the context of process mining, it discusses in detail how it is achieved using demonstrations from a system and a process mining tool called ProM framework, then we discuss the most challenging problems and propose a framework for process mining for a semi structured application system.

## Thesis statement

We present a framework for process mining of semi-structured systems from data collection, cleaning, conversion, mining & analysis of the event logs using the Prom Framework Tool among other tools. It is meant to guide others in going through the process as there are huge numbers of similar systems in use currently.

# Literature Review

## Process Mining

Process mining starts by gathering information about processes as they take place as W.M.P. van der Aalst and A.J.M.M. Weijters rightfully say in their research “Process mining – A Research Agenda”. Any transactional system would give this information in one form or another. Many information systems have a timestamp on events as they happen and usually these events are the first source of information for this process. The basic idea of process mining is to extract knowledge from event logs recorded by an information system. Example of a log is on Table 1. It contains five cases, each case one or more tasks depending on how far along the process the case is.

case 1	task A
case 2	task A
case 3	task A
case 3	task B
case 1	task B
case 1	task C
case 2	task C
case 4	task A
case 2	task B
case 2	task D
case 5	task E
case 4	task C
case 1	task D
case 3	task C
case 3	task D
case 4	task B
case 5	task F
case 4	task D

**Table 1: Example of a log file**

Until recently, the information in these event logs was rarely used to analyze the underlying processes. Process mining aims at improving this by providing techniques and tools for discovering process, control, data, organizational, and social structures from event logs. Fuelled by the omnipresence of event logs in transactional information systems (cf. WFM, ERP, CRM, SCM, and B2B systems), process mining has become a vivid research area. The process mining group at Eindhoven University of Technology in Netherlands have been specifically involved in process mining research and have many publications and tools that they have come up with.



Many vendors are now pushing technologies such as Business Process Analysis (BPA) and Business Activity Monitoring (BAM). These systems typically aim at basic performance indicators such as cycle time and frequencies. The goal of researchers on process mining is to allow for more advanced concepts where knowledge is extracted from logs and causalities can be discovered.

There are many workflow tools that have come up over time however there is one that is interesting enough because of the kind of information and analysis it provides. The ProM Framework is an extensible framework that supports a wide variety of process mining techniques and many researchers e.g. W.M.P. van der Aalst and H.T. de Beer and B.F. van Dongen in their research “Process Mining and Verification of Properties: An Approach based on Temporal Logic” have used it especially the mining plug-in, which implements a mining algorithm that constructs a Petri net based on an event log, the export plug-in, which implement a “save as” functionality for objects e.g. graphs etc.

Process mining should not be confused with Business Intelligence (BI). Business intelligence is a broad category of applications and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions. BI applications include the activities of decision support systems, query and reporting, online analytical processing (OLAP), statistical analysis, forecasting, and data mining. While process mining digs deeper at the event logs that are produced by enterprise systems in order to come up with meaningful information on process performance, audit etc.

BI is a term that was used as early as September, 1996, when a Gartner Group report said: By 2000, Information Democracy will emerge in forward-thinking enterprises, with Business Intelligence information and applications available broadly to employees, consultants, customers, suppliers, and the public. The key to thriving in a competitive marketplace is staying ahead of the competition. Making sound business decisions based on accurate and current information takes more than intuition. Data analysis, reporting, and query tools can help business users wade through a sea of data to synthesize valuable information from it - today these tools collectively fall into a category called “Business Intelligence”.

Today’s Business Intelligence (BI) tools used in many industries, like Cognos, Business Objects, or SAP BI, typically look at aggregate data seen from an external perspective (frequencies, averages, utilization, service levels, etc.), as R.S. Mans, M.H. Schonenberg, M. Song, W.M.P. van der Aalst, and P.J.M. Bakker indicate in their research “Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital”. These BI tools focus on performance indicators e.g. the number of knee operations, the length of waiting lists, and the success rate of surgery. Process mining looks “inside the process” at different abstraction levels. So, in the context of a hospital, unlike BI tools, the concern is mainly with the care paths followed by individual patients and whether certain procedures are followed or not. Process mining would also focus on process performance e.g. time taken in each

activity, what process paths were frequently followed, what can be improved from this to make the process easier on patients.

This research topic has recently gained a lot of interest as information systems come of age and computerization is not seen as the only answer, more and more benefits are being sought out of these systems.

An interesting class of information systems that produce event logs are the so-called *Process-Aware Information Systems* (PAISs). Examples are ERP systems (e.g. SAP), case handling systems (e.g. FLOWer), CRM systems (e.g. Microsoft Dynamics CRM), middleware (e.g., IBM's WebSphere), hospital information systems (e.g., Chipsoft), etc. These systems provide very detailed information about the activities that have been executed.

Process mining addresses the problem that most “process/system owners” have limited information about what is actually happening. In practice, there is often a significant gap between what is prescribed or supposed to happen, and what *actually* happens. Only a concise assessment of reality, which process mining strives to deliver, can help in verifying process models, and ultimately be used in system or process redesign efforts.

*The idea of process mining is to discover, monitor and improve real processes (i.e., not assumed processes) by extracting knowledge from event logs. We consider three basic types of process mining; discovery, conformance, and extension.*

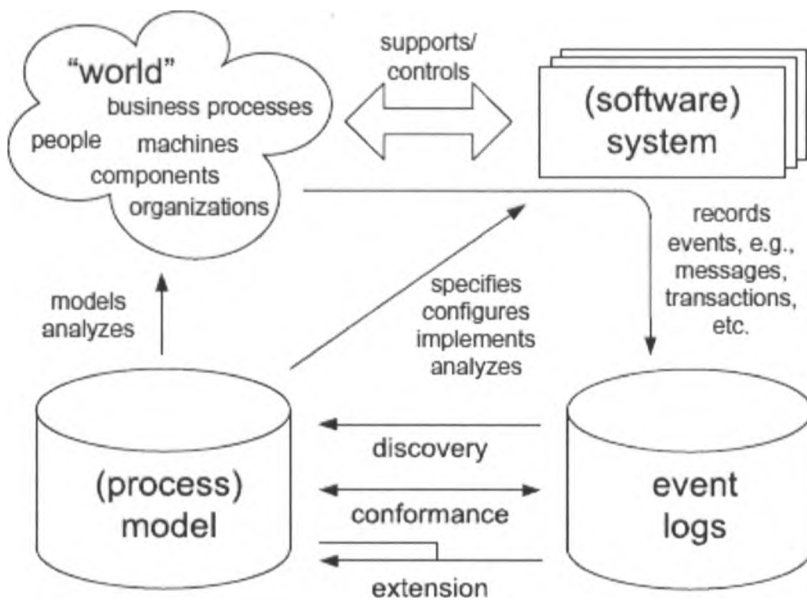
### **Framework reviewed**

The framework to be developed in this study will be based on the framework developed by R.S. Mans, M.H. Schonenberg, M. Song, W.M.P. van der Aalst, and P.J.M. Bakker in their research “Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital”. They considered three basic types of process mining: *discovery*, *conformance*, and *extension*.

**Discovery:** Traditionally, process mining has been focusing on *discovery*, i.e., deriving information about the original process model, the organizational context, and execution properties from enactment logs. An example of a technique addressing the control flow perspective constructs a Petri net model describing the behaviour observed in the event log. It is important to mention that there is no a-priori model, i.e., based on an event log some model is constructed. However, process mining is not limited to process models (i.e., control flow) and recent process mining techniques are more and more focusing on other perspectives, e.g., the organizational perspective, performance perspective or the data perspective. For example, there are approaches to extract social networks from event logs and analyze them using social network analysis. This allows organizations to monitor how people, groups, or software/system components are working together. Also, there are approaches to visualize performance related information, e.g. there are approaches, which graphically shows the bottlenecks and all kinds of performance indicators, e.g., average/variance of the total flow time or the time spent between two activities.

**Conformance:** This is an a-priori model. This model is used to check if reality conforms to the model. Conformance checking may be used to detect deviations, to locate and explain these deviations, and to measure the severity of these deviations.

**Extension:** This is an a-priori model. This model is extended with a new aspect or perspective, i.e., the goal is not to check conformance but to enrich the model with the data in the event log. An example is the extension of a process model with performance data, i.e., some a-priori process model is used on which bottlenecks are projected. At this point in time with mature tools such as the ProM framework, featuring an extensive set of analysis techniques which can be applied to real-life logs while supporting the whole spectrum depicted in the figure below..



**Figure 1: Process mining framework**

### Critique of the framework

R.S. Mans, M.H. Schonenberg, M. Song, W.M.P. van der Aalst, and P.J.M. Bakker' model, captures the process and the relationships from the business requirements and reality to the system software which does the actual logging as processes are done through it, to the event logs which are then mined and finally to the process model, which again feeds into the world realities and business requirements.

However, in many semi-structured systems there is most often than not a lack of process model, so it is expected that once the process mining process is through then a process model will be produced and fed into the system.

Another handicap with semi-structured systems is the data cleanliness, this is a huge aspect as very often there will be data which does not conform perhaps it was testing data or otherwise. Many systems are debugged and correctly on live environments. Thus cleanup and preparation of data should be considered as an important part of the process.

Finally, the best way to analyse is to look at the logs with a view of getting the control flow analysis, performance analysis and role analysis. With this kind of information, then all manner of decisions touching on how to control the work flow, how to improve the organizational/role allocation and how to improve the performance of the processes themselves can be done.

One of the tools to support process mining is the process mining framework ProM. It is plugin-based to support new areas and techniques. In the last decade, process mining evolved from control flow discovery to a broad area of research to get all kinds of information from a log, which resulted in more than 250 different plugins within ProM. This shows that many different techniques exist to apply process mining. Since there are so many, it is not clear anymore when to use which plugin. Although many case studies, have been performed, the main problem with these case studies is that they were all done case-by-case on the insights and knowledge of the researcher performing the case study. I.e. it is hard to make process mining a repeatable service as said in a research titled "Process Diagnostics: a Method Based on Process Mining" by Melike Bozkaya, Joost Gabriels & Jan Martijn van der Werf LaQuSo.

## The ProM Framework Tool

**ProM** is an **extensible** framework that supports a wide variety of process mining techniques in the form of plug-ins. It is **platform independent** as it is implemented in Java. **ProM** framework is issued under an **open source** license, namely the Common Public License (CPL), and researchers and developers are invited to **contribute** in the form of new plug-ins. It was developed by the Process Mining Group, Eindhoven Technical University.

ProM is a generic framework for implementing process mining tools in a standard environment. The ProM framework receives as input logs in the Mining XML (MXML) format. Currently, this framework has plug-ins for process mining, analysis, monitoring and conversion.

ProM is available as binary distribution files for the Windows, Mac OS X and UNIX platforms, and as source code under the terms of the CPL license. It requires a present installation of the Java Runtime Environment, version 1.5/5.0 or greater (Version 5.0 is recommended for Windows, Linux, and Mac OS X).

The following are the different types of plugins that are found in ProM;

**Mining plug-ins**, such as:

1. Plugins supporting control-flow mining techniques (such as the Alpha algorithm, Genetic mining, Multi-phase mining, etc.)

2. Plugins analysing the organizational perspective (such as the Social Network miner, the Staff Assignment miner, etc.)
3. Plugins dealing with the data perspective (such as the Decision miner, etc.)
4. Plugins for mining less-structured, flexible processes (such as the Fuzzy Miner)
5. Elaborate data visualization plugins (such as the Cloud Chamber Miner)
6. (and many more)

**Analysis plug-ins** dealing with:

1. The verification of process models (e.g., Woflan analysis)
2. Verification of Linear Temporal Logic (LTL) formulas on a log
3. Checking the conformance between a given process model and a log
4. Performance analysis (Basic statistical analysis, and Performance Analysis with a given process model)

**Export plug-ins**, which implement “save as” functionality for some objects (such as graphs). For example, there are plug-ins to save EPCs, Petrinets, spreadsheets, etc.

**Import plug-ins**, which implement an “open” functionality for exported objects, e.g., load instance-EPCs from ARIS PPM.

**Conversion plug-ins**, which implement conversions between different data formats, e.g. EPCs to Petri nets.

Finally, ProM sports a large array of **log filters**, which are a valuable tool for cleaning logs from undesired, or unimportant, artefacts.

There are other tools in the market that are used in the process mining field, examples are below though this are offered commercially;

1. Futura Reflect, a Process Mining and Process Intelligence suite developed by Futura Technology
2. Interstage Automated Process Discovery, a Process Mining service offered by Fujitsu, Ltd. as part of the Interstage Integration Middleware Suite.
3. BPM|one, offering both basic process mining functionality as well as a more comprehensive process mining module as part of the Pallas Athena BPM|one software suite.
4. Nitro is a tool by Fluxicon for easily converting CSV and XLS event logs for ProM.
5. ARIS Process Performance Manage, a Process Mining and Process Intelligence Tool offered by Software AG as part of the Process Intelligence Solution.

The following reasons led me to choose ProM;

1. It is an open source software
2. It has been used extensively for research, with several researchers adding to its functionality.
3. It is rich in functionalities and different mining & analysis plugins.
4. It is widely seen as a leader in process mining tools.

## MXML

**MXML** is an XML-based user interface markup language first introduced by Macromedia in March 2004. Application developers use MXML in combination with ActionScript to develop Rich Internet applications. This is a vendor-independent format to store event logs. One MXML file can store information about multiple processes. Per process, events related to particular process instances (cases) are stored. Each event refers to an activity.

### *Data model*

MXML has three kinds of objects: the documents, the nodes and the iterators. A document has mainly a root node, which holds all the top-level document nodes. Even if only one root -tag- node is allowed in a valid XML file, it is possible to have more non-tag nodes, like comments, processing instructions, and directives. In a document object, MXML also stores eventual errors and the line where an error has been encountered. The document is also used to carry around "formatting style" requests in input or output.

A node is the minimal unit of information; MXML distinguishes among 6 kinds of nodes:

1. Document nodes: they are just "transparent" holder for top-level nodes. The root node in a document is always a document type node.
2. Tag nodes: they are the nodes used to store information in the XML file: an example could be `<item> data </item>`
3. Data nodes: They are the data held by an item. An item can have more than a data node child, as it can, as in this example:

```

<bibliography>
  <book isbn="12345678" publisher = "pb1">
    <author>
      [language = English]
        <firstname> Manolis <firstname>
        <lastname> Gergatsoulis <lastname> [ ]
      [language = Greek]
        <firstname> Μ Ο & <firstname>
        <lastname> " Ο & <lastname> [ ]
    <author>
    <author>
      [language = English]
        <firstname> Panos <firstname>
        <lastname> Rondogiannis <lastname> [ ]
      [language = Greek]
        <firstname> Ο&&<firstname>
        <lastname> Ρ Ο Ο & <lastname> [ ]
    <author>
    <title>
      [language = English]
        Multidimensional Programming Languages [ ]
      [language = Greek]
        Ο " & ! " & Ο Ο [ ]
    <title>
    <price currency="USD">
      [period = discount client = regular]
        100 [ ]
      [period = normal client = regular]
        120 [ ]
      [period in {discount, normal} client = special]
        100 [ ]
    <price>
    <year> 1999 <year>
  <book>
  ...other book elements ...

  <publisher id = "pb1">
    [language = English] NCSR Demokritos [ ]
    [language = Greek] ΕΚΕ Ε Ο ο& [ ]
  <publisher>
  ...other publisher elements ...

</bibliography>

```

4. Comment nodes: they are comments put by the document designers to state something meaningful about what the document (or document part) are meant for. They assume the form of a string inside a block that begins with `<!--` and ends with `-->`: `<!-- This is an XML comment -->`

5. Processing instructions: they are blocks that are passed as they are to 3d party processors; php escape (`<?php ... code ... ?>`) is an example. Also the `<?xml` declaration is treated by XML as a processing instruction; this could change in next releases of MXML.
6. Directives: they are nodes inside a `<! ... >` block. The DOCTYPE and all the DTD declarations, along with the internal and external entity declarations are directives. MXML does not parse them in any way, and handles them untranslated to the application.

Iterators are meant to access easily the document nodes structure, allowing partitioning the tree in subtrees.

### *Nodes in-depth*

Each node has a name, a set of attributes (each of which is a couple of string, the name and the value), a data and the links to the siblings nodes. Some node types can have some of this values not set. In example, both processing instructions and directives has a name (the string immediately following the opening tag) and a data (the content of the tag), but they have not an attribute set. Data and comments notes have only data, while tag nodes have pretty everything. Document nodes are empty, "transparent". To simplify the work of programmers having to scan configuration oriented XML documents, if a tag node has only one data node among its children, its data element will be merged with the data element of that child, and the data node will be removed thus "flattening" the structure. Is it also possible to create a new tag node with its data element set to a string. If, later on, there is the necessity to add a new data node to the item, it does not matter if it has the data element merged: on output, both the data element and the entire child data node will be correctly written. The data element is always written after all the children nodes.

Navigation in the node tree is guaranteed by four node attributes, pointing respectively to the next node in the same level of the hierarchy (and with the same parent), to the previous one, to the first node of the node children, and finally to the parent. So, to traverse the whole tree, one has to start from the root node and then descend recursively in the first child node: then all the "brothers" of that node are scanned up to the last node in the tree.

### *Utilities for node management*

The API is discussed in detail in the documentation; what is necessary to attention, is that some utilities are provided to access/modify the content of the attribute list of the nodes, and to retrieve the depth and, eventually, the path of a node. The depth of a node is its distance from the root node, counted in steps necessary to reach it or his first brother. The path of a node is the list of the name of all its ancestors, plus the node name, separated by a "/" character: `/main/item` is a path indicating that the node named `item` can be found immediately below the node named `"main"`, which is at top-level. Node paths are not unique, moreover, only a node having a name, and whose parents are all named, can have a path.



### Naming conventions

MXML naming conventions are rigid ways in which each function, structure and data is called. Each function in the library begins with "mxml\_", the name of the object that the function refers to and the operation on that object. In example, the function to create a new node is "mxml\_node\_new()". Mxml function namings are not necessarily restricted to structures, but could also refer to abstract objects, like "mxml\_path\_\*" (which operates on strings representing node paths), or "mxml\_attribute\_\*" (which handle the attribute lists for nodes). Every symbol defined in .h file begins with "MXML\_". So, a document is typed as an MXML\_DOCUMENT, and a node is a structure named MXML\_NODE.

### Structure of an MXML file

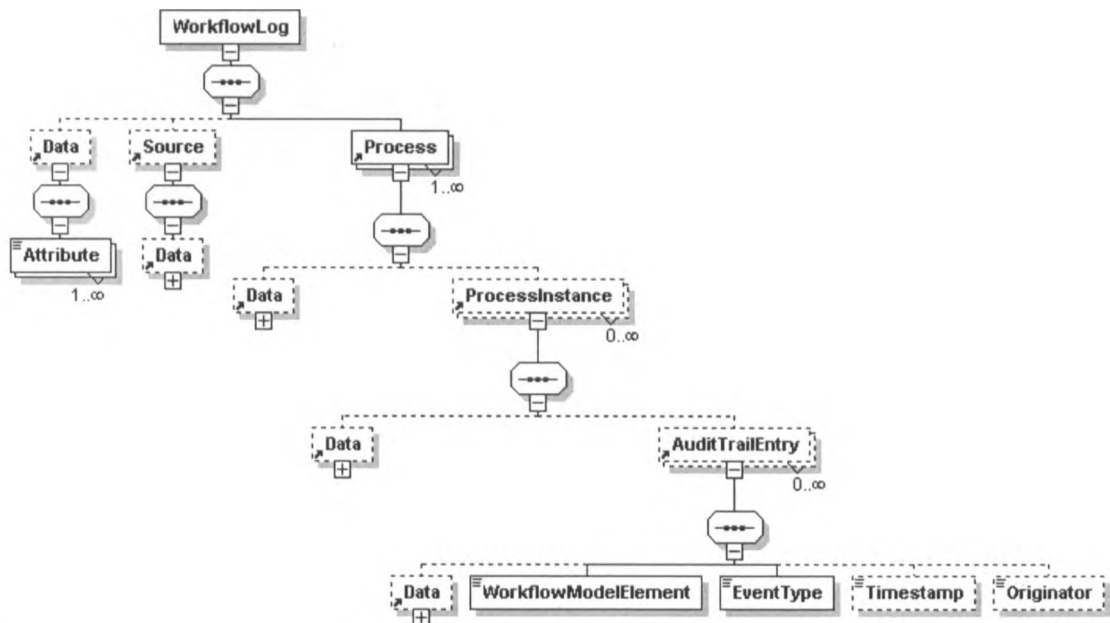


Figure 2: Structure of an MXML file

# Data Collection and Preparation

As the research is data-heavy a lot of emphasis was placed on data collection & preparation for its analysis.

Below is the detailed description of the procedures and techniques that were used in the data collection, conversion and eventual analysis. As the main objective of the paper is to come up with a framework the procedures below should be given extra attention as they give the information required to carry out this exercise given any system.

## Data Collection

The first thing is to understand the makeup of the system's database and especially the files or tables that contain the workflow information & audit trails.

The system currently under study has a web-based front-end, and a back-end of Microsoft SQL database. It is a procurement system currently in use in a professional services firm. It has several modules; however the one in use currently is the LPO approval process. Therefore the process that can be derived from it is as follows,

1. Raising the LPO,
2. Submitting the LPO for endorsement to the first partner
3. Endorsement of the LPO
4. Request for approval to the second partner
5. Approval of the LPO

The events that signify that each task has completed are as follows;

1. Submitted for Endorsement
2. Endorsed
3. Approved

With "Submitted for Endorsement" is the event that signifies the start of the process and "Approved" signifies the end of the process. After the approval the LPO is printed and given to the supplier.

The system contains several workflow tables, however after thoroughly understanding what each table does we picked up on the following three tables which were of interest to the research;

1. workflow\_history
2. workflow\_document
3. workflow\_definition

*work\_history* – this is the main audit trail table, it contains the tasks, timestamp & originator id

Its definition is as follows;

```
[workflow_history_id] [int] IDENTITY(1,1) NOT NULL, PK
[document_id] [int] NOT NULL, FK
[workflow_definition_id] [int] NULL, FK
[workflow_date] [datetime] NULL,
[current_status_description] [varchar](50) NULL,
[workflow_action] [varchar](50) NULL,
[staff_id] [varchar](50) NULL,
[comments] [varchar](1000) NULL
```

It has all the required parameters for analysis except for the eventtype(start,complete) which we had to add for purposes of the process instance beginning & end being recognized automatically. The other parameters are; process instance id, audit trail id, timestamp, originator, instance action.

*workflow\_documents* – This is the documents table, when a new document is raised, its record is stored here e.g when it was created, the person who initiated it etc. when it undergoes tasks such as requests for approvals, endorsement & approvals, this information is saved in the *workflow\_history* table but the main connecting factor is the *document\_id*. This table was used as basis for the cases (process instances) because of this tying factor.

Its definition is as follows;

```
[workflow_document_id] [int] IDENTITY(1,1) NOT NULL, PK
[item_category_id] [int] NULL, FK
[workflow_document_source_id] [int] NULL, FK
[workflow_job_code] [varchar](100) NULL,
[workflow_document_initiator] [varchar](255) NULL,
[workflow_document_date_created] [datetime] NULL CONSTRAINT [DF_WorkFlow_Documents_date_created]
DEFAULT (getdate()),
[workflow_document_date_due] [datetime] NULL,
[workflow_document_current_state_id] [int] NULL,
[workflow_document_priority] [int] NULL,
[workflow_job_codew] [varchar](100) NULL
```

E.g. In the image below the *document\_id* field is the one that contains the unique id for the audit trail record; the *document\_id* is similar for all the tasks executed for that document.

1	workflow history id	document id	workflow definition id	workflow date	workflow action	staff id
2	564	78	41	3/8/2010 7:19	Declined	KE000359
3	563	78	42	3/8/2010 7:18	Submitted for Endorsement	KE000359
4	565	78	42	3/8/2010 7:23	Submitted for Endorsement	KE000359
5	566	78	41	3/8/2010 7:23	Endorsed	KE000359
6	616	78	44	3/8/2010 10:56	Approved	KE000229
7	567	79	42	3/8/2010 7:39	Submitted for Endorsement	KE000359

**Table 2: Identifying cases**

Therefore this is the table to be used to get the process instances. The parameters required for analysis from this table were only two; document\_id and description. However, since there was no description available from the table thus we used the document\_id as description.

*workflow\_definitions* – this table contains the definitions of the processes

Its definition is as follows;

```
[workflow_definition_id] [int] IDENTITY(1,1) NOT NULL, PK
[workflow_id] [int] NULL, FK
[item_category_id] [int] NOT NULL, FK
[state_name] [varchar](50) NOT NULL,
[state_order] [int] NOT NULL CONSTRAINT [DF_Item_Category_WorkFlow_state_order] DEFAULT ((0)),
[state_enabled] [int] NOT NULL CONSTRAINT [DF_Item_Category_WorkFlow_state_enabled] DEFAULT ((1)),
[state_owner] [varchar](50) NULL,
[state_is_final_approval] [int] NOT NULL CONSTRAINT
[DF_Item_Category_WorkFlow_state_is_final_approval] DEFAULT ((0)),
[go_back_state_id] [int] NULL,
[go_next_state_id] [int] NULL,
[go_back_advice_description] [varchar](50) NULL,
[go_next_advice_description] [varchar](50) NULL,
[current_status_description] [varchar](50) NOT NULL,
[go_back_document_result_status] [varchar](50) NULL,
[go_next_document_result_status] [varchar](50) NULL,
```

E.g. when an approval is declined it is indicated within the same process definition, a new definition is not created for that event, therefore the field relating to declines and deletions has to be added for that purpose.

After analysis of the database and how the tables relate to each other we went to data cleanup. We exported the data to Ms Excel for the cleanup.

## Event logs data cleanup

After data collection, the data has to be cleaned up. Before applying any mining technique to an event log, you may want to remove unnecessary information from the log before you start the mining. For instance, you may be interested in mining only information about the cases that are completed or conform. For the system we are analysing some data had to go as it did not conform to the rest, we also had to make sure that the timestamp was in the correct format. We used Excel to do most of the cleanup as we could easily isolate what we did not want. The cleaning step is usually a projection of the data to consider only the data you are interested in.

The next step was to remove the columns that were unnecessary for mining ensuring that the data does not lose integrity. Below is an image of the workflow\_history data before and after the cleanup.

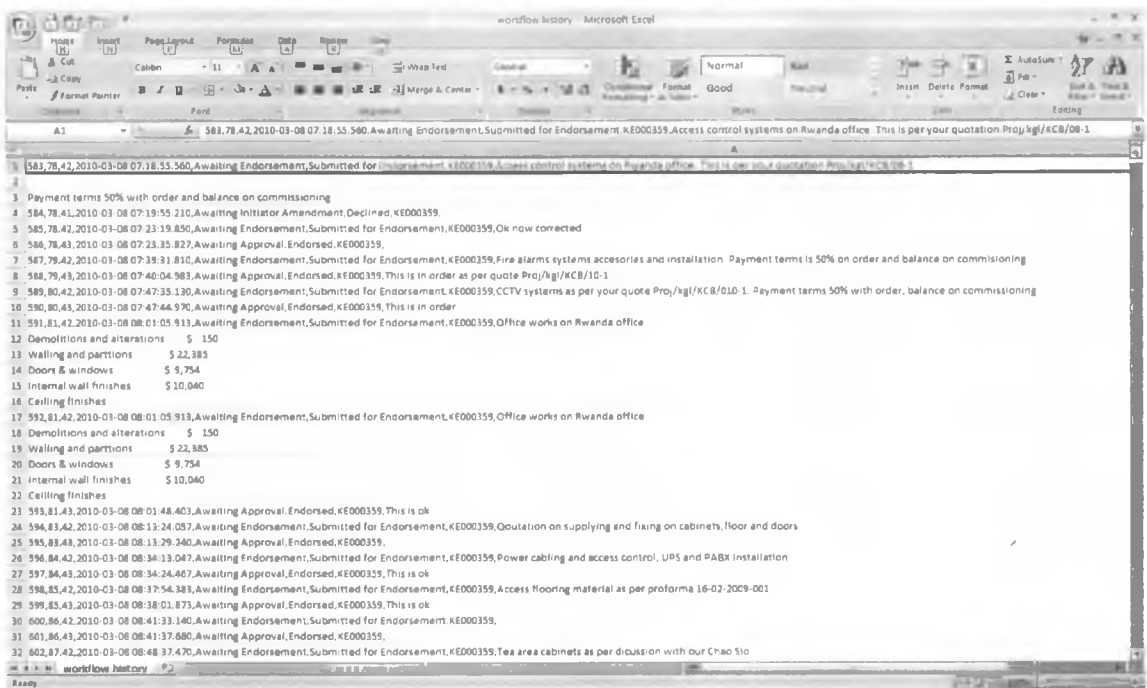
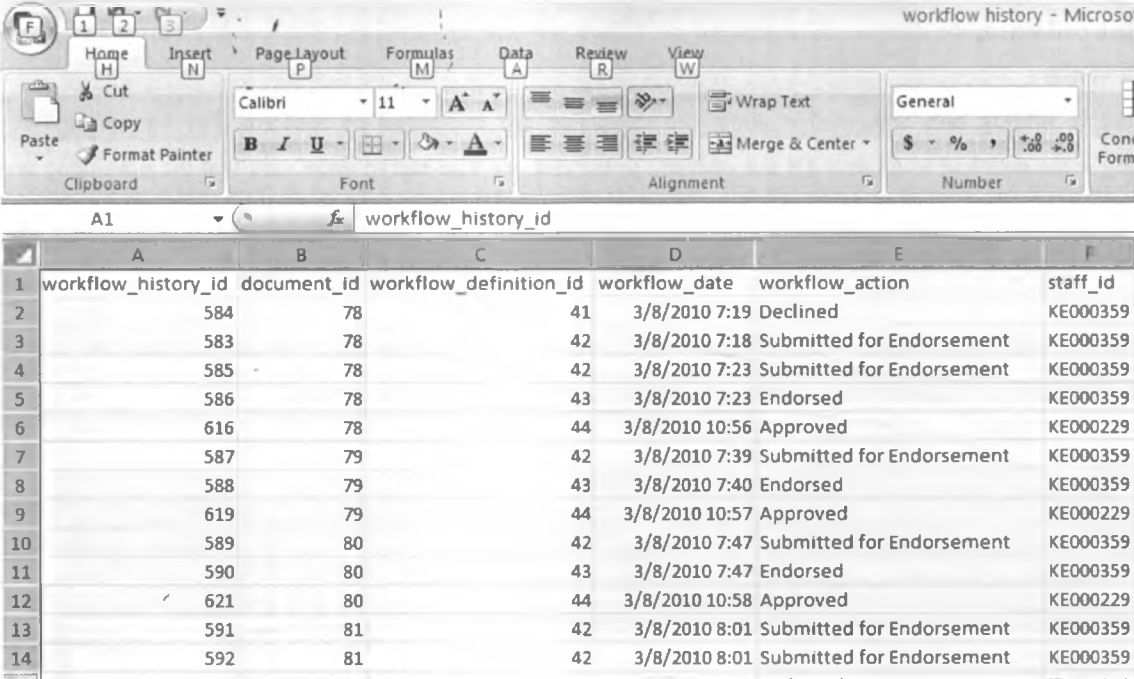


Figure 3: Data before cleanup



	A	B	C	D	E	F
1	workflow_history_id	document_id	workflow_definition_id	workflow_date	workflow_action	staff_id
2		584	78	41	3/8/2010 7:19 Declined	KE000359
3		583	78	42	3/8/2010 7:18 Submitted for Endorsement	KE000359
4		585	78	42	3/8/2010 7:23 Submitted for Endorsement	KE000359
5		586	78	43	3/8/2010 7:23 Endorsed	KE000359
6		616	78	44	3/8/2010 10:56 Approved	KE000229
7		587	79	42	3/8/2010 7:39 Submitted for Endorsement	KE000359
8		588	79	43	3/8/2010 7:40 Endorsed	KE000359
9		619	79	44	3/8/2010 10:57 Approved	KE000229
10		589	80	42	3/8/2010 7:47 Submitted for Endorsement	KE000359
11		590	80	43	3/8/2010 7:47 Endorsed	KE000359
12		621	80	44	3/8/2010 10:58 Approved	KE000229
13		591	81	42	3/8/2010 8:01 Submitted for Endorsement	KE000359
14		592	81	42	3/8/2010 8:01 Submitted for Endorsement	KE000359

Figure 4: Data after cleanup

## Conversion of the event logs to MXML

For data to be uploaded in ProM, it has to be in MXML format. MXML is a **pure C** library that is meant to help developers implementing XML file interpretation in their projects. The compact design is easy to put it in any project, for it is very small, an average program will grow from 15 to 30 kb when it is included.

There were several solution to convert the data, however we chose to use Ms Access database as an in between since it there was an available plugin that could change this data to MXML. With Ms Access the data has to be prepared in a manner that can be easily converted to MXML. Note that MXML is node-based thus the data had to be in a manner that it can be picked as nodes and child nodes within it.

To convert data in a Microsoft Access database about cases and tasks that have been executed to the ProM MXML format, *four* tables have to be defined which have a similar structure to that of the fields in the MXML format. The elements in the MXML format that can contain information about cases and tasks that have been executed are respectively the *Process Instance* element and the *Audit Trail Entry* element. Furthermore, both the *Process Instance* element and the *Audit Trail Entry* element can have *Data* as sub element which can contain respectively additional information about process instances and audit trail entries. An image of the MXML format can be found is above.

Therefore, the first table is *Process\_Instances* which needs to be filled with the identifier of a certain process instance (field *PI-ID*) and, if available, its accompanying description (field *description*). Furthermore, it is important to note that the *PI-ID* field has to be a primary key in the table. The second table *Data\_Attributes\_Process\_Instances* needs to be filled with additional information about each process instance, the so called data attributes. Therefore, this table contains the fields *PI-ID*, *Name* and *Value*. The *PI-ID* field is needed to identify to which process instance each data attribute belongs (actually, this is a foreign key for the *PI-ID* field in table *Process\_Instances*). Also, each process instance can have zero or more data attributes that belong to it. Furthermore, the field *name* represents the name of the data attribute and the field *value* represents the value of the data attribute.

The third table *Audit\_Trail\_Entries* needs to be filled with data about tasks that have been performed during the execution of a process instance. Not surprisingly, this table contains fields with name *WFMElt* (for the name of the task), *EventType* (the task event type, e.g. start, complete), *Timestamp* (the time in which the task changed its state), and *Originator* (the person or system that caused the change in the task state). However, we also have the columns *PI-ID* and *ATE-ID*. The field *ATE-ID*, which is a unique identifier for each audit trail entry (so, this has to be a primary key in the table). The reason for introducing this field is because additional information about each audit trail entry that is relevant could exist, but does not fit in the other fields of table *Audit\_Trail\_Entries*. For this additional information, we have table *Data\_Attributes\_c* which is set up in a similar way as table *Data\_Attributes\_Process\_Instances*. Another reason for introducing the *ATE-ID* field is because it is needed in table *Data\_Attributes\_Audit\_Trail\_Entries* to be able to identify to which audit trail entry each data attribute belongs to.

After creating the above tables, we populated them as follows;

*Process\_Instances (PI-ID, Description)*: we put *document\_id* for both fields as we lacked the description

*Audit\_Trail\_Entries (ATE-ID,PI-ID,WFMElt,EventType,Timestamp,Originator)*: we put the necessary data from the *workflow\_history* table which contained this information.

A populated *audit\_trail\_entries* table in Ms Access

ATE-ID	PI-ID	WFMEIT	EventType	Timestamp	Originator
78		Submitted for Endorsement	start	3/8/2010 7:18:00 AM	KE000359
584	78	Declined		3/8/2010 7:19:00 AM	KE000359
585	78	Submitted for Endorsement	start	3/8/2010 7:23:00 AM	KE000359
586	78	Endorsed		3/8/2010 7:23:00 AM	KE000359
587	79	Submitted for Endorsement	start	3/8/2010 7:39:00 AM	KE000359
588	79	Endorsed		3/8/2010 7:40:00 AM	KE000359
589	80	Submitted for Endorsement	start	3/8/2010 7:47:00 AM	KE000359
590	80	Endorsed		3/8/2010 7:47:00 AM	KE000359
591	81	Submitted for Endorsement	start	3/8/2010 8:01:00 AM	KE000359
592	81	Submitted for Endorsement	start	3/8/2010 8:01:00 AM	KE000359
593	81	Endorsed		3/8/2010 8:01:00 AM	KE000359
594	83	Submitted for Endorsement	start	3/8/2010 8:13:00 AM	KE000359
595	83	Endorsed		3/8/2010 8:13:00 AM	KE000359
596	84	Submitted for Endorsement	start	3/8/2010 8:34:00 AM	KE000359
597	84	Endorsed		3/8/2010 8:34:00 AM	KE000359
598	85	Submitted for Endorsement	start	3/8/2010 8:37:00 AM	KE000359
599	85	Endorsed		3/8/2010 8:38:00 AM	KE000359
600	86	Submitted for Endorsement	start	3/8/2010 8:41:00 AM	KE000359
601	86	Endorsed		3/8/2010 8:41:00 AM	KE000359
602	87	Submitted for Endorsement	start	3/8/2010 8:48:00 AM	KE000359
603	87	Endorsed		3/8/2010 8:48:00 AM	KE000359
604	88	Submitted for Endorsement	start	3/8/2010 8:56:00 AM	KE000359
605	88	Endorsed		3/8/2010 8:56:00 AM	KE000359
606	89	Submitted for Endorsement	start	3/8/2010 8:59:00 AM	KE000359
607	89	Endorsed		3/8/2010 8:59:00 AM	KE000359
608	90	Submitted for Endorsement	start	3/8/2010 9:09:00 AM	KE000359
609	90	Endorsed		3/8/2010 9:09:00 AM	KE000359
610	91	Submitted for Endorsement	start	3/8/2010 9:12:00 AM	KE000359
611	91	Endorsed		3/8/2010 9:12:00 AM	KE000359
612	90	Approved	completed	3/8/2010 10:51:00 AM	KE000229
613	91	Approved	completed	3/8/2010 10:53:00 AM	KE000229

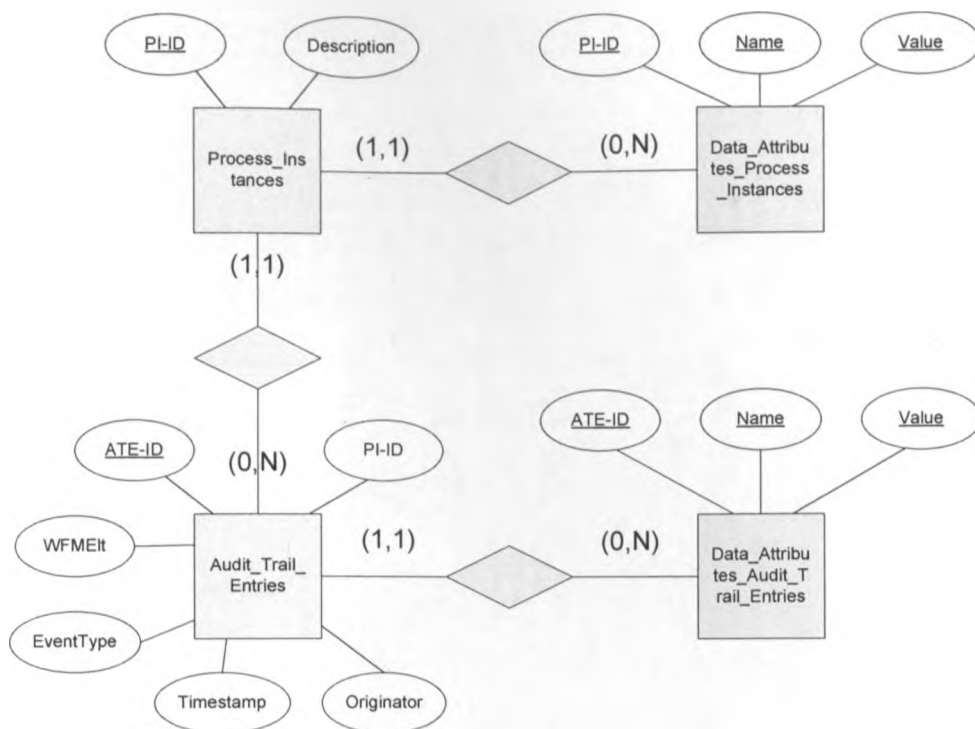
Figure 5: Audit\_Trail\_Entries Table

The next step was to search for tools that can do the conversion to MXML. There exists an open source program known as ProMImport, it is a tool that was developed to convert logs to the acceptable formats that ProM uses. We used it to convert the MsAccess database that we had created to MXML.

For *Data\_Attributes\_Process\_Instances*, *Data\_Attributes\_Audit\_Trail\_Entries* tables we did not require any additional information so we left them without any data. As can be seen from the image below, they do not necessarily have to have records in them. It depends on the system that is being analysed.



Figure 6 is the ERD of the process mining tables;



**Figure 6: Process mining table ERD**

The database is now ready for upload.

The next step was to ensure that an ODBC connection for the access database has to be setup on the computer. See appendix 1.

***The ProM Import Framework***

The ProM Import contains several plugins e.g. Apache2, general CSV file, PeopleSoft, which can be used to convert data of these formats to MXML format, however with Ms Access the plugin we used is the MS Access Database plugin that comes with the latest version 7.0. We entered the information as can be seen from the diagram below then run the program.

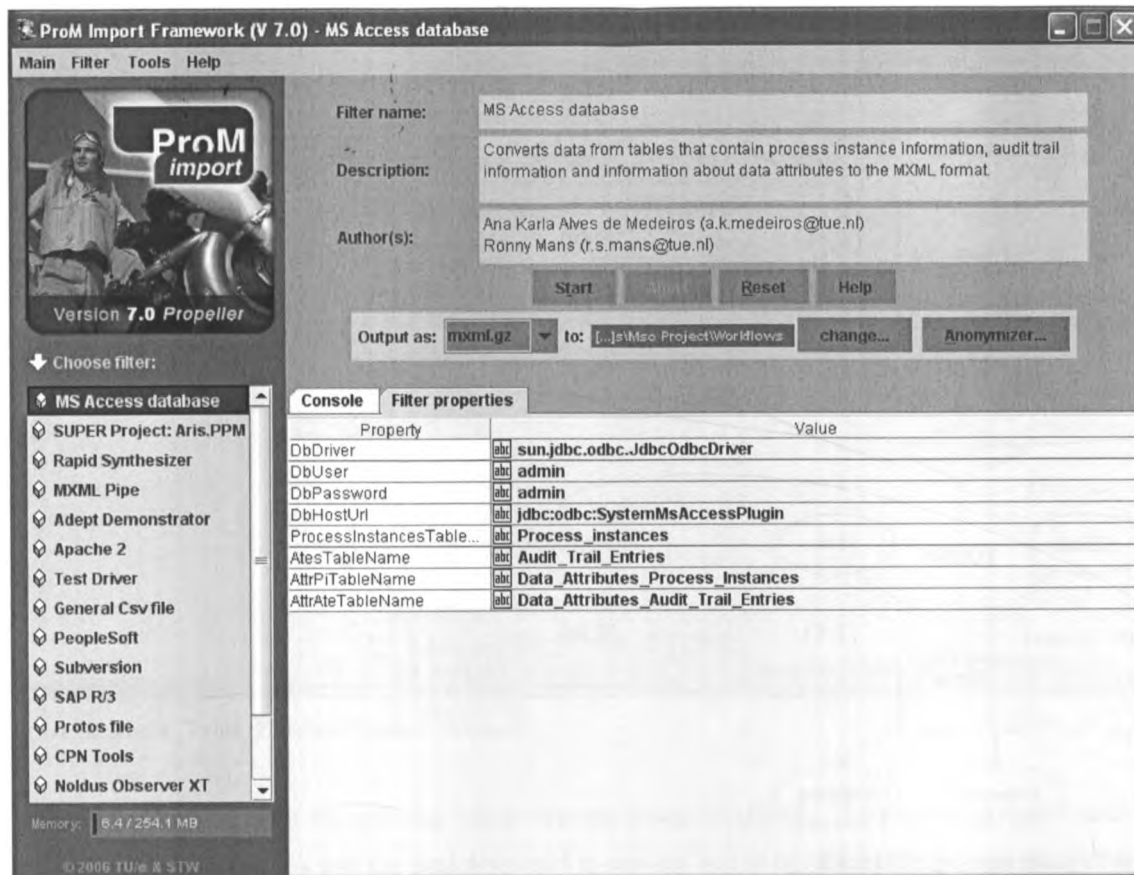


Figure 7: ProM Import Framework (V 7.0)

The MXML file was now ready for analysis.

I used ProM [5.2.] to do the process mining. It is an open source system which many researchers have worked on and is considered to be one of the best in the field. See the image below on loading the MXML log to ProM.

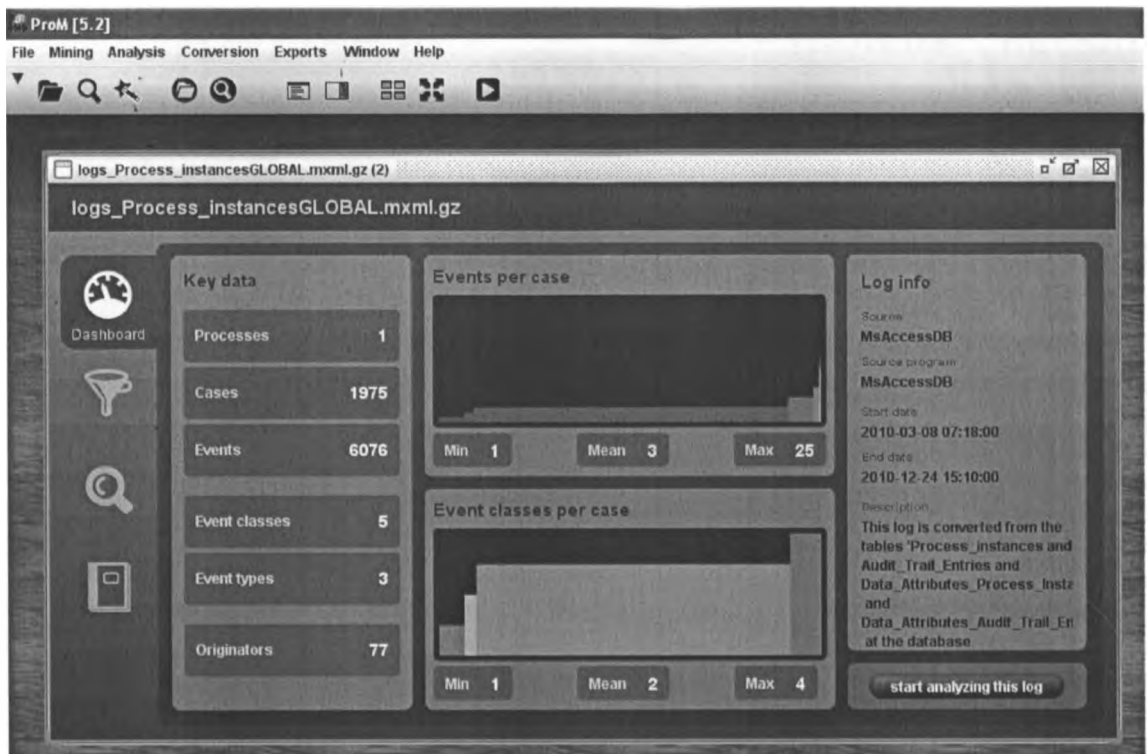
## Mining and Analysis

In this section we present some results obtained from analysis of the e-procurement logs. We concentrate on the information that can be derived using ProM.



Figure 8: Opening a Log file

Immediately after loading ProM does an initial filtering of data that quickly captures the information in the file and arranges it according to processes, cases and events. The MXML file has to have been created from properly prepared data otherwise ProM will read it incorrectly. See the image below for the dashboard that comes up automatically after the initial loading.



**Figure 9: Log Dashboard**

We can quickly deduce the following from the dashboard above;

No. of Processes	1
No. of Cases	1975
No. of Event Types	3
No. of Originators	77
Total no. of events	6076
Minimum no. of events per case	1
Maximum no. of events per case	25
Average no. of events per case	3
No. of Event Classes	5
Minimum no. of event classes	1
Maximum no. of events classes	4
Average no. of event classes	2
Absolute Start date of the log	2010-03-08 07:18
Absolute End date of the log	2010-12-24 15:10

**Table 3: Dashboard Statistics**

More analysis will reveal more information on these events. See appendix 2, which contains the log summary of the event log that we have uploaded into the ProM Framework. It is as a result of a default log filter that identifies the cases, whether they are complete or not and provides summaries of the entry logs.

Given the system I'm mining we can answer the following questions after mining;

1. What is the most frequent path for every process model?

This is the model that is followed most frequently by many cases. Using the Fuzzy Miner plug-in ProM has assisted in quickly coming up with the model and the numbers in each step indicate the frequency of the cases that pass through this route.

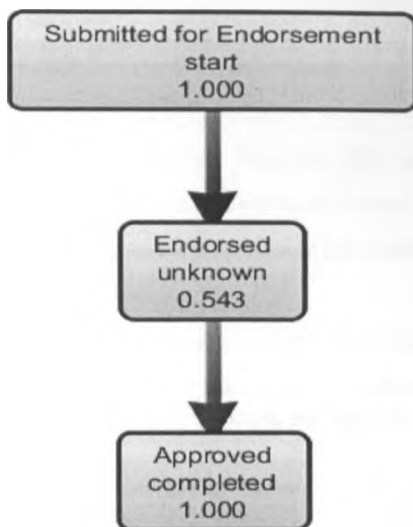


Figure 10: Process Model 1

Other models as mined are as follows; the one below shows many different phases many processes may follow especially if there at any one point a certain request is declined.

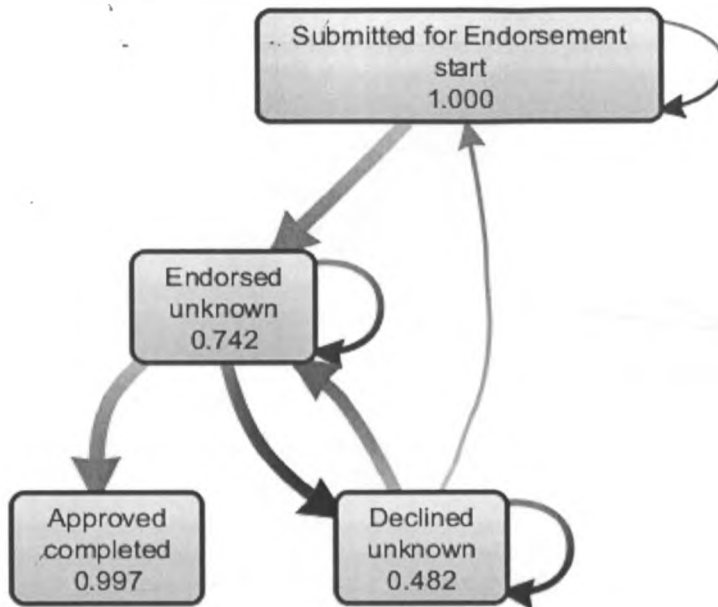


Figure 11: Process Model 2

2. How is the distribution of all cases over the different paths through the process?

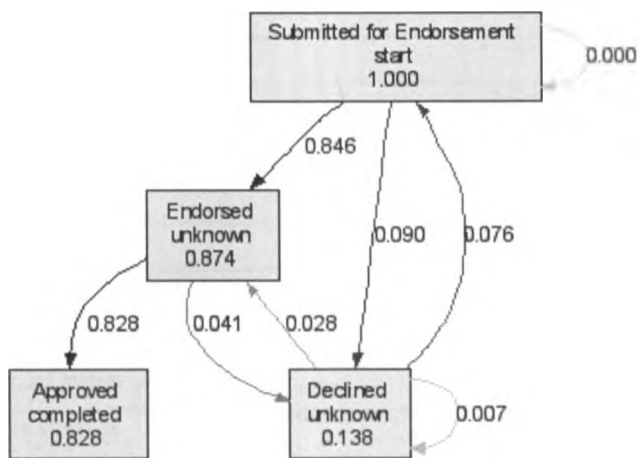


Figure 12: Distribution of Cases obtained from the frequency abstraction miner

The above diagram shows the following

From Activity	To Activity	Frequency
Submitted for Endorsement	Submitted for Endorsement	0%
Submitted for Endorsement	Endorsed	84.6%
Submitted for Endorsement	Declined	9%
Declined	Submitted for Endorsement	7.6%
Endorsed	Approved	82.8%
Endorsed	Declined	4.1%
Declined	Endorsed	2.8%
Declined	Declined	0.7%

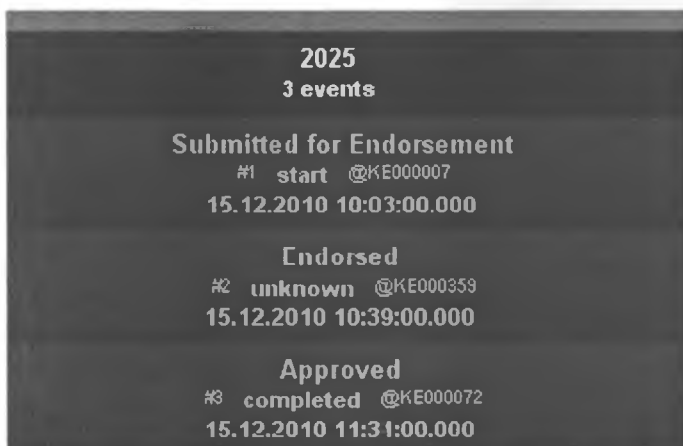
**Table 4: Distribution of cases**

3. How compliant are the cases (i.e. process instances) with the deployed process models? Where are the problems? How frequent is the (non-) compliance?

The process model set out in the system as explained earlier is as follows;

- Raising the LPO
- Submitting the LPO for endorsement to the first partner, the partner may decline at this point and it goes back to the initiator of the LPO, the initiator may submit it again or it ends at this point
- Endorsement of the LPO
- Request for approval to the second partner, the partner may decline and the LPO goes back to the endorser. It may be endorsed again or it ends at this point
- Approval of the LPO

Most of the cases are compliant; they follow the most frequent path of 3 tasks in the process, i.e. submission for endorsement, endorsement, approval, as in the diagram below;



**Figure 13: compliant case following the most frequent path**

Other cases whereby there is decline follow the process as in the example below;

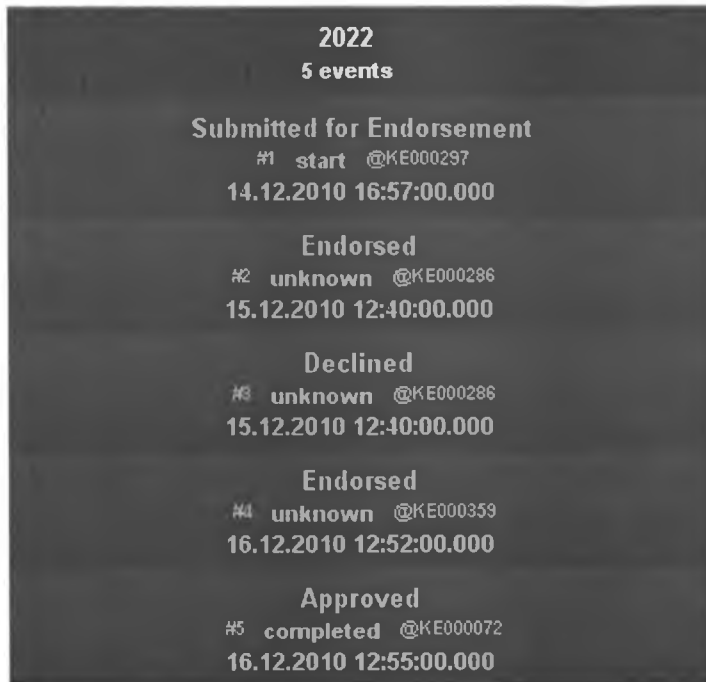
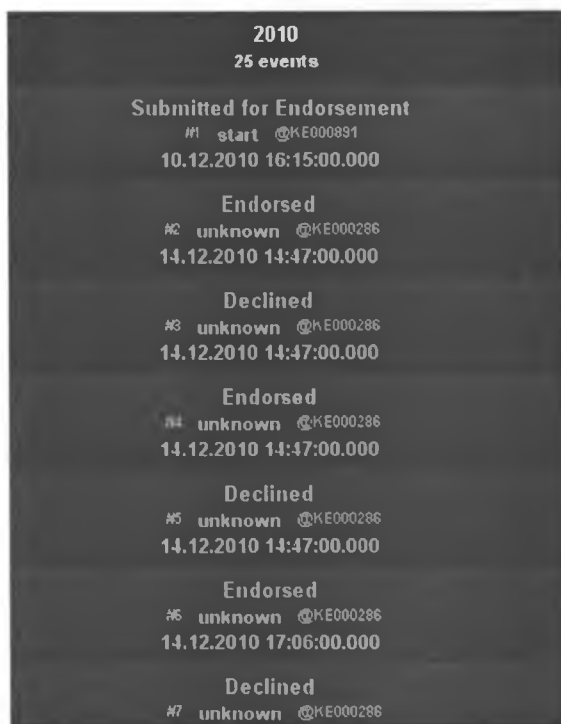


Figure 14: Compliant case with the path covered by some of the cases

However, there are a few exceptions. As in case no. 2018, there were 9 events. This is unusual as there were several tasks for endorsing and declining by the same partner. This cannot be the case since if a partner endorses the LPO goes to another partner for an approval. Another exception noted was case no. 2010, which had 25 events of the same nature as the one above. See the diagram below;



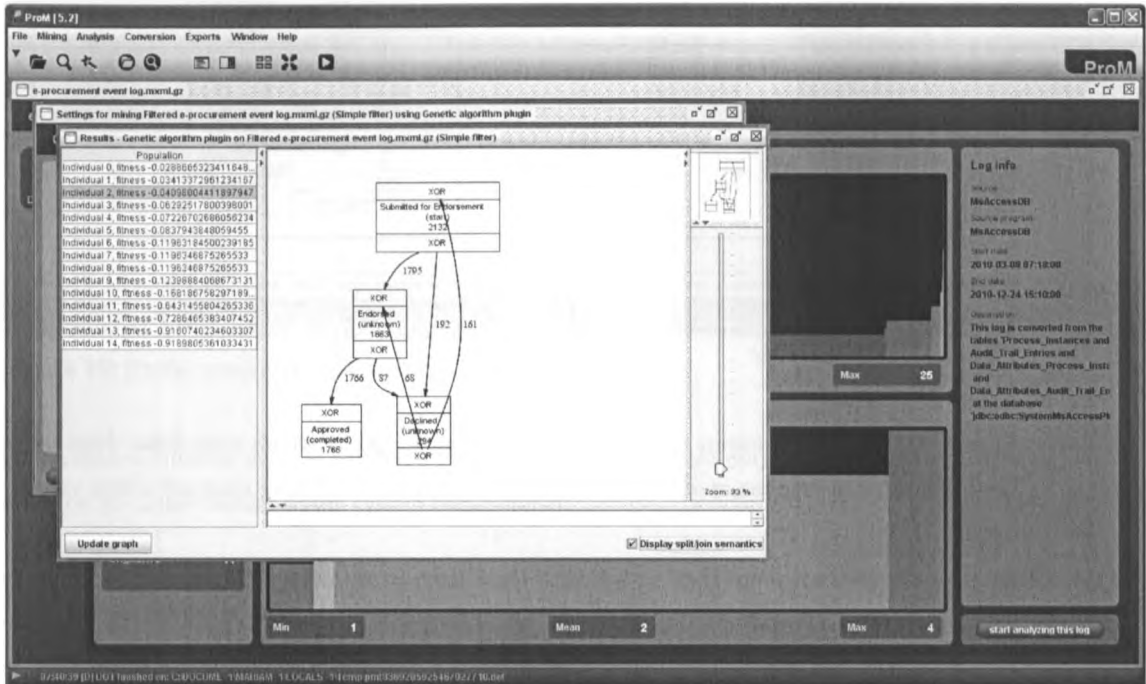


**Figure 15: Non-compliant case with 25 events, depicting abnormality**

Abnormalities are an indication of a bug, fraud etc. The frequency of the exceptions was however not much in this log.

4. What are the routing probabilities for each split task (XOR or OR split/join points)?

The diagram below indicates the routing probabilities of the events in the process. E.g. at the start stage an event can either be endorsed or declined.



**Figure 16: Routing probabilities**

5. Are the events following laid down policies

Organizations have different procedures which must be followed, e.g. a person cannot approve his own request. In the diagram below, the plug-in that assists in checking whether the policies have been followed is shown, it is known as LTL checker.

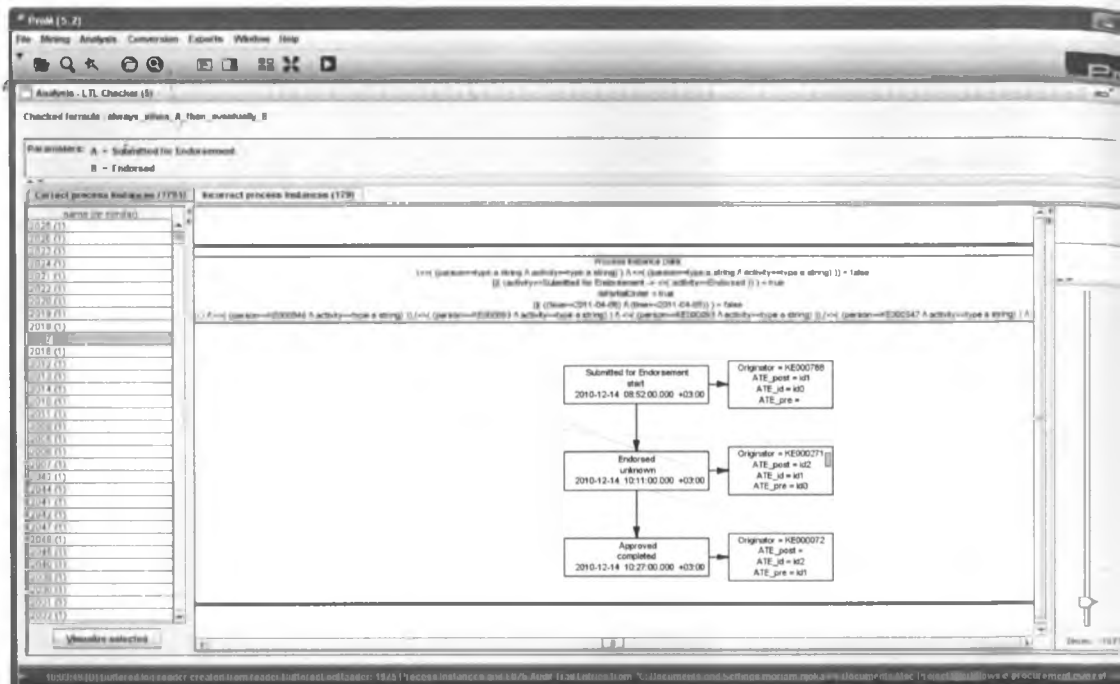


Figure 17: LTL checker

In this case, given two events that follow one another, it checks which cases have these events follow one another. 1791 have returned a true value while 179 of the cases have returned a false value.

- Which paths take too much time on average? How many cases follow these routings? What are the critical sub-paths for these paths?

Using the performance sequence diagram, 19 patterns came up and upon filtering off those patterns that time in minutes less than zero and those that had only one case, we were left with 15 patterns. This shows that the cases can follow any of the 15 patterns, the requirement here is to evaluate what patterns are acceptable and get rid of the unacceptable ones.

E.g. in the diagram below pattern one takes an average of 59 hours and follows the most frequent path shown in figure 10.

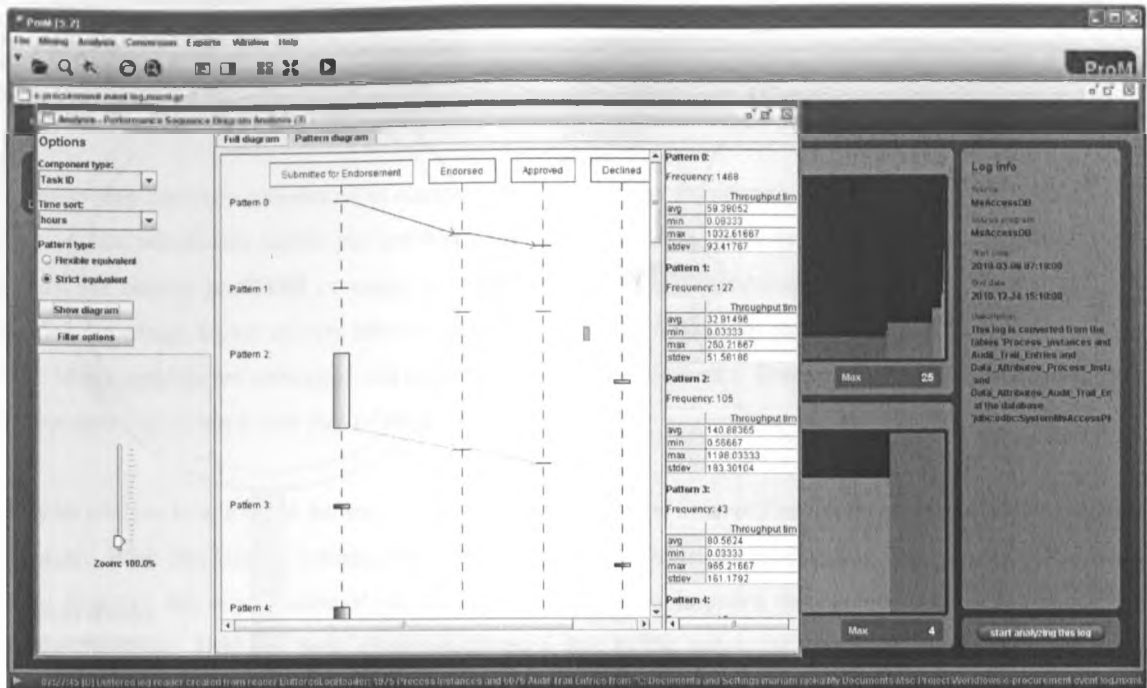


Figure 18: Performance on different patterns

Given the information derived using the above mining and analysis plug-ins, then recommendations can be made on the improvement of the system being mined.

## The Proposed Framework

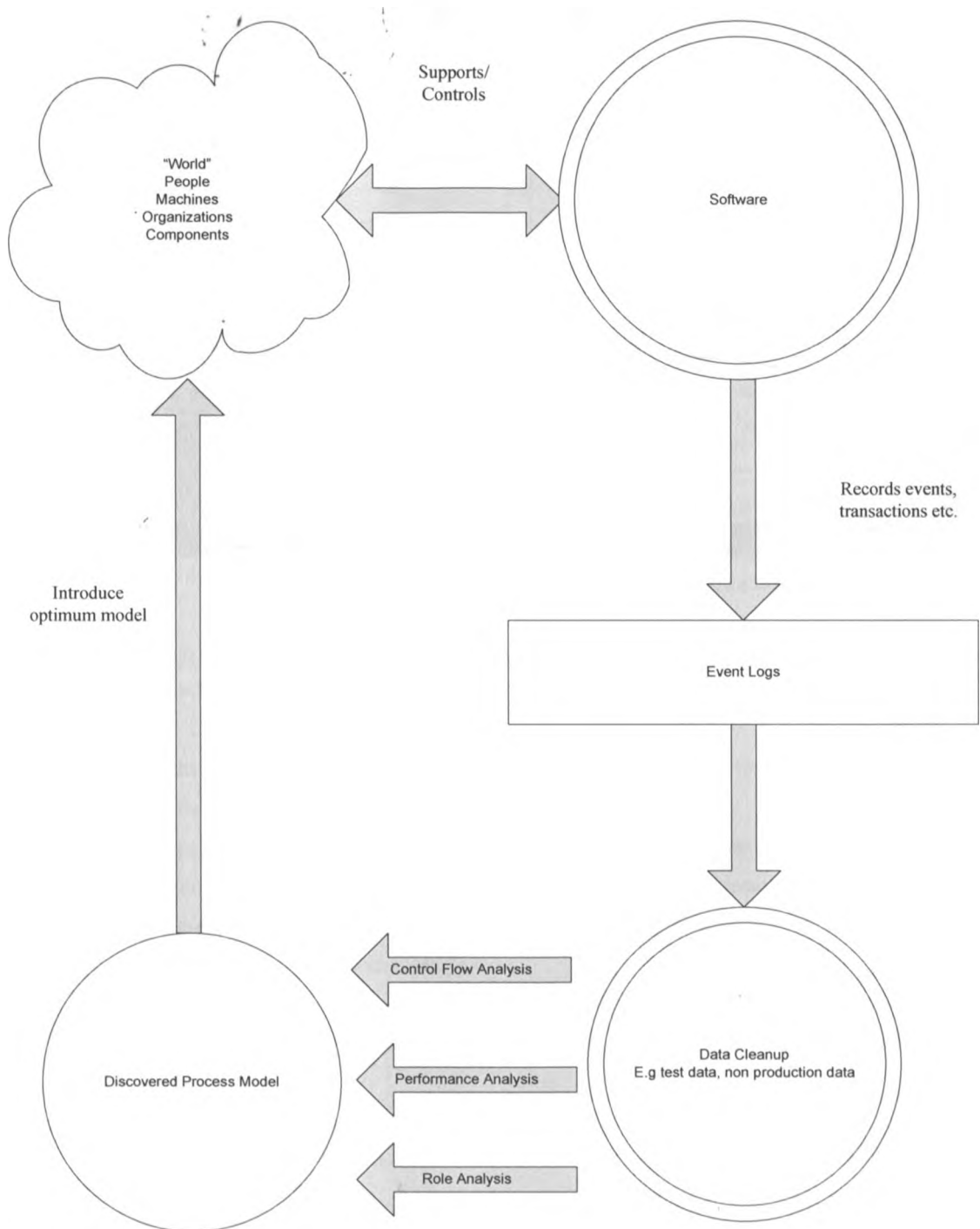


Figure 19: The proposed framework

The proposed framework takes in the vision of the world, the business processes, which are then converted into systems. Through the systems, event logs are created. Event logs are then mined and the next step is conducted.

The data cleanup step has been introduced as compared to the previous framework. Given that most semi structured systems contain data which may not be just production data, e.g. test data. Cleanup in this case becomes prominent to ensure only the correct processes or cases are picked for analysis. This handicap in semi-structured systems makes data cleanup a huge aspect as very often there will be data which does not conform perhaps it was testing data or otherwise. Many systems are debugged and corrected on live environments. Thus cleanup and preparation of data should be considered as an important part of the process.

The best way to analyse is to look at the logs with a view of getting the *control flow analysis*, *performance analysis* and *role analysis*. With this kind of information, then all manner of decisions touching on how to control the work flow, how to improve the organizational/role allocation and how to improve the performance of the processes themselves can be done. This has been given prominence due to the nature of discovery that is done on semi structured systems. These systems in many cases will not have had previous statistical information of these crucial data. Considering there may be no documented process models to follow, or even performance tests done on the system previously and recorded then these analysis methods will give us the information.

### **Control flow analysis**

The data is analyzed to discover all the different process patterns from the event logs. These are then analyzed to get the most optimum one. This will give us the different paths that the processes in the system follow.

E.g.

Submit for endorsement – Decline – Submit for endorsement – Endorsed – Approved

Submit for endorsement – Endorsed – Approved

All these paths can then be analyzed in terms of frequency, correctness among others.

### **Performance analysis**

Given the timestamp available in logs, information such as average time for the process, average times per pattern, minimum, maximum etc are discovered. In performance analysis all the aspects of what could affect the success of the process in analyzed. Things to be considered are like the length of the path a process follows, the time it takes to complete, the average for the processes.

With this kind of statistics, simulations can be done and the most optimum model derived.

### **Role analysis**

With the tools that discover the relationships between the people carrying out the activities, detailed role analysis can be carried out and streamlined to improve the process. E.g. Roles being carried out by several people may show some inefficiencies thus reductions can be done etc.

In addition, workloads and the flow from one person to another can be analyzed, with this knowledge, organizations are then able to tell the bottlenecks and steps are taken to clear them.

The result of the analysis is the discovered process model which will be incorporated into the organization either through a soft process change or business process reengineering given the weight. The software will have to be changed accordingly to reflect the new business process and the cycle continues.

## Implications of Research

A process model can have positive or negative effects on an organization. If properly designed then efficiencies are increased otherwise there are many organizations that have introduced computerized systems and are yet to feel the benefits due to the poor structure of the process model on that system.

Given that many semi-structured systems did not follow a structured way when they were developed, then its most likely they will be having a negative effect on the organization they are running in.

This research can be used to assist many organizations in cleaning up their systems and coming up with optimal processes to run their operations.

## Conclusion

In this research, the concentration has been on the process of process mining and how applicable it can be using the context of a procurement system. It takes you through the process of data collection and its preparation thereof. It also describes how to convert the data to the MXML format acceptable in the ProM framework. The research takes you through the process of mining and analysis and the result is the discovery of the optimum process mining framework by going through control flow analysis, performance analysis and role analysis using the tools available in the ProM Framework. These tools assisted in production different types of information such as model patterns, frequencies etc.

The literature review gives a good understanding of process mining, the ProM framework and MXML technologies. A good understanding of the technologies is required to implement process mining. Further research can be carried out as it is a relatively new area. Future work can focus on developing more tools that can be used to analyze entry logs and come up with meaningful information.

This research has demonstrated the implementation of process mining on a system that is already in use, this has led to coming up with a framework that can be used to analyze and eventually improve these systems. The framework introduced was initially derived from the framework developed by R.S. Mans, M.H. Schonenberg, M. Song, W.M.P. van der Aalst, and P.J.M. Bakker. The framework looks at the special circumstance of semi-structured systems which are common in the developing world. It is needed urgently in streamlining processes within the application systems developed locally due to their lack of structure during development.

By following this framework, systems developers will be able to analyze the impact of their systems after they have gone into production and therefore come up with better process models and improvements. It will assist in structuring systems that have already started being used.



## Bibliography

1. Peter Van Den Brand. (2004). *Architecture of the process mining framework*
2. B.F. van Dongen, A.K.A. de Medeiros, H.M.W. Verbeek, A.J.M.M. Weijters and W.M.P. van der Aalst. (2005). *The ProM framework: A new era in process mining tool support*, 26th International Conference on Applications and Theory of Petri Nets, G. Ciardo and P. Darondeau, LNCS 3536, pages 444-454, 2005
3. <http://prom.win.tue.nl/tools/prom/> by the Process Mining Group, Eindhoven Technical University.
4. W.M.P. van der Aalst and A.J.M.M. Weijters. *Process Mining: A Research Agenda*. Department of Technology Management, Eindhoven University of Technology, P.O. Box 513, NL-5600 MB, Eindhoven, The Netherlands.
5. W.M.P. van der Aalst and H.T. de Beer and B.F. van Dongen. *Process Mining and Verification of Properties: An Approach based on Temporal Logic*. Department of Technology Management, Eindhoven University of Technology, P. O. Box 513, NL-5600 MB, Eindhoven, the Netherlands.
6. Gregory A. Hansen. (1997). *Automating Business Process Reengineering, Second Edition*. Prentice hall PTR
7. W.M.P. van der Aalst and A.J.M.M. Weijters. *Process Mining: A Research Agenda*. Department of Technology Management, Eindhoven University of Technology, P.O. Box 513, NL-5600 MB, Eindhoven, The Netherlands.
8. W.M.P. van der Aalst and H.T. de Beer and B.F. van Dongen. *Process Mining and Verification of Properties: an Approach based on Temporal Logic*. Department of Technology Management, Eindhoven University of Technology, P. O. Box 513, NL-5600 MB, Eindhoven, The Netherlands.
9. Gregory A. Hansen. (1997) *Automating Business Process Reengineering, Second Edition*. Prentice hall PTR
10. Melike Bozkaya, Joost Gabriels, Jan Martijn van der Werf LaQuSo, *Process Diagnostics: a Method Based on Process Mining* Laboratory for Quality Software
11. Yannis Stavarakas 1, Manolis Gergatsoulis 1, and Panos Rondogiannis. *Multidimensional XML*. Institute of Informatics & Telecommunications
12. R.S. Mans1, M.H. Schonenberg1, M. Song1, W.M.P. van der Aalst1, and P.J.M. Bakker2. *Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital*.
13. Luca Rossetti (2006) What is Business Intelligence <http://searchdatamanagement.techtarget.com/definition/business-intelligence>

## Appendices

### Appendix 1: How to configure an ODBC connection to a Microsoft Access Database on your computer

The procedure to load set up an ODBC connection to a Microsoft Access database located at your computer is:

1. Open the "Control Panel" window by clicking on "Start/Setting/Control Panel".
2. Double-click "Administrative Tools".
3. Double-click "Data Sources (ODBC)".
4. Select the tab "System DSN".
5. Click on the button "Add..." of the "System DSN" tab. You should get a window like the one in Figure 12.

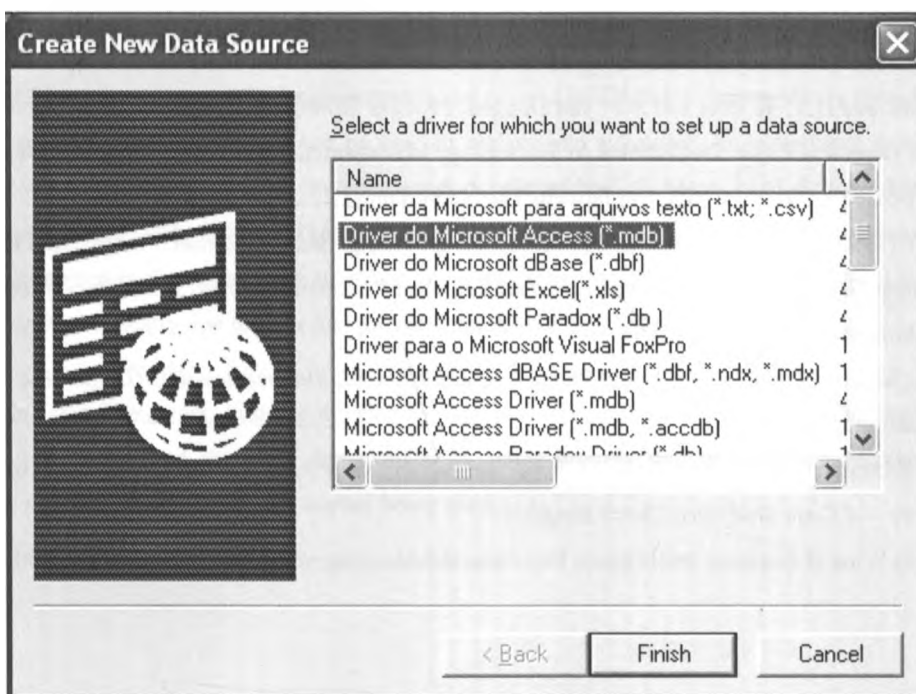


Figure 20: Screenshot of the window to create a new data source

6. Select "Driver to Microsoft Access (\*.mdb)" and click the button "Finish".
7. The next window is like the one in Figure 13. This window allows you to define the Provide the "Data Source Name", click on the button "Select..." to inform where the Microsoft Access database is located (see Figure 14), and on the button "Advanced..." to set up the "Username" and "Password" to be able to access this database. See Figure 15.

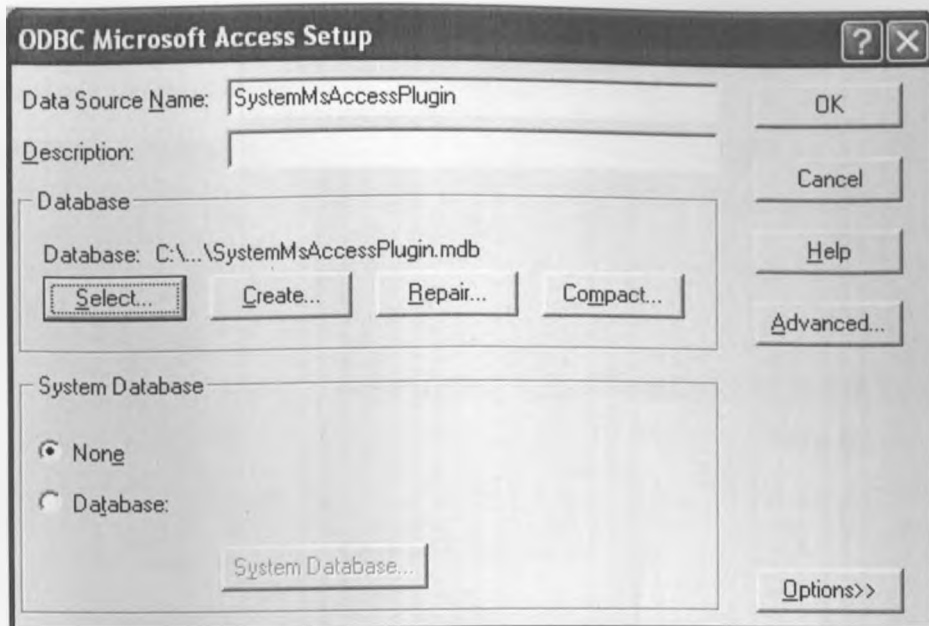


Figure 21: Screenshot of the window to set up the name of the database that has to be provided after jdbc:odbc: in the field DbHostUrl in the MS Access database plugin

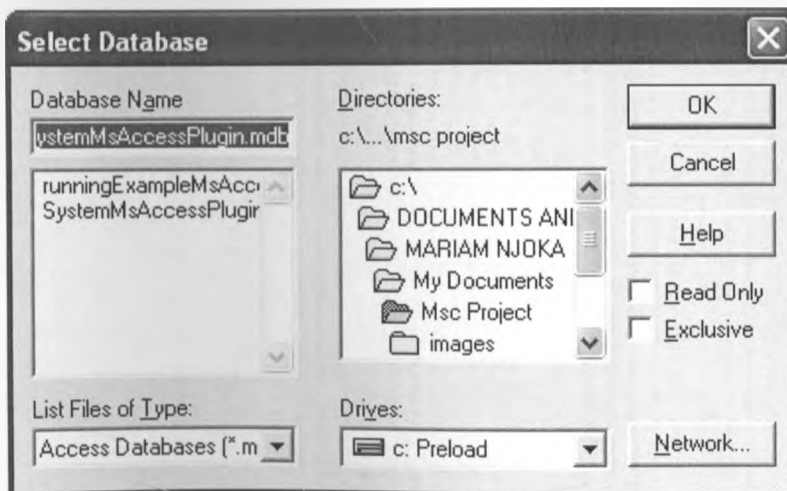
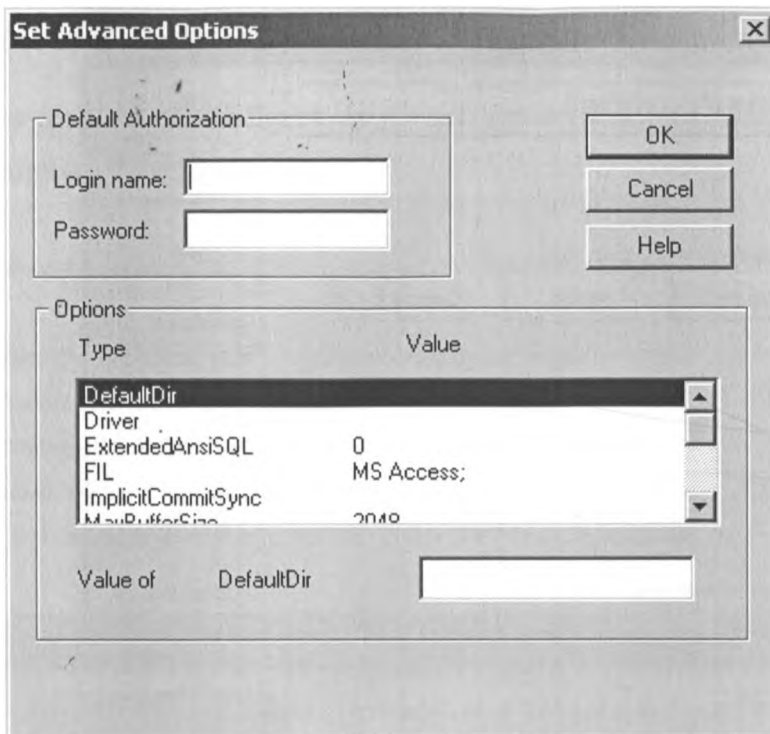


Figure 22: Screenshot of the window to set up the database location



**Figure 23:** Screenshot of the window to set up the "username" and "password" provided to MS Access database plugin

## Appendix 2: Log summary of the event log

The log has been truncated in some parts to avoid it being too long.

Log Summary

Number of processes: 1

Total number of process instances: 1975

Total number of audit trail entries: 6076

Name: logs\_Process\_instancesGLOBAL.mxml.gz

Description: This log is converted from the tables 'Process\_instances and Audit\_Trail\_Entries and Data\_Attributes\_Process\_Instances and Data\_Attributes\_Audit\_Trail\_Entries' at the database 'jdbc:odbc:SystemMsAccessPlugin'

Attribute name	Value
os.version	5.1
os.arch	x86
user.name	Mariam Njoka
mxml.creator	MXMLib ( <a href="http://promimport.sf.net/">http://promimport.sf.net/</a> )
java.version	1.6.0_17
mxml.version	1.1
java.vendor	Sun Microsystems Inc.
os.name	Windows XP
app.name	ProM Import Framework
app.version	7.0 (Propeller)

Source

Name: MsAccessDB

Description:

Attribute name	Value
program	MsAccessDB

## Process Instances

Number of process instances entries: 1975

Process Instance	Occurrences (absolute)	Occurrences (relative)
100, 1000, 1001, 1002, 1003, 1004, 1005, 1006, 1007, 1008, 1009, 101, 1010, 1011, 1012, 1013, 1014, 1015, 1016, 1017, 1018, 1019, 102, 1020, 1021, 1022, 1023, 1024, 1025, 1026, 1027, 1028, 1029, 103, 1030, 1031, 1032, 1033, 1034, 1035, 1036, 1037, 1038, 1039, 104, 1040, 1041, 1042, 1043, 1044, 1045, 1046, 1047, 1048, 1049, 105, 1050, 1051, 1052, 1053, 1054, 1055, 1056, 1057, 1058, 1059, 106, 1060, 1061, 1062, 1063, 1064, 2044, 2045, 2046, 2047, 2048, 2049, 205, 2050, 2051, 2052, 2053, 2054, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304	1	0.051%

## Log events

Number of audit trail entries: 5

Model element	Event type	Occurrences (absolute)	Occurrences (relative)
Submitted for Endorsement	start	2132	35.089%
Endorsed	unknown	1868	30.744%
Approved	completed	1770	29.131%
Declined	unknown	298	4.905%
Submitted for Itinerary Advice	start	8	0.132%

## Starting Log Events

Number of audit trail entries: 2

Model element	Event type	Occurrences (absolute)	Occurrences (relative)
Submitted for Endorsement	start	1970	99.747%
Submitted for Itinerary Advice	start	5	0.253%

## Ending Log Events

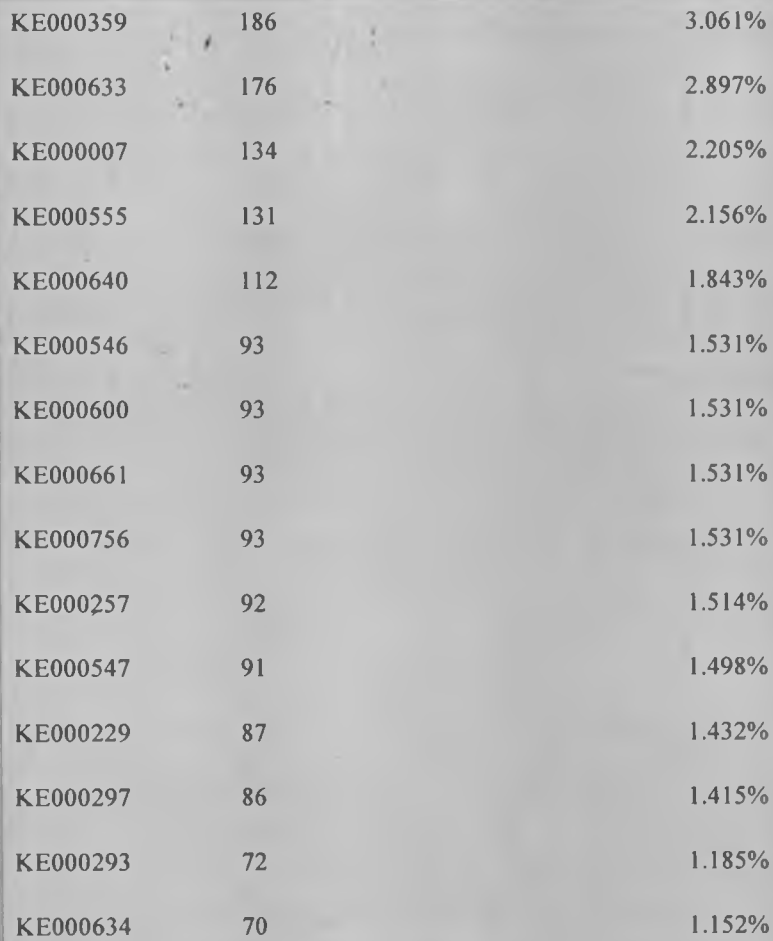
Number of audit trail entries: 5

Model element	Event type	Occurrences (absolute)	Occurrences (relative)
Approved	completed	1770	89.62%
Submitted for Endorsement	start	135	6.835%
Declined	unknown	59	2.987%
Endorsed	unknown	10	0.506%
Submitted for Itinerary Advice	start	1	0.051%

## Originators

Number of originators: 77

Originator	Occurrences (absolute)	Occurrences (relative)
KE000072	1275	20.984%
KE000258	847	13.94%
KE000485	344	5.662%
KE000204	262	4.312%
KE000286	252	4.147%
KE000093	212	3.489%



KE000359	186	3.061%
KE000633	176	2.897%
KE000007	134	2.205%
KE000555	131	2.156%
KE000640	112	1.843%
KE000546	93	1.531%
KE000600	93	1.531%
KE000661	93	1.531%
KE000756	93	1.531%
KE000257	92	1.514%
KE000547	91	1.498%
KE000229	87	1.432%
KE000297	86	1.415%
KE000293	72	1.185%
KE000634	70	1.152%

**Figure 24: Log Summary**





**University of Nairobi  
School of Mathematics (Chiromo Campus)**

**HIV AND HSV-2 CO-INFECTION AMONG FISHERMEN ALONG  
LAKE VICTORIA, KENYA: AN APPLICATION OF MULTIPLE  
LOGISTIC MODEL //**

**By  
Margaret W. Mburu  
156/73062/2009**

**Internal Supervisors:**

**1<sup>st</sup> Dr. Thomas Achia  
2<sup>nd</sup> Mrs. Anne Wang'ombe**

University of NAIROBI Library



0478782 6

Project submitted in partial fulfillment of the requirements for the degree of  
Master of Science in Biometry 2009 -2010

*Date Presented: 26<sup>TH</sup> July 2011*

**Certification**

This is to certify that this report was written by Margaret W. Mburu under our supervision as her original work which has not been submitted for award of a degree in any other University.

**Margaret W. Mburu**



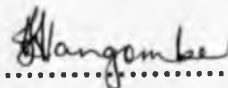
.....  
**Student**

**Dr. Thomas Achia**



.....  
**1st Supervisor**

**Mrs. Anne Wang'ombe**



.....  
**2nd Supervisor**

<b>Table of Contents</b>	<b>Page</b>
CHAPTER 1:INTRODUCTION.....	10
1.1 BACKGROUND OF THE STUDY.....	10
1.1.1 HIV and HSV-2 Co-infection .....	10
1.1.2 HIV.....	10
Acquisition of HIV .....	13
Symptoms and Diagnosis of HIV .....	13
Resistance to Treatment.....	14
1.1.3 HSV2 Virology.....	16
1.2 OBJECTIVES OF THE STUDY.....	18
1.2.1 Primary objective .....	18
1.2.2 Secondary objectives .....	19
1.3 SIGNIFICANCE OF THE STUDY .....	19
1.4 LITERATURE REVIEW.....	19
1.5 DATA/DESIGN OF THE STUDY.....	27
CHAPTER 2: METHODOLOGY .....	29
2.1 THEORETICAL MODEL .....	29
2.2 PARAMETER ESTIMATION: MAXIMUM LIKELIHOOD .....	31
2.3 MODEL SELECTION.....	35
2.3.1 Study Variables .....	36
CHAPTER 3: RESULTS .....	38
3.1 RESULT TABLES .....	43
CHAPTER 4: DISCUSSION AND CONCLUSION.....	50
List Of References .....	<b>5Error! Bookmark not defined.</b>

## ABSTRACT

**Background** HSV-2 infection enhances the transmission and acquisition of HIV. Identifying risk factors for HIV and HSV-2 co-infection provides the opportunity to develop health promotion programmes based on primary prevention.

**Objective** To establish risk factors for the presence of HIV and HSV-2 Co-Infection Among Fishermen along Lake Victoria, Kenya.

**Methods** Two-stage cluster sampling to select 168 fishermen for the structured interviews and STI testing was used. Subjects were tested for HIV, Syphilis and HSV-2 infections. In addition, the demographic and behavioural characteristics of the subjects were assessed through face to face interviews. The prevalence of HIV was tested using rapid methods and cases confirmed by ELISA while Syphilis was tested using the rapid plasma reagin (RPR) card test. HSV-2 was tested using ELISA. Fishermen with equivocal HSV-2 test result are removed from this analysis leaving a samole 126 fishermen for this project.

**Results** A total of 28(22.2%) fishermen were co-infected with HIV-1 and HSV-2 viruses. Overall 72(57.1%) fishermen were infected with HSV-2 (95% CI,48.4-65.9%) and 31(24.6%) were infected with HIV-1 (95% CI,17.5-31.7%). The risk factors for HIV-1 and HSV-2 co-infection are marital status, previously infected with gonorrhoea or syphilis and number of sexual partners in the last one month.

**Conclusion** Approaches to reducing the transmission of HSV-2 and HIV-1 are needed. From both clinical and public health perspectives, there is a clear imperative to test different approaches to interrupting the synergistic link between HSV-2 and HIV-1.

## **GLOSSARY**

**HIV**-Human Immuno-Deficiency Virus

**HSV-2**-Herpes Simplex Virus Type 2

**GUD**-Genital Ulcer Disease

**HSV-1**- Herpes Simplex Virus Type 1

**STI**-Sexually Transmitted Disease

**GLM**-Generalized Linear Model

**AIC**-Akaike Information Criterion

**CI**-Confidence Interval

**RR**-Relative Risk

**AIDS**- Acquired Immunodeficiency Syndrome or Immune Deficiency Syndrome

**LAT**- Latency-Associated Transcript

**HLAs** -Histocompatibility Antigens

**DISC** -Disabled Infectious Single Cycle

**RCTP** -Research Care and Training Program

**KEMRI**-Kenya Medical Research Institute

**RPR** -Rapid Plasma Reagin

**ELISA**- Enzyme-linked immunosorbent assay

**HPV** -Human Papilloma Virus

**ML**- Maximum Likelihood

**ANOVA**-Analysis of Variance

**CD-4**-Cluster of Differentiation 4

**RNA**-Ribonucleic Acid

**DNA**-Deoxyribonucleic acid

**FDA**-Food and Drug Administration

**LNF**-Laparoscopic Nissen Fundoplication

**ONF**- Open Nissen Fundoplication

**NRTI**-Nucleoside Reverse Transcriptase Inhibitor

## **ACKNOWLEDGEMENT**

Thanks to the MSC colleagues for their various inputs and support during the course. Without their support, the task could have been insurmountable.

To my parents Keziah Muthoni and James Mburu, family members and friends, I say thank you for your understanding, moral support, prayers, and encouragement that kept me going even when things appeared tough.

I can not forget to thank the late Dr. Rosemary Nguti for her invaluable input in the whole entire course, may she rest in peace.

**DEDICATION**

This piece of work is dedicated to my son David Gideon Mburu for his understanding, care, prayers, moral support and his love.

Further I salute him for his perseverance during the long hours of my absence whose fruits are evident through this project paper.

God richly bless you, David.



**ABOVE ALL I THANK THE ALMIGHTY GOD FOR THIS FAR HE HAS TAKEN ME  
FAITHFULLY**

**HE HAS MADE ALL THINGS POSSIBLE**

**GLORY BE TO HIS NAME**

## CHAPTER 1.0 INTRODUCTION

### 1.1 BACKGROUND OF THE STUDY

HSV-2 infection enhances the transmission and acquisition of HIV. Herpes simplex virus type 2 (HSV-2) is the most common cause of genital ulcer disease (GUD) worldwide. HSV-2 prevalence in sub-Saharan Africa ranges from 27-57% in men to 30-74% in women [18,19] and is higher in Human Immuno-Deficiency Virus (HIV) infected persons [20]. Epidemiologic studies suggest synergy between HIV-1 and HSV-2 that facilitates the spread of both viruses, with HSV-2 increasing HIV-1 susceptibility and infectiousness [21] and HIV-1 infection increasing HSV-2 reactivation frequency [22].

#### 1.1.1 HIV and HSV-2 Co-infection

HIV-1 is often acquired in the presence of pre-existing co-infections, such as HSV-2, tuberculosis, chronic diarrhoea, oral candidiasis, cytomegalovirus, hepatitis B virus and hepatitis C virus. The effects of co-infection on clinical course of HIV-1 infection may be complex and vary by the co-infecting agent. The name herpes comes from the greek *herpein*—"to creep." HSV-2 infection enhances the transmission and acquisition of HIV. Genital herpes is associated with a two- to three-fold increased risk of HIV acquisition and an up to five-fold increased risk of HIV transmission per-sexual act, and may account for 40% to 60% of new HIV infections in populations where HSV-2 has a high prevalence [1]. Significant numbers of HIV infections could potentially be averted if HSV-2 could either be prevented or suppressed. However, two recent trials reported that HSV-2 suppressive therapy currently available did not reduce HIV acquisition [2,3].

HIV in turn, increases the risk of HSV-2 transmission [4]. Outbreaks of HSV-2 are generally more severe, extensive, persistent, and invasive for those with more advanced HIV disease [5]. In fact, persistent HSV-2 infection was one of the original opportunistic infections that resulted in the identification of AIDS [6].

However, few data are available on the interaction of HIV and HSV-2 infections.

### **1.1.2 HIV**

HIV which leads to AIDS is the pandemic of our time, the world has come a long way since the disease was first discovered, but we still have far to go to stop this global threat. The virus can remain dormant in human body for ten years and produce 10 million copies of itself so you can not fight infection and get sick easily.

HIV is a virus that attacks a person's CD-4 cells which are also called T-cells. It gets inside a CD-4 cell which protects against germs and make up your immune system. Then, it makes copies by injecting its viral RNA into these cells. The RNA soon becomes DNA by using reverse transcriptase enzymes. The new DNA combines with the old CD-4 cell DNA, redirects the production of the cell and the cell begins to make a long string of several different viral proteins which are then cut into smaller pieces and used to coat the new viral DNA which is the HIV RNA combined with the CD-4 cells DNA. Soon the individual viral proteins are combined with the viral DNA and the newly made HIV is released into the bloodstream to infect other CD-4 cells and continue copying itself. HIV can make 20 million copies of itself each day. AIDS which stands for Acquired Immunodeficiency Syndrome or Immune Deficiency Syndrome is an advanced stage of HIV when the immune system is already damaged from the HIV virus not allowing the CD-4 cells to work properly.

There are medicines that alter the HIV cycle causing the HIV to not be able to reproduce the way it normally would. There are five different types all designed to work differently. Usually someone would use a combination of more than one medicine or more than one medicine family. One family is called Nucleoside reverse transcriptase inhibitors these are called NRTIs these medicines work by affecting the building blocks that the HIV uses to make copies of itself. The NRTIs family contains Retrovir (zidovudine) which is the much famed AZT which was one of the first HIV AIDS treatment approved in the United States. NtRTIs are similar to NRTIs, but NrRTIs are chemically pre-activated allowing them to get into the bloodstream more rapidly. Non-nucleoside reverse transcriptase inhibitors are also called NNRTIs. NNRTIs work by affecting the enzyme that allows the cell to make RNA then the cell can not produce the DNA or complete the reproductive cycle. Fusion inhibitors (FIs) work by blocking the HIV from attaching to or getting inside a CD-4 cell. The last group Protease inhibitors (PIs) stop the protease enzyme from cutting the viral protein chains into smaller pieces to coat the new viral DNA. It can not reproduce so it does not affect any other

CD-4 cell. In all these medicines there is no cure just treatment to help stop HIV from becoming AIDS as fast as without treatment.

Viral load is the number of HIV copies that are in one milliliter of human blood. Viral load is also called viral level. High viral load number near one million or more copies mean HIV is getting worse and may lead to AIDS sooner than a low viral load. An undetectable viral load means it is below 50 and this shows that the HIV in the blood is very low or none, but not that it is gone. HIV never goes away though it can remain dormant.

Viral load can change the medication a doctor subscribes to control viral load. Viral load in HIV AIDS patients is tested usually every three to four months through a blood test. CD-4 or T-Cell counts also affect the viral load. CD4 cell count and viral load cut-offs have been defined to guide clinicians in when to start antiretroviral treatment. A person can also know if the medicine they are taking is working or if they need medication changes by the viral load. Resistance is when the medication quits working, the HIV mutates or it becomes immune to the medication. This can cause a viral load to sky rocket and it can alert a doctor that the medication needs to be changed. It can be started by not taking medicines properly, being re-infected with HIV, or because of mutation. Resistance is silent and called the silent side effect.

HIV was brought to attention of the general public in 1981. It had been affecting those in Africa for years where it was called "Slim" because a person's body would deteriorate making them appear "slim" before death. Scientists hypothesize that it is spread by intimate contact until when the term AIDS came into use.

Institut Pasteur in France discovered the viruses HIV1 and HIV2. 1985 brought us the first HIV antibody test that is FDA approved.

Finally in 1987, the first HIV treatment was released which was AZT (zidovudine, Retrovir) from Glaxo Wellcome (Aegis). The World Health Organization also established their Global Program on AIDS. World AIDS day was conceived by one hundred forty countries when health ministers from all over the world met in London to discuss HIV AIDS. The first world AIDS day which was December 1 1988 was themed "A World United Against AIDS". The Food and Drug Administration approved the use of pentamidine mist against PCP which is an opportunist infection that AIDS allows to be let into the system because of the low immune

system level. In 1991, a new drug ddI (didanosine, Videx) was released for use it is a nucleoside reverse transcriptase inhibitor or NRTI. The World Health Organization estimates that ten million people have HIV worldwide.

The AIDS threat in Africa continues to grow causing more deaths and more homeless children.

### ***Acquisition of HIV***

HIV and/or AIDS can be transmitted by having unprotected sex both heterosexual and homosexual, coming in contact with HIV infected semen or bodily fluids, direct blood contact, injection drug use, oral sex, using unclean syringes, sharing needles, blood transfusions, pregnancy (mother to infant), breast feeding (mother to baby), open mouth kissing if there is an open sore in the infected mouth, or in a health care setting. It is not a contagious disease, a sneeze or cough can not harm. Sharing razors and tooth brushes are frowned on by the Center for Disease Control because of the chance of direct blood contact.

Ways HIV and/or AIDS is not transmitted is hugging, touching, sharing a bathroom, sharing household items, contact with sweat or tears, sharing swimming pools, saunas, hot tubs, toilets, sinks, and facilities.

HIV and AIDS can be Prevented by using latex or polyurethane condoms with a water based lubricant for men, females using polyurethane condoms, finger cot, spermicides which might be able to help kill the HIV virus, but this has not been proven yet, using a dental dam during oral sex or a rubber barrier, washing or sterilizing sex toys, not engaging in risky behavior, having more than one sexual partner, not sharing needles or syringes to inject drugs or for any matter, abstaining from sex altogether. Condoms inhibit fluid contact and prevents STI infection. Having safe sex or using safe sex behaviors can decrease the risk of contracting HIV AIDS. Prevention is very important in not contracting or transmitting and treating this disease.

### ***Symptoms and Diagnosis of HIV***

HIV is diagnosed using blood, saliva, or cells from the inside of the cheek.

HIV infection comes in three stages: acute infection, chronic infection, and AIDS.

Acute HIV infection is the earliest and shortest stage of HIV infection. Not everyone gets symptoms, but most people come down with a flu-like illness three to six weeks after infection. The symptoms are the same as flu or mononucleosis: fever and fatigue lasting for a week or two. There may or may not be other symptoms; A blotchy red rash, usually on the upper torso, that does not itch, Headache, Aching muscles, Sore throat, Swollen lymph glands, Diarrhea, Nausea and Vomiting.

Sensitive new tests can tell whether you have acute HIV infection. Treatment during the acute stage of HIV infection works much, much better than later treatment. Be sure to tell your doctor about your HIV risk. If you don't, you may not get the right tests. Standard HIV tests -- either home tests or lab tests -- won't detect acute HIV infection. The body puts up a terrific struggle against HIV. At the end of this struggle, the body reaches a kind of standoff with the virus. This is chronic HIV infection, which begins three to six months after a person gets HIV. There aren't any symptoms. For most people, this stage of HIV infection lasts about 10 years. Even though there are no symptoms, the immune system slowly runs down. A normal person has a CD4 T-cell count of 450 to 1,200 cells per microliter. When people with HIV have their T-cell counts drop to 200 or lower, they have reached the stage of AIDS. AIDS itself has no symptoms. Because the immune system is devastated, disease symptoms are specific to the kind of infections a person may have. When a person's T cells get very low, doctors prescribe drugs to prevent infections.

AIDS patients may have some of the following symptoms:

- Being tired all the time
- Swollen lymph nodes in the neck or groin
- Fever lasting more than 10 days
- Night sweats
- Unexplained weight loss
- Purplish spots on the skin that don't go away
- Shortness of breath
- Severe, long-lasting diarrhea
- Yeast infections in the mouth, throat, or vagina
- Easy bruising or unexplained bleeding

### **Resistance to Treatment**

HIV weakens the immune system and without treatment, it leads most infected people to develop AIDS. When CD4 cell counts are below 200, it means the infected person has progressed from HIV infection to AIDS. There is no cure for AIDS. Antiretroviral treatment can suppress HIV and can delay illness for many years.

HIV resistance to zidovudine (AZT) is conferred by several mutations in the HIV-1 reverse transcriptase (RT) gene, including Thr-215–Tyr (T215Y), K70R, D67N, M41L, and L210W [25]. The T215Y mutation is a primary mutation observed after AZT treatment. T215Y alone reduces the susceptibility for AZT 16-fold and is the first mutation seen in the majority of patients receiving combination therapy with AZT and other nucleoside analogs such as didanosine (ddI) or zalcitabine (ddC) [26,27]. The 215Y mutation also has been found in some patients treated with stavudine (d4T), and its presence in AZT-experienced patients may compromise the response to subsequent treatments with d4T [28-30]. The widespread use of antiretroviral drugs to treat HIV-1- infected persons has raised concerns regarding transmission of drug-resistant HIV-1. Surveillance of drug-resistant HIV-1 in recently infected persons has documented transmission of HIV-1 carrying several resistance mutations including 215Y. In addition, a unique set of mutations at codon 215 of HIV-1 RT has been found in treatment-naïve HIV-1-infected persons. These mutations are mainly 215C(TGC) and 215D(GAC), although other amino acids such as 215N(AAC) and 215S(TCC) also have been seen. Studies of HIV-1 seroconverters infected with viruses carrying the T215Y mutation have shown that 215D, 215C, 215N, and 215S represent revertants of HIV-1215Y. HIV-1215D, HIV-1215S, and HIV-1215N all have been shown to have higher replicative fitness than HIV-1215Y in the absence of AZT *in vitro*, which may explain why the 215Y mutation reverts through these intermediates *in vivo*. However, the determinants that influence selection of a particular revertant are not fully understood. Longitudinal studies of small numbers of patients have shown that 215C, 215D, 215N, and 215S are stable mutations that can persist in the absence of antiretroviral therapy. For instance, Yerly *et al.* reported persistence of 215D and 215C for a period of 1–2 years in four patients, whereas de Ronde *et al.* found persistence of 215D, 215C, 215S, and 215N for 1–3 years in five patients. The stability of these mutations has been explained by small differences in replicative fitness between these viruses and the wild-type (WT) HIV-1T215. Information on the effect of these mutations on susceptibility to nucleoside analogs is limited to four patients who had viruses with 215C or 215D and were found to be sensitive to AZT. However, because these mutations

differ from 215Y by a 1-nt change they may potentially be prone to evolve rapidly to 215Y under drug-selective pressure. Therefore, it is important to assess the impact of these intermediate mutations on the rate of acquisition of 215Y. A better understanding of the evolution of these viruses in the presence of AZT *in vitro* therefore might shed light on the clinical significance of these mutations and their potential impact on the response to AZT treatment.

### 1.1.3 HSV2 Virology

HSV-2, together with HSV-1 and the varicella-zoster virus (chickenpox), belongs to the subfamily Alphaherpesvirinae in the family Herpesviridae. These are large, complex enveloped viruses with an outer lipid envelope studded with at least 10 viral glycoproteins, an intermediate tegument layer comprising at least 15 viral proteins, and an icosahedral nucleocapsid containing the double-stranded DNA genome. The genome is organized into a 126-kb long and a 26-kb short region of double-stranded DNA bracketed by inverted repeat sequences that readily allow isomerization or recombination of the two regions. The genome comprises some 84 open reading frames. These have been divided into immediate-early genes, whose transcription depends on a virally-encoded activating protein, VP16, and which encode the viral  $\alpha$  proteins; the early genes, which are turned on by the  $\alpha$  proteins and whose products ( $\beta$  proteins) are involved in DNA replication; and the late genes, the products of which ( $\gamma$  proteins) are virion structural proteins and proteins needed for virus particle assembly and egress. Some of the viral envelope glycoproteins (gD) are antigenically related to those of HSV-1, whereas most are type-specific (particularly gG1 and gG2). Numerous viral gene products, which are dispensable for virus growth *in vitro*, can be considered as virulence genes that are involved in preventing apoptosis in the infected host cell, blocking the induction of interferons, or downregulating the presentation of viral antigens in the context of class I histocompatibility antigens (HLAs).

When the latent state is established in the neural ganglia, transcription is severely restricted such that a single transcript is produced from the latency-associated transcript (LAT) promoter, and only a few viral proteins are made. At intervals, changes in neuronal physiology induced by trauma, hormones, stress or immune suppression, render the neurones permissive to virus replication, resulting in full transcription of the genome and a burst of progeny virions.

5.3.3. Vaccines The prospect for developing a vaccine against HSV-2 that



could provide sterilizing immunity is thought to be unrealistic. The goals of the vaccines under development are rather to prevent the establishment of latent infection by blocking access of the virus to sensory ganglia, to reduce the severity of the symptoms, and/or to reduce the frequency of recurrences. The correlates of protective immunity against HSV-2 are not entirely understood. Passive maternal antibody seems important in preventing infection of the newborn and CD4<sup>+</sup> Th1 T-cells appear to be crucial to the immune response. IFN- $\gamma$  secretion and CD8<sup>+</sup> CTL may also play a major role, particularly in the prevention of recurrences.

HSV-2 subunit vaccines were developed based on the use of viral envelope glycoproteins.

- A two-component gB2 and gD2 recombinant glycoproteins subunit vaccine formulated in MF59 adjuvant was developed by Chiron. The 2-component vaccine induced high antibody titres and showed 26% efficacy in women for a period of six months but protection did not persist and male volunteers were not protected.
- GSK developed a single component gD2 vaccine formulated in AS04 adjuvant (alum + monophosphoryl lipid A). The gD2 vaccine induced good Th1 immunity in mice, including high IFN- $\gamma$  secretion, and provided good protection against vaginal HSV-2 challenge in female guinea pigs. The vaccine was tested in two large, double-blind, controlled Phase III trials on volunteers with a partner with genital herpes disease. In the first study, 847 subjects were selected as seronegative for both HSV-1 and HSV-2, whereas in the second study the 2491 selected subjects were selected only on the basis of HSV-2 seronegativity. The vaccine was 73% efficacious against genital herpes disease in doubly seronegative women. Trends towards protection against infection were also observed, but the figures were not statistically significant (less than 48% efficacy). Most unexpectedly, however, the vaccine was not effective in women previously seropositive for HSV-1 and in men, regardless of their HSV seropositivity status. This suggests that HSV-1 immunity is protective against HSV-2, but no satisfactory explanation is available of why subunit vaccines seem to provide only gender-specific protection. Further Phase III efficacy trials of the gD2 vaccine (Herpevac) are in progress in collaboration with the NIH, involving about 7500 persons from 18 to 30 years of age, double HSV-1/HSV-2 seronegative women. A vaccine that protects women could be expected to decrease the rate of neonatal HSV infection and have an impact on the epidemic spread of genital herpes. Lack of

efficacy of vaccines in HSV-1 infected individuals would however render the vaccine of little use in developing countries, where HSV-1 infection is ubiquitous.

- A novel, live attenuated HSV-2 candidate vaccine has been developed by Xenova/GSK using a replication-impaired virus mutant that lack the gene of the essential glycoprotein gH (ICP8 gene mutation) as a disabled infectious single cycle (DISC) virus vaccine. The vaccine was tested in Phase II trials in the USA as a therapeutic vaccine in HSV-2 seropositive symptomatic patients. It was well tolerated and induced neutralizing antibodies and CTL in 83% of the vaccinees, but no difference in time to recurrence and no difference in virus shedding were observed as compared with controls. The development of the DISC vaccine has been refocused towards its use as a prophylactic vaccine.
- Another live, replication-impaired vaccine is currently under development by Avant Immunotherapeutics. Other viral mutants that are defective for replication and impaired for establishment of latency, such as mutant d15–29, are at a preclinical stage of development.
- A live attenuated vaccine based on a replication-competent ICP10 mutant of HSV-2 developed by AuRix is in Phase II clinical study.

The impact of HSV-2 infection on the human immune system and its ability to counter other infections such as HIV-1 may be complex [8]. HSV-2 infection alters the function and phenotype of monocytes [9]–[11], impeding their maturation, which may in turn alter the phenotype and function of interacting CD4+ and CD8+ T cells. HSV-2 is a ligand for certain TLR molecules [15], and may mediate the effect on dendritic cells. HSV-2 may alter the innate immune system, with subsequent impacts on the adaptive immune system.

Individuals experiencing incident HSV-2 infections are at the greatest risk of HIV acquisition, compared with individuals not infected with HSV-2 or who have prevalent HSV-2 infection. The individuals with serologic evidence of recent incident HSV-2 infection had the highest HIV incidence, illustrating that recent infection with HSV-2 is independently associated with HIV acquisition.

Identifying risk factors for HIV and HSV-2 co-infection provides the opportunity to develop health promotion programmes based on primary prevention. Therefore, we sought to determine the risk factors for HIV and HSV-2 co-infections in a group of fishermen along

Lake Victoria in Kenya, where research has shown HSV-2 to be high among fishermen population [7].

## **1.2 OBJECTIVES OF THE STUDY**

### **1.2.1 Primary objective**

To establish risk factors for the presence of HIV and HSV-2 Co-Infection Among Fishermen along Lake Victoria, Kenya.

### **1.2.2 Secondary objectives**

- To predict HIV and HSV-2 Co-Infection Probabilities among Fishermen along Lake Victoria, Kenya.
- To predict HIV and HSV-2 Infection Probabilities among Fishermen along Lake Victoria, Kenya.

## **1.3 SIGNIFICANCE OF THE STUDY**

Opportunistic Infections and co-infections are the major cause of deaths amongst HIV infected individuals and this mostly depends upon the risk factors, type of exposure and geographic region. The commonest types of infections reported are tuberculosis, chronic diarrhoea, oral candidiasis, HSV-2, cytomegalovirus, hepatitis B virus and hepatitis C virus. Worldwide, HSV-2 is a common infection in adults [12], and among persons with HIV-1 or at risk for HIV-1 infection. Because of frequent co-infection of HIV-1 and HSV-2, HSV-2 is now also considered a serious public health threat especially in sub Saharan Africa where its prevalence may be greater than 50%. HSV-2 control strategy (through drugs or vaccine when available) may reduce HIV-1 transmission, the potential implications in public health. Information on risk factors of HIV-1 and HSV-2 co-infection is essential for program interventions may require specific approaches to curb the problem.

## **1.4 LITERATURE REVIEW**

[23] conducted a cross-sectional analysis of risk factors for HSV-2 seropositivity in >3300 HIV-1 uninfected members of African HIV-1 serodiscordant couples in which the HIV-1 infected partner was dually-infected with HSV-2. The aim of the study was to assess whether HIV-1 plasma RNA level or CD4 count in the HIV-1 infected partner, as indicators of higher

HIV-1 infectiousness and more advanced immunosuppression, modifies HSV-2 infectiousness, and if male circumcision in the HIV-uninfected partners of women with HSV-2/HIV-1 dual infection protected them from HSV-2 acquisition. Univariate and multivariate analyses were performed using Poisson regression with HSV-2 status of the partner participant as the outcome variable. In order to examine male circumcision and to allow for other possible gender differences in transmission risk, separate analyses were performed for male and female HIV-1 infected participants. Backward elimination, starting with all variables significant at  $p < 0.1$  in univariate regression, was performed to select a final multivariate model. Results showed that among male HIV-1 uninfected partners of HIV-1/HSV-2 dually infected females, older age (adjusted prevalence ratio [aPR] 1.11,  $p < 0.001$ ), a greater number of children (aPR 1.02 per 1 child increase,  $p < 0.001$ ) and greater number of years living together (aPR 1.07, per 1 year increase,  $p = 0.009$ ) were significantly associated with HSV-2 infection. Characteristics of the female HIV-1 infected partner associated with HSV-2 infection in their male HIV-1 uninfected partners included having herpetic lesions observed on genital exam at enrollment (aPR 0.92,  $p = 0.04$ ), other sexual partners (aPR 1.28,  $p = 0.05$ ), and using vaginal drying agents (aPR 1.08,  $p = 0.05$ ). Circumcised HIV-1 uninfected men were at slightly decreased risk for HSV-2 (aPR 0.94,  $p = 0.06$ ), although this did not achieve statistical significance. Among female HIV-1 uninfected partners of HIV-1/HSV-2 dually-infected males, only older age (aPR 1.08,  $p < 0.0001$ ) and greater number of years of education (aPR 0.90,  $p = 0.001$ ) of the HIV-1 infected partner were associated with HSV-2 seropositivity.

In a study where participants were female sex workers who had previously enrolled in an HIV prevention study in rural Zimbabwe, the relationship between HSV-2 and HIV-1 infection and in particular whether genital shedding of HSV-2 had any effect on genital shedding of HIV-1 was carried out by [31]. The specific objectives were to determine: (1) the prevalence of HSV-2 among sex workers in rural Zimbabwe; (2) the extent of recognized symptomatic genital herpes; (3) the extent of genital HSV-2 shedding; and (4) the extent of genital HIV-1 shedding and its relationship to HSV-2 status. Factors associated with HIV-1 and HSV-2 co-infection were examined on univariate analyses. Factors associated with HIV-1 and HSV-2 genital shedding were examined on both univariate and multivariate analyses. Categorical data were analysed using chi-squared tests or Fisher's exact test where appropriate, with multivariate analysis by logistic regression. Continuous data were analysed using t-tests if distributed normally. Continuous data such as age, time as a sex worker and number of

partners, were categorized so that a roughly equal number of individuals would be in each category. Categories for CD4 cell count, PVL and CVL viral load were defined based on clinically accepted categories. All variables that were statistically significant ( $P < 0.15$ ) on univariate analysis were included in the multivariate analyses, as were variables in which prior knowledge suggested that it would be appropriate to include these factors in the model. Confidence limits for this correlation coefficient were calculated using Fisher's transformation. Linear regression was used to analyse the determinants of genital HIV-1 shedding. 124 women were co-infected and 58 women were infected with HSV-2 alone. Co-infected women were more likely to be younger, unmarried, to have had more sexual partners, and to have infection with *Trichomonas vaginalis* than those with HSV-2 alone ( $P < 0.05$  for each variable). They also appeared more likely to report having genital herpes than those with HSV-2 infection alone ( $P = 0.093$ ).

In a study to measure the effect of HSV-2 seropositivity and HSV-2 genital shedding on the risk of perinatal HIV transmission among HIV positive pregnant women by [24], risk factors for HIV transmission were measured using the  $\chi^2$  test, *t*-test, and Wilcoxon rank sum test. Adjusted odds ratios (ORs) with 95% confidence limits (CLs) for HSV-2 transmission were calculated using logistic regression. HSV-2 seropositivity and HSV-2 genital shedding both were evaluated as risk factors for HIV transmission, and known predictors of transmission, including HIV viral load, were included in the multivariable analysis. The findings were that HSV-2 seropositivity increases the risk of perinatal HIV transmission, and HSV-2 genital shedding increases the risk of intrapartum HIV transmission.

In a longitudinal study, to estimate risk of HSV-2 infection in women with bacteria vaginosis, involving 1207 women, aged 18-30 years carried out by [25], factors associated with HSV-2 antibody seroprevalence included the presence of Group B *Streptococcus* on vaginal culture, a history of intercourse with an uncircumcised male, and the diagnosis of BV. Previously identified factors that were confirmed by the study included black race, older age, and the use of douching products. During follow-up 32 women turned HSV-2 positive and factors associated with acquisition of HSV-2 included having a high school education or less (HR 3.0; 95% CI 1.5-6.2), having a new sexual partner (HR 2.8; 95% CI 1.3-5.9), and having a diagnosis of BV during the interval prior to seroconversion (HR 2.2; 95% CI 1.0-4.6;  $P = .04$ ).

[26] carried out a study involving 186 antiretroviral-naive people whose HIV seroconversion date lay within the previous 170 days. At the time of HIV diagnosis, 101 people (54%) had a positive test for HSV-2. Through 24 months of follow-up, CD4 counts drifted downward in people without HSV-2 but remained stable in those with HSV-2.

[13] used longitudinal mixed effects models to assess the relationship of co-infection status on CD4+ T cell counts and HIV-1 RNA levels, with random effects for time and intercept in a study to examine the impact of HSV-2 status at the time of HIV-1 acquisition for its impact on subsequent clinical course, and total CD4+ T cell phenotypes. Results showed that HIV-1/HSV-2 co-infected treatment naïve adults in early HIV-1 infection had higher CD4+ T cell counts than those infected with HIV-1 alone.

In a study to establish a unique HSV-2 macaque model to facilitate research to define how HSV-2 increases HIV transmission, and enable more rigorous evaluation of candidate antiviral approaches by [14], it was noted that modeling is of fundamental importance to better understand the interplay of HSV-2 and HIV, and provides a useful means of testing the efficacy of preventative and therapeutic strategies.

[16] presented study results at the 13th Conference on Retroviruses and Opportunistic Infections. In the trial, 140 women from Burkino Faso with HIV/HSV-2 coinfection were randomized to valacyclovir (VACV) suppressive therapy (n = 70) or placebo (n = 70). Valacyclovir (1.0 g) was given daily for 3 months. None of the women were eligible for highly active antiretroviral therapy. The mean CD4 count was 519 cells/uL in the VACV group and 482 cells/uL in the placebo group. The most important finding of the study was that it demonstrated for the first time that HSV-2 increases HIV-1 replication.

[17] carried out a study to provide a clearer picture regarding infections occurring amongst HIV seropositive individuals so that the scientific findings could be translated into sustainable prevention programmes and improved public health policies. Analysis of the different spectrum of OIs/Co-infections were carried out with 204 HIV sero-positive patients (142 males and 62 females) who visited the HIV/AIDS Apex Clinic in a tertiary care hospital from March 2006 to March 2009. Statistical analysis was performed using student's t test. The Null Hypothesis was also tested and a P value < 0.05 was considered to be statistically significant. Mean, median and mode were also estimated. The common co-infections/opportunistic infections were Oral Candidiasis (53.43%), Chronic Diarrhoea (47.05%), HSV-2 (36.76%),

Tuberculosis (35.29%), Cytomegalovirus (26.96%), Hepatitis B Virus (15.19%) and Hepatitis C Virus (7.35%). Dual infections, like HSV-2 & Cytomegalovirus (15.38%), HSV-2 & Tuberculosis (14.61%), HSV-2 & oral candidiasis (24.61%) and CMV & oral candidiasis (14.61%) were significant in follow-up patients.

In discussing the impact of prevalent and incident HSV-2 infection upon the acquisition of HIV using data published in the *Journal of Infectious Diseases*, involving the search for HSV-2 antibodies in stored serum samples from a cohort of 2,732 HIV-negative patients attending four clinics in Pune, India [50], results were as follows: Of the 2,732 individuals enrolled, 2,260 were male, 463 were female, and 9 were eunuchs. The prevalence of HSV-2 at enrollment was 43%. The HSV-2 incidence was 11.4 per 100 person-years, and the HIV incidence was 5.9 cases per 100 person-years. The HIV incidence was 3.6 per 100 person-years among persons without evidence of HSV-2 infection, 7.5 per 100 person-years among persons with prevalent or remote incident HSV-2 infection, and 22.6 per 100 person-years among persons with recent incident HSV-2 infection. Using a proportional hazards model, the investigators found that the presence of asymptomatic prevalent HSV-2 infection conferred an adjusted hazard ratio for HIV infection of 2.14 (compared with no genital ulceration and negative results of serologic testing for HSV-2). Symptomatic prevalent HSV-2 infection conferred an adjusted hazard ratio of 5.06. This study demonstrated that individuals experiencing incident HSV-2 infections are at the greatest risk of HIV acquisition, compared with individuals not infected with HSV-2 or who have prevalent HSV-2 infection. The individuals with serologic evidence of recent incident HSV-2 infection had the highest HIV incidence, illustrating that recent infection with HSV-2 is independently associated with HIV acquisition.

In an article to review various strands of epidemiological evidence linking HSV-2 and HIV included a consideration of the similarity of the sexual risk factors and behaviours associated with acquisition of these two infections, i.e. younger age at coitarche, higher number of sexual partners, women at greater risk than men, homosexual males at greater risk than heterosexual males, previous sexually transmitted infections (STIs), a greater number of years of sexual activity and contact with female sex workers, studies looking at the prevalence and incidence of these infections in the general population and also in populations at increased risk for HIV acquisition, studies showing that HSV-2 is acquired before HIV and finally that individuals with pre-existing HSV-2 are more likely to acquire HIV and that the prevalence of HSV-2 infection in the general population has a major impact on the sexual transmission of HIV. Cohort and nested



case-control studies provided information about pre-existing HSV-2 and HIV acquisition and the relative risk (RR) ratio was 2.1 (95% confidence interval, 1.4-3.2). By using this estimate, it has been calculated that in HSV-2-positive individuals, 52% of sexually transmitted risk can be attributed to HSV-2. In addition, the calculated population-attributable risk percentage (also known as the aetiological fraction) varied with the HSV-2 seroprevalence in the population. In populations where HSV-2 prevalence is 80% or more, almost half of the sexually acquired HIV can be attributed to HSV-2. [38]

[27] in a study to examine the risk factors of malnutrition among children whose mothers are infected with HIV in sub-Saharan Africa applied Multilevel logistic regression models to Demographic and Health Survey data collected during 2003–2008 from 18 countries in sub-Saharan Africa, where the DHS Demographic and Health Survey included HIV test data for adults of reproductive age. Results showed that the risk of malnutrition among children whose mothers are infected with HIV is particularly high among children aged one, boys, multiple/twin births, those who were smaller than average at birth, or whose mothers had no education, or in poorest or single parent households.

[32] directly studied the influence of HSV infection on HIV-1 replication in vivo by administering chronic daily therapy with acyclovir to HIV/HSV-2–coinfecting persons and measuring plasma HIV-1 RNA levels before and after administration of acyclovir. Acyclovir reduced plasma RNA levels by an average of one-third of a log; a reduction in plasma HIV RNA levels was observed in 11 of 12 persons, and HIV-1 RNA levels returned to previous baseline upon discontinuation of therapy. These studies may provide some explanation of the older studies with zidovudine monotherapy that showed increased survival with concomitant acyclovir use.

HIV-1 infection appears to be fueling the HSV-2 epidemic. Studies of male factory workers in Zimbabwe and men and women in rural Rakai have shown markedly higher acquisition rates for HSV-2 in HIV-1–seropositive compared with HIV-1–seronegative persons, with relative risks of 4.7 and 3.7, respectively [33,34]. HIV-1–infected persons appear to have an increased risk of acquisition of HSV-2, although it is not known whether these observations represent increased susceptibility to HSV-2 infection or are a marker for sexual exposure to HIV-1 and HSV-2 coinfecting persons who shed HSV from mucosal surfaces more frequently than HIV-1 seronegative persons with HSV-2 infection.



The data on the mucosal interactions of HIV and HSV-2 suggest that HIV-1-seropositive, HSV-2-seropositive persons may transmit HIV infection more frequently than HIV-1-seropositive persons who are HSV-2 seronegative. A study of HIV-discordant couples in Uganda (2001) found similar probabilities of HIV transmission irrespective of whether the HIV-positive partner was HSV-2 seropositive or seronegative [35]. However, the high prevalence of HSV-2 antibodies in HIV-positive persons (85%) limited the power to detect an effect. It is noteworthy that the presence of recent symptomatic genital ulceration in the HIV-positive source partner significantly increased the transmission probability per act (0.0041) when compared with no ulceration (0.0011) [36]. Because HSV-2 is the most common cause of genital ulceration in this population, it suggested that HSV-2 may enhance transmission of HIV from symptomatic dually infected persons.

[37] carried out a cross sectional study of sample size 2000 to estimate age and sex specific HSV-2 prevalence in urban African adult populations and to identify factors associated with infection in Contonuo (Benin), Yaounde (Cameroon), Kisumu and Ndola (Zambia). Analysis of HSV-2 infection were stratified by city and by gender, as the effect of various exposures on risk of HSV-2 infection may be modified by both these factors. Logistic regression was used to calculate age-adjusted odds ratios and 95% confidence intervals for the association of HSV-2 with socia-demographic and sexual behaviour risk factors. Variables significant at the 95% significance level in the age adjusted analysis for any city were included in multivariate analysis, which again were stratified by city and gender. Statistical significance of odds ratios was assessed using the likelihood ratio test, and variables were retained in the model if they were statistically significant in any city. The association of HSV-2 infection with other STIs (HIV, syphilis, gonorrhoea, chlamydia and *T.vaginallis* and history of genital pain/sores in the past 12 months) were analyzed by including these variables in the final multivariate model. The results showed that HSV-2 prevalence was significantly associated with older age, ever being married and number of life time sexual partners, in almost all cities and both sexes. There was also a strong, consistent association with HIV infection. Among women, the adjusted odds ratios for the association between HSV-2 and HIV infections ranged from 4.0 (95% CI=2.0-8.0) in Kisumu to 5.5 (95 CI=1.7-1.8) in Yaounde, and those among men ranged from 4.6 (95% CI=2.7-7.7) in Ndola to 7.9 (95% CI=4.1-15) in Kisumu.

[51] used Multiple Logistic Regression to evaluate risk factors associated with anemia and Iron Deficiency in a sample of children participating in or applying for the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC). Maternal WIC participation during pregnancy, child age, and the intake of  $\geq 125$  mL orange or tomato juice/d were negatively associated with Iron Deficiency, and being male and living in an urban location were positively associated with Iron Deficiency.

[52] used univariate and multiple logistic regressions to investigate the frequency of reoperation and factors that might influence its occurrence, in a retrospective, follow up cohort study of all children below years, who underwent laparoscopic Nissen fundoplication (LNF) or open Nissen fundoplication (ONF) from January 1, 1997 to December 31, 2002 at Pediatric Gastroenterology and Nutrition, Emory University School of Medicine, Atlanta, Georgia. Univariate Logistic regression showed that initial laparoscopic surgery, a history of prematurity, and reflux alone tended to be associated with increased risk of reoperation. In multiple logistic regression the risk for reoperation showed to be higher in patients who underwent initial LNF and children with history of prematurity; however, the independent impact of these two risk factors did not reach statistical significance.

[53] used multiple logistic regressions to estimate the combined effects of three variables (AAP, IAP, and vector number) on plant infection status, including interactions between them in a study to describe the transmission ecology of the bacterium *Xylella fastidiosa* Wells et al., the causal agent of Pierce's disease in grapevines, by its leafhopper vectors.

[40] in a study to examine correlates of condom use among a national random probability sample of sexually experienced young adults aged 15 to 24 years (n=7686) in South Africa, used multiple logistic model to examine predictors of condom use by gender. Results showed that those who had used a condom at their sexual debut were more likely to have used a condom during their most recent sexual intercourse as those who had not used a condom at their sexual debut.

[39], in a cross-sectional study in 2004 in Xi'an City, China to explore the sociodemographic and environmental factors at community, school, and household levels associated with physical inactivity, hierarchical multiple logistic models were developed based on a conceptual framework of factors related to physical activity. It was found that gender, age of adolescent, Paternal

education, lack of extracurricular exercise and fewer sports meetings to be associated with physical inactivity.

[41] used a multiple logistic regression model for Predicting the development of *Phytophthora ramorum* symptoms in Tanoak (*Lithocarpus densiflorus*) California.

[42] used multiple logistic model in a study to describe oncology nurses' use of National Comprehensive Cancer Network (NCCN) clinical practice guidelines for chemotherapy-induced neutropenia (CIN) and febrile neutropenia (FN). The model showed factors related to oncology nurse participants' use of NCCN clinical practice guidelines for CIN and FN. The professional characteristic shown to predict use of the NCCN clinical practice guidelines was holding an advanced oncology nurse certification versus generalist or no certification.

The multiple logistic model method has potential drawbacks which are not generally recognized [43]. Specifically the so-called main effect logistic model assumes that the probability of developing disease is linearly and additively related to the risk factors on the logistic scale. This assumption stipulates that for each risk factor, the odds ratio is constant over all reference exposure levels, and that the odds ratio exposed to two or more factors is equal to the product of individual risk factor odds ratios. If the observed odds ratios in the data follow this pattern, the model-predicted odds ratios will be accurate, and the meaning of the odds ratio for each risk factor will be straightforward. But if the observed odds ratios deviate from the model assumption, the model will not fit the data accurately, and the model-predicted odds ratios will not reflect those in the data. Although satisfactory fit can always be achieved by adding to the model polynomial and product terms derived from the original risk factors, the odds ratios estimated by such an interaction logistic model are difficult to interpret, viz., the odds ratio for each risk factor depends not only on the reference exposure levels of that factor, but also on the exposure level in other factors. Empirical evidence suggests that the actual relationship between risk factors and disease is likely to be nonlinear and nonadditive on the logistic scale. In most instances, an interaction logistic model is required to analyse epidemiological data satisfactorily.

Multiple logistic function and Mantel-Haenszel procedure are capable of controlling confounding in a large set of data.

Logistic regression is applicable to a broader range of research situations than discriminant analysis.

## **1.5 DATA/DESIGN OF THE STUDY**

The data being used for this project is obtained from a cross sectional study of 168 fishermen screened for participation in a male microbicide study carried out in Kisumu in the year 2007/2008. This study was carried out by Research Care and Training Program (RCTP) under Kenya Medical Research Institute (KEMRI). Two-stage cluster sampling to select 168 fishermen for the structured interviews and STI testing was used. Subjects were tested for HIV, Syphilis and HSV-2 infections. In addition, the demographic and behavioural characteristics of the subjects were assessed through face to face interviews. The prevalence of HIV was tested using rapid methods and cases confirmed by ELISA while Syphilis was tested using the rapid plasma reagin (RPR) card test. Infection with Herpes simplex virus type 2 (HSV-2) was tested using ELISA. Some subjects had inconclusive(equivocal) results for HSV-2 and a further test was required to determine HSV-2 status (positive or negative) which could not be done due to limited funds. As a result, cases with equivocal HSV-2 result have been deleted from this project analysis. The cases considered for this project analysis are therefore 126 in total, which had either HSV-2 positive or negative results. All variables being used in this project for analysis have been answered for all cases ( meaning the data doesnt have missing entries).

## CHAPTER 2.0 METHODOLOGY

### 2.1 Theoretical Model

Multiple Logistic regression is a generalized linear model (GLM) used for binomial regression. The GLM is an extension of the General Linear Model to include response variables that follow any probability distribution in the exponential family of distributions such as the Binomial. Hypothesis tests applied to the GLM do not require normality of the response variable, nor do they require homogeneity of variances. Hence, GLMs can be used when response variables follow distributions other than the Normal distribution, and when variances are not constant. Parameter estimates are obtained using the principle of maximum likelihood; therefore hypothesis tests are based on comparisons of likelihoods.

Multiple Logistic regression is an accepted statistical method for assessing association between an antecedent characteristic (risk factor) and a quantal outcome (probability of disease occurrence), statistically adjusting for potential confounding effects of other covariates. It is also useful for situations in which you want to be able to predict the presence or absence of a characteristic or outcome based on values of a set of predictor variables. In addition it can be used to estimate odds ratios for each of the independent variables in the model.

This model has been used to analyze the risk factors associated with HIV and HSV-2 co-infection.

The general multiple logistic regression model for  $p$  predictors is as follows:

$$\text{logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

which presents a relationship between the natural logarithm of the odds ratio of coinfection with HIV-1 and HSV-2. The response variable of interest,  $Y_i$ , is a dichotomous variable taking value 1 if a fisherman is coinfecting and 0 otherwise, and  $p$  is the probability that the  $i$ th individual is HIV-1 and HSV-2 co-infected. Here,  $\beta_0, \beta_1, \dots, \beta_p$  are regression parameter estimates, where  $\beta_i$  refers to the effect of  $x_i$  on the odds that the individual is co-infected, controlling the other covariates, that is at fixed levels of the other predictor variables.  $\beta_1$  represents also the change in the odds of an outcome for an increase in one unit of  $x_i$ .

The parameter estimate can be conceptualized as how much mathematical impact a unit changes in the value of the independent variable has on increasing or decreasing the probability that the dependent variable will achieve the value of one in the population from which the data are assumed to have been randomly sampled.

Exponentiation of the parameter estimates for the independent variable in the model by the number  $e$  yields the odds ratio, which is a more intuitive and easily understood way to capture the relationship between the independent and dependent variables. The odds ratio gives the increase or decrease in probability that a unit change in the independent variable has in the probability that the event of interest will occur. The parameter estimates and associated odds ratios are *point estimates* of the true value of these quantities in the population from which the data under analysis are assumed to have been randomly sampled.

The null hypothesis,  $H_0$  of main interest is that  $\beta_i$  equals zero, i.e there is no relationship between the binary response variable and the predictor variables.

95% confidence intervals (CI) for the parameters are also calculated from the product of the asymptotic standard error and standard normal distribution using the R command (not shown here). Both the standard errors and confidence intervals are considered approximate. If the 95% CI includes 1; then the odds ratio is not statistically significant. When the lower bound of the 95% CI is so close to 1, then p-value is very close to 0 .05. The width of a CI provides a measure of the precision with which inferences can be made. To calculate a CI, the sampling distribution of the estimator is required.

Parameters of a logistic response function are often estimated using the method of maximum likelihood (ML). One of the problems with ML estimation is that, no closed-form solution exists for the values of the parameters that maximize the log-likelihood function. Hence sophisticated computer-intensive numerical search procedures (i.e: Newton Raphson) are required to find ML estimates of parameters.

## 2.2 Parameter Estimation: Maximum Likelihood Estimation

The following steps are followed to estimate parameters from distributions in the exponential family. Consider independent random variables  $Y_1, \dots, Y_N$  satisfying the properties of a GLM. We wish to estimate parameters  $\beta$  which are related to the  $Y_i$ 's through  $E(Y_i) = \mu_i$  and  $g(\mu_i) = \mathbf{x}_i^T \beta$ .

For each  $Y_i$ , the log-likelihood function is

$$l_i = y_i b(\theta_i) + c(\theta_i) + d(y_i) \quad (2.1)$$

where  $b(\theta_i)$  is the natural parameter of the distribution.

Also

$$E(Y_i) = \mu_i = \frac{-c'(\theta_i)}{b'(\theta_i)} \quad (2.2)$$

$$\text{var}(Y_i) = \frac{b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)}{[b'(\theta_i)]^3} \quad (2.3)$$

and

$$g(\mu_i) = \mathbf{x}_i^T \beta = \eta_i \quad (2.4)$$

where  $\mathbf{x}_i$  is a vector with elements  $\mathbf{x}_{ij}, j = 1, \dots, p$ .

The log-likelihood function for all the  $Y_i$ 's is

$$l = \sum_{i=1}^N l_i = \sum_{i=1}^N y_i b(\theta_i) + \sum_{i=1}^N c(\theta_i) + \sum_{i=1}^N d(y_i)$$

To obtain the maximum likelihood estimator for the parameter  $\beta_i$ , we need

$$\frac{\partial l}{\partial \beta_i} = U_i = \sum_{i=1}^N \frac{\partial l_i}{\partial \beta_i} = \sum_{i=1}^N \left[ \frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \beta_i} \right], \quad (2.5)$$

which follows on using the chain rule for differentiation.

We now consider each term on the right hand side of this equation separately.

First,

$$\frac{\partial l_i}{\partial \theta_i} = y_i b'(\theta_i) + c'(\theta_i) = b'(\theta_i)(y_i - \mu_i),$$

by differentiating (2.1) and substituting (2.2). Next

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\frac{\partial \mu_i}{\partial \theta_i}}$$

Differentiation of (2.2) gives

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{-c'(\theta_i)}{b'(\theta_i)} + \frac{c'(\theta_i)b''(\theta_i)}{[b'(\theta_i)]^2} = b'(\theta_i) \text{var}(Y_i)$$

from (2.3).

Finally, from (2.4)

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij}$$

Hence the score, given in (2.5), is

$$U_i = \sum_{j=1}^N \left[ \frac{(y_i - \mu_i)}{\text{var}(Y_i)} x_{ij} \left( \frac{\partial \mu_i}{\partial \eta_i} \right) \right] \tag{2.6}$$

The variance-covariance matrix of the  $U_j$ 's has terms

$$\mathfrak{I}_{jk} = E[U_j U_k]$$

which form the **information matrix**  $\mathfrak{I}$ . From (2.6)

$$\begin{aligned} \mathfrak{I}_{jk} &= E \left\{ \sum_{i=1}^N \left[ \frac{(Y_i - \mu_i)}{\text{var}(Y_i)} x_{ij} \left( \frac{\partial \mu_i}{\partial \eta_i} \right) \right] \sum_{i=1}^N \left[ \frac{(Y_i - \mu_i)}{\text{var}(Y_i)} x_{ik} \left( \frac{\partial \mu_i}{\partial \eta_i} \right) \right] \right\} \\ &= \sum_{i=1}^N \frac{E[(Y_i - \mu_i)^2] x_{ij} x_{ik} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2}{[\text{var}(Y_i)]^2} \end{aligned} \tag{2.7}$$

because  $E[(Y_i - \mu_i)(Y_l - \mu_l)] = 0$  for  $i \neq l$  as the  $Y_i$ 's are independent. Using

$E[(Y_i - \mu_i)^2] = \text{var}(Y_i)$ , (2.7) can be simplified to

$$\mathfrak{I}_{jk} = \sum_{i=1}^N \frac{x_{ij} x_{ik}}{\text{var}(Y_i)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2 \tag{2.8}$$



The estimating equation  $\theta^{(m)} = \theta^{(m-1)} + \frac{U^{(m-1)}}{\mathfrak{I}^{(m-1)}}$  for the method of scoring generalizes to

$$\mathbf{b}^{(m)} = \mathbf{b}^{(m-1)} + [\mathfrak{I}^{(m-1)}]^{-1} \mathbf{U}^{(m-1)} \tag{2.9}$$

where  $\mathbf{b}^{(m)}$  is the vector of estimates of the parameters  $\beta_1, \dots, \beta_p$  at the  $m$ th iteration. In equation (i),  $[\mathfrak{I}^{(m-1)}]^{-1}$  is the inverse of the information matrix with elements  $\mathfrak{I}_{jk}$  given by (h) and  $\mathbf{U}^{(m-1)}$  is the vector of elements given by (f), all evaluated at  $\mathbf{b}^{(m-1)}$ . If both sides of equation (h) are multiplied by  $\mathfrak{I}^{(m-1)}$  we obtain

$$\mathfrak{I}^{(m-1)}\mathbf{b}^{(m)} = \mathfrak{I}^{(m-1)}\mathbf{b}^{(m-1)} + \mathbf{U}^{(m-1)} \tag{2.10}$$

From (h)  $\mathfrak{I}$  can be written as

$$\mathfrak{I} = \mathbf{X}^T \mathbf{W} \mathbf{X},$$

where  $\mathbf{W}$  is the  $N \times N$  diagonal matrix with elements

$$w_{ii} = \frac{1}{\text{var}(Y_i)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2 \tag{2.11}$$

The expression on the right-hand side of (2.10) is the vector with elements

$$\sum_{k=1}^p \sum_{i=1}^N \frac{x_{ik} x_{ik}}{\text{var}(Y_i)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2 b_k^{(m-1)} + \sum_{i=1}^N \frac{(y_i - \mu_i) x_{ij}}{\text{var}(Y_i)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)$$

evaluated at  $\mathbf{b}^{(m-1)}$ ; this follows from equations (2.8) and (2.6). Thus the right-hand side of equation (2.10) can be written as  $\mathbf{X}^T \mathbf{W} \mathbf{z}$ , where  $\mathbf{z}$  has elements

$$z_i = \sum_{k=1}^p x_{ik} b_k^{(m-1)} + (y_i - \mu_i) \left( \frac{\partial \eta_i}{\partial \mu_i} \right) \tag{2.12}$$

with  $\mu_i$  and  $\frac{\partial \eta_i}{\partial \mu_i}$  evaluated at  $\mathbf{b}^{(m-1)}$

Hence the iterative equation (2.10), can be written as

$$\mathbf{X}^T \mathbf{W} \mathbf{X} \mathbf{b}^{(m)} = \mathbf{X}^T \mathbf{W} \mathbf{z} \tag{2.13}$$

This is the same form as the normal equations for a linear model obtained by weighted least squares, except that it has to be solved iteratively because, in general,  $\mathbf{z}$  and  $\mathbf{W}$  depend on  $\mathbf{b}$ . Thus for generalized linear models, maximum likelihood estimators are obtained by an **iterative weighted least squares** procedure (Chames et al., 1976).

Most statistical packages that include procedures for fitting generalized linear models have an efficient algorithm based on (m). They begin by using some initial approximation  $\mathbf{b}^{(0)}$  to evaluate  $\mathbf{z}$  and  $\mathbf{W}$ , then (m) is solved to give  $\mathbf{b}^{(1)}$  which in turn is used to obtain better approximations for  $\mathbf{z}$  and  $\mathbf{W}$ , and so on until adequate convergence is achieved. When the difference between successive approximations  $\mathbf{b}^{(m-1)}$  and  $\mathbf{b}^{(m)}$  is sufficiently small,  $\mathbf{b}^{(m)}$  is taken as the maximum likelihood estimate.

### 2.3 Model Selection

Model selection for multiple logistic regression faces the same issues as for ordinary regression for normal data. The selection process becomes harder as the number of explanatory variables increases, because of the rapid increase in possible effects and interactions. There are two competing goals. The model should be complex enough to fit the data well. On the other hand, it should be simple to interpret, smoothing rather than overfitting the data. In an attempt to balance the goals, saturated models were fitted and the insignificant covariates and interactions were removed at 5% level of significant. A related aim is to find the best model, one that provides the maximum fit for the fewest predictors. The criteria for assessing different models include information criteria like Akaike's. The Akaike Information Criterion (AIC) adjusts the  $G^2$  (deviance) for a given model for the number of predictor variables. It is expressed as

$$AIC = G^2 - n + 2p,$$

where  $n$  is the number of observations and  $p$  is the number of predictors. Models with low AICs are the best fit and if many models have similarly low AICs, the one with the fewest model terms is chosen.

The model achieved is the tentative model for analyzing HIV-1 and HSV2 co-infection. Hosmer and Lemeshow recommend including any covariate which has P-value up to 0.25 in multivariate analysis .

Logistic regression coefficients can be used to estimate odds ratios for each of the independent variables in the model.

Our analyses of the Fishermen data will focus on establishing relationships between the binary fishermen outcome co-infection (measured by the variable co-infected with "1" indicating co-infection with HIV-1 and HSV-2 and "0" not co-infected) and the following fishermen characteristics that might have affected the chances of co-infection, namely Age, number of lifetime sexual partners, marital status, infection with an STI before, condom use, alcohol intake, religion, level of education, time away from home and wealth. We begin by using simple descriptive tools to provide initial insights into the structure of the data. In our case we have used frequencies and cross tabulations to look at associations between categorical explanatory variables and fishermen co-infection with HIV-1 and HSV-2.

Once we are satisfied with the model, it can be used for prediction of HIV-1 and HSV-2 co-infection seroprevalence as follows

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}$$

All statistical analyses have been performed using SPSS version 18.0 in running summaries and descriptives and R in running all the logistic models.

### 2.3.1 Study Variables

Table 1. Description of Study Variables

Variable Name	Description of the Variable Name
beach	Beach lived in the 4th quarter of the last 12 mnths
hivs	Currently infected with HIV
hsv2s	Currently HSV2 Infection
coinfected	Currently HIV and HSV-2 coinfecting
syphs	Currents SYPHILLIS Infection
anysti	Currently infected with either HIV, HSV-2 or Syphilis
age	Age in 5 year age categories
religion	Religion
marital	Marital status
education	Level of education
yearsch	How many years of school have you completed?
awayfhome	Time way from home
liveaway	Are there times you live alone away from your wife/wives or your regular partner(s)?
monwife	If yes, months per year you live together with wife
monregpt	If no, months per year you live together with regular partner
nchild	how many children do you have?
childcat	Category of Childre
totearn	Total Earnings per month
earning	Total earnings
mobile	Do you have a mobile phone?

fishboat	Do you have your own fishing boat?
fishnet	do you have your own fishing net?
currlive	where do you currently live?
wkbeach4	In the fourth quarter of the last 12 months,in what beach were you working?
nspntawy	How many times in the last month have you travelled and spent away from your current house?
paysex	Have you ever paid someone, given gifts, fish, favours or the right to buy your fish in exchange for sex?
hpv	Diagnosed with human papilloma virus Before
herpes	Diagnosed with herpes Before
gc	Diagnosed with gonorrhoea Before
chlam	Diagnosed with chlamydia Before
warts	Diagnosed with genital warts Before
hiv	Diagnosed with HIV Before
syph	Diagnosed with Syphilis Before
none	Diagnosed with NONE Before
sexdebut	How old were you when you first had sexual intercourse?
debutcat	age category of sex debut
nsexptn	How many sexual partners have you had in the last one month?
pat1	last one month partners category
sexptltm	How many sexual partners have you had in your lifetime?
lifecat	Category of lifetime partners
lklyhiv	How likely is it that you will become infected with HIV/AIDS?
immbath	Immediately before you had sex did you bathe?
aftbath	Immediately after you had sex did you bathe?
tkalcobf	Did you drink alcohol before the last sexual encounter?
condus3	Did you put on a condom before you started having sex in the last sexual act?
likely	How likely do you think it is that this person had sex with other people during the past one month?
liklyinf	How likely do you think it is that this person had a sexually transmitted infection?

## CHAPTER 3.0 RESULTS

Overall 72(57.1%) fishermen were infected with HSV-2 (95% CI,48.4-65.9%) and 31(24.6%) were infected with HIV-1 (95% CI,17.5-31.7%), showing HSV-2 to be highly prevalent than HIV-1. A total of 28(22.2%) fishermen were co-infected with HIV-1 and HSV-2. Out of 31 HIV-1 infected fishermen, only three (9.7%) were not infected with HSV-2. *Table 2* shows the characteristics of Sociodemographic, Socioeconomic, Behavioral and Proximate determinants of the 126 fishermen. Eighty-one (64.3%) fishermen are married, 39(31%) single and 6(4.8%) divorced/separated/widowed. Majority of the fishermen 64(50.8%) are below 25yrs of age, 30(23.8%) are aged between 25-29yrs of age, 15(11.9%) are between 30-34yrs of age and 17(13.5%) are 35 and above years of age. Majority, 67(53.2%) are on high monthly income, 33(26.2%) are on medium income and 26(20.6%) are on low monthly income. Majority, 72(57.1%) of the fishermen had their first sexual intercourse at the age between 15-20yrs of age, 51(40.5%) at the age of below 15yrs and 3(2.4%) at the age of above 20 yrs of age. The last time they had sexual intercourse, most fishermen 112(88.9%) did not put on a condom before having sex, only 11.1% had put on a condom before having sex. Ninety-one (72.2%) have ever paid for sex while 35(27.8%) had not. None of the fishermen had been diagnosed with warts, HIV and HPV. Those ever diagnosed with HSV-2 were 8(6.3%), syphilis 26(20.6%), gonorrhea 29(23%) and chlamydia 10(7.9%). Majority, 75(59.5%) had  $\leq 1$  sexual partners for the last one month, 34(27%) had 2 sexual partners and 17(13.5%) had  $\geq 3$  sexual partners.

From bivariate cross-tabulations with HIV-1 and HSV-2 co-infection as a response variable, we see associations between age and HSV-2 ( $\chi^2(3)=25.418$ ,  $p<0.001$ ) but no association with HIV ( $\chi^2(3)=7.062$ ,  $p=0.070$ ) or coinfection( $\chi^2(3)=5.696$ , $p=0.127$ ). Age 25-29yrs (80%) and over 35yrs (88.2%) are more at risk of HSV-2 acquisition as compared to ages less than 25yrs (35.9%) and 30-34yrs (66.7%).

Strong correlations of all three, that is HIV ( $\chi^2(2)=17.797$ ,  $p<0.001$ ), HSV-2 ( $\chi^2(2)=31.071$ ,  $p<0.001$ ) and Co-infection ( $\chi^2(2)=14.593$ ,  $p=0.001$ ), with marital status are evident. Divorced/Separated or married fishermen are more at risk as shown on the table having higher percentages of infection as compared to single fishermen.

There is also association between HSV-2 and the number of children that a fisherman has ( $\chi^2(3)=22.893$ ,  $p<0.001$ ) but no association with HIV ( $\chi^2(3)=2.809$ ,  $p=0.422$ ) and Co-infection ( $\chi^2(3)=2.952$ ,  $p=0.399$ ) with HIV and HSV-2.

Having a fishing net or a fishing boat is associated with HSV-2 infection, ( $\chi^2(1)=6.129$ ,  $p=0.013$ ), ( $\chi^2(1)=8.312$ ,  $p=0.004$ ) respectively, but it is not associated with HIV infection and co-infection with HIV-1 and HSV-2.

The number of sexual partners that a fisherman had for the last one month is associated with HIV-1 infection ( $\chi^2(2)=6.262$ ,  $p=0.044$ ), but not with HSV-2 infection or co-infection with HIV and HSV-2, ( $\chi^2(2)=2.329$ ,  $p=0.312$ ) and ( $\chi^2(2)=4.897$ ,  $p=0.086$ ) respectively.

Strong correlations of all three, that is HIV ( $\chi^2(1)=14.961$ ,  $p<0.001$ ), HSV-2 ( $\chi^2(1)=20.181$ ,  $p<0.001$ ) and Co-infection ( $\chi^2(1)=16.898$ ,  $p<0.001$ ) with having been diagnosed with an STI before are evident. The STIs in question are human papilloma virus (HPV), hsv-2, Gonorrhea, Chlamydia, Warts, HIV and syphilis. None had ever been diagnosed with warts or HIV before.

Most co-infected fishermen are aged between 25-29yrs and 30-34yrs each at 33.3% prevalence, while those aged <25yrs and 35+ yrs have a prevalence of 14.1% and 23.5% respectively, ( $\chi^2(3)=5.696$ ,  $p=0.127$ ). Age at first sexual intercourse is not associated with HIV and HSV-2 co-infection ( $\chi^2(2)=0.908$ ,  $p=0.635$ ). Condom use is also not associated with co-infection ( $\chi^2(1)=0.006$ ,  $p=0.940$ ).

*See Table 3 for cross tabulations.*

Running univariate logistic regression with co-infection with HIV-1 and HSV-2 as a response variable and all the other variables as independent variables, ANOVA fit shows high significance with marital status ( $p<0.0015$ ), being diagnosed with gonorrhea previously ( $p=0.002$ ), Chlamydia ( $p=0.007$ ) and syphilis ( $p<0.0013$ ) before. Number of sexual partners for the last one month and alcohol use before sex are marginally significant,  $p=0.06$  and  $p=0.07$  respectively. Secondly, running univariate logistic regression with HIV-1 as a response variable and all the other variables as independent variables, ANOVA fit shows high significance with marital status ( $p<0.0015$ ), diagnosed with gonorrhea (0.006), syphilis( $p<0.0012$ ) and chlamydia ( $p=0.01$ ) before, and number of sexual partners one had in the last one month. Lastly, running univariate logistic regression with HSV-2 as a response

variable and all the other variables as independent variables. ANOVA fit shows high significance with age ( $p < 0.001006$ ), marital status ( $p < 0.0010001$ ), diagnosed with gonorrhoea ( $p < 0.0019$ ), chlamydia ( $p = 0.02$ ) and syphilis ( $p < 0.0011$ ) before, and the perception of infection with HIV/AIDS (0.03).

These significant variables further give the following results on logistic regression.

The risk of coinfection is lower amongst the married ( $OR = 0.18$ , 95%  $CI = [0.03, 1.09]$ ,  $p = 0.0624$ ) and also among the Single ( $OR = 0.03$ , 95%  $CI = [0.003, 0.25]$ ,  $p = 0.001$ ) as compared to the divorced/separated ( $OR = 2.000$ , 95%  $CI = [0.37, 10.92]$   $P = 0.42$ ). The risk of infection with HIV-1 is lower amongst the married ( $OR = 0.19$ , 95%  $CI = [0.03, 1.04]$ ,  $p = 0.06$ ) and also among the single ( $OR = 0.03$ , 95%  $CI = [0.003, 0.25]$ ,  $p = 0.001$ ) as compared to the divorced/separated.

The risk of coinfection among those previously diagnosed with Gonorrhoea is higher ( $OR = 4.44$ , 95%  $CI = [1.77, 11.10]$ ,  $p < 0.001$ ) than for those not. The risk of infection with HIV-1 among those previously diagnosed with Gonorrhoea is higher ( $OR = 3.57$ , 95%  $CI = [1.46, 8.71]$ ,  $p = 0.005$ ).

The risk of coinfection among those previously diagnosed with Chlamydia is higher ( $OR = 6.41$ , 95%  $CI = [1.67, 24.66]$ ,  $p < 0.007$ ) than for those not. The risk of infection with HIV-1 among those previously diagnosed with Chlamydia is higher ( $OR = 5.46$ , 95%  $CI = [1.43, 20.86]$ ,  $p = 0.92$ ).

The risk of co-infection among those previously diagnosed with syphilis is higher ( $OR = 5.67$ , 95%  $CI = [2.203, 14.57]$ ,  $p < 0.0013$ ) than those not. The risk of infection with HIV-1 among those previously diagnosed with syphilis is higher ( $OR = 5.70$ , 95%  $CI = [2.25, 14.45]$ ,  $p < 0.0012$ ).

The risk of co-infection among those who consumed alcohol before sex is higher ( $OR = 3.07$ , 95%  $CI = [0.97, 9.75]$ ,  $p = 0.06$ ) than those who did not.

The risk of infection with HIV-1 is lower amongst those having 2 sexual partners for the last one month ( $OR = 0.94$ , 95%  $CI = [0.30, 2.98]$ ,  $p = 0.92$ ) and higher amongst those having  $\geq 3$  sexual partners ( $OR = 0.22$ , 95%  $CI = [0.06, 0.79]$ ,  $p = 0.02$ ) as compared to those who had  $\leq 1$  partner.



The risk of infection with HSV-2 is lower among those aged 30-34yrs (OR=0.5,95% CI=[0.12,2.02],p=0.33) and also among those aged < 25yrs (OR=0.14,95% CI=[0.05,0.39],p<0.0012) and higher among those aged 35+yrs (OR=1.88,95% CI=[0.33,10.53],p=0.48) as compared to those aged 25-29yrs.

The risk of infection with HSV-2 is higher among those married (OR=1.43,95% CI=[0.24,8.38],p=0.69) and lower among those who are single (OR=0.13,95% CI=[0.02,0.83],p=0.03) as compare to those who are divorced/separated.

The risk of infection with HSV-2 is higher among those previously diagnosed with gonorrhea (OR=4.90,95% CI=[1.73,13.90],p=0.003), chlamydia (OR=7.57, 95% CI=[0.93,61.70],p=0.06) and syphilis (OR=7.98, 95% CI=[2.25,28.28],p=0.001) as compared to those not.

The risk of infection with HSV-2 among those who said its impossible to be infected with HIV-1 is higher (OR=4.67, 95% CI=[1.24,17.56],p=0.02) and lower among those saying its likely and unlikely (OR=0.82, 95% CI=[0.32,2.12],p=0.69) and (OR=0.67, 95% CI=[0.24,1.86],p=0.44).

In model selection of the best fitting model, we find that marital status, previously infected with gonorrhea or syphilis and number of sexual partners in the last one month are risk factors for HIV-1 and HSV-2 co-infection. This model has the least AIC 113.7. The risk of co-infection with HIV-1 and HSV-2 is lower among the married (OR=0.24,95% CI=[0.03,1.73],p=0.16) and also among those who are single (OR=0.04,95% CI=[0.003,0.47],p=0.01) as compared to those who are divorced/separated. The risk of co-infection is higher among those who were previously diagnosed with gonorrhea (OR=3.11,95% CI=[1.10,8.74],p=0.03) or syphilis (OR=5.16,95% CI=[1.78,14.93],p=0.002) as compared to those not. The risk of co-infection is higher among those who have  $\geq 3$  sexual partners for the last one month (OR=2.31,95% CI=[0.56,9.52],p=0.23) and lower among those who had 2 sexual partners (OR=0.36,95% CI=[0.08,1.54],p=0.17) as compared to those who had  $\leq 1$  sexual partner.

The best fitting model for HIV-1 infection has marital status, previously infection with Chlamydia and number of sexual partners for the last 1 month as the risk factors with AIC as 126.3. The risk of infection with HIV-1 is lower amongst the married (OR=0.10,95% CI=[0.01,0.95],p=0.04) and also among the single (OR=0.02, 95% CI=[0.002,0.29],p=0.003)

as compared to those who are divorced/separated. The risk of infection with HIV-1 is higher among those previously infected with chlamydia (OR=4.75, 95% CI=[1.10,20.51],p=0.03) than those not. The risk of infection with HIV-1 is higher among those with  $\geq 3$  sexual partners for the last one month (OR=1.37, 95% CI=[0.37,5.15],p=0.64) and lower among those with 2 sexual partners (OR=0.29,95% CI=[0.08,1.13],p=0.07) as compared to those who had  $\leq 1$  sexual partner.

The risk factors for HSV-2 infection are marital status, previous infection with gonorrhea or syphilis with AIC as 136.01. The risk of HSV-2 infection is lower amongst the married (OR=0.19 95% CI=[0.26,13.24],p=0.54) and also the single (OR=0.03 95% CI=[0.002,1.58],p=0.13) as compared to those who are divorced/separated. The risk of HSV-2 infection is higher among those previously infected with gonorrhea (OR=2.00,95% CI=[0.83,8.13],p=0.10) and lower among those previously infected with syphilis (OR=0.19,95% CI=[1.57,25.21],p=0.01) as compared to those not.

*See Table 5*

## 3.1 RESULT TABLES

Table 2. univariate descriptive analysis of the variables

Variable	N (%), median, range
<b>Marital status</b>	
Single(Never Married)	39 (31%)
Married	81 (64.3%)
Separated/divorced/Widowed	6(4.8%)
<b>Education level</b>	
5-8 yrs	65 (51.6%)
9-14 yrs	61 (48.4%)
<b>Age in 5 year category</b>	
< 25 yrs	64 (50.8%)
25-29yrs	30 (23.8%)
30-34yrs	15 (11.9%)
35 yrs +	17 (13.5%)
<b>Religion</b>	
Protestant	59 (46.8%)
Catholic	39 (31%)
Independent Churches	28 (22.2%)
<b>Children (Range)</b>	0 (Minimum) 15 (Maximum)
<b>Total Income Per Month</b>	
Low	26 (20.6%)
Medium	33 (26.2%)
High	67 (53.2%)
<b>Where do you Currently Live?</b>	
Rented House	46 (36.5%)
Ancestral Home	76 (60.3%)
A friend/Relative's House	3 (2.4%)
Own House Away from Ancestral Home	1(0.8%)
<b>Age at first Sexual Intercourse</b>	
<15 yrs	51 (40.5%)
15-20yrs	72 (57.1%)
>20 yrs	3 (2.4%)
<b>Ever Paid someone for Sex?</b>	
Yes	91 (72.2%)
No	35 (27.8%)
<b>How likely is it that you will become infected with HIV/AIDS?</b>	
Impossible	20 (15.9%)
Very Unlikely	20 (15.9%)
About 50% Chance	62 (49.2%)
Very Likely	24 (19%)

**Did you drink Alcohol before the last Sexual intercourse?**

Yes	14 (11.1%)
No	112 (88.9%)

**Did you put on a Condom before you started having sex during the last Sexual Intercourse?**

Yes	14 (11.1%)
No	112 (88.9%)

**How likely do you think it is that this person had a Sexually transmitted Infection?**

Very Unlikely	73 (57.9%)
Unlikely	40 (31.7%)
50% Chance	7 (5.6%)
Likely	5 (4%)
Very Likely	1 (0.8%)

**Diagonosed with the following STIs Before**

Human Papilloma Virus	None
HSV-2	8 (6.3%)
Gonorrhea	29 (23%)
Chlamydia	10 (7.9%)
Genital warts	None
HIV	None
Syphilis	26 (20.6%)
None	74 (58.7%)

**Currently Infected with the following STIs**

HIV	31 (24.6%)
HSV-2	72 (57.1%)
Syphilis	2 (1.6%)
None	50 (39.7%)

**Currently HIV and HSV-2 Co-infected**

Yes	28 (22.2%)
No	98 (77.8%)

**Time lived away from home**

No	55 (43.7%)
Less than 3months	10 (7.9%)
3-6 Months	14 (11.1%)
7 Months +	37 (29.4%)
Other	10 (7.9%)

**Do you have a mobile phone?**

Yes	62 (49.2%)
No	64 (50.8%)

**Do you have Your Own Fishing Boat?**

Yes	27 (21.4%)
No	99 (78.6%)

**Do you have Your Own Fishing Net?**

Yes	30 (23.8%)
No	96 (76.2%)

**Number of Sexual partners for the last one month?**

≤ 1 partner	75 (59.5%)
2 Partners	34 (27%)
≥ 3 Partners	17 (13.5%)

**Lifetime Sexual Partners**

≤ 5 Partners	43 (34.1%)
6-10 Partners	33 (26.2%)
11-15 Partners	13 (10.3%)
16-20 Partners	16 (12.7%)
>20 Partners	21 (16.7%)

*Table 3. Bivariate cross-tabulations with HIV, HSV-2 and Co-infection*

	Predictor	Response variable		
		HIV status (%)	HSV-2 status	coinfection
Sociodemographic	Age	$\chi^2(3)=7.062, p=0.070$	$\chi^2(3)=25.418, p<0.001$	$\chi^2(3)=5.696, p=0.127$
	< 25 yrs	10(15.6%)	23(35.9%)	9(14.1%)
	25-29 yrs	11(36.7%)	24(80%)	10(33.3%)
	30-34 yrs	6(40%)	10(66.7%)	5(33.3%)
	35 +yrs	4(23.5%)	15(88.2%)	4(23.5%)
Religion	Religion	$\chi^2(2)=0.311, p=0.856$	$\chi^2(2)=0.403, p=0.818$	$\chi^2(2)=0.840, p=0.657$
	Catholic	9(23.1%)	23(59%)	8(20.5%)
	Protestant	14(23.7%)	32(54.2%)	12(20.3%)
	Independent	8(28.6%)	17(60.7%)	8(28.6%)
Marital status	Marital status	$\chi^2(2)=17.797, p<0.001$	$\chi^2(2)=31.071, p<0.001$	$\chi^2(2)=14.593, p=0.001$
	Single	3(7.7%)	8(20.5%)	2(5.1%)
	Married	23(28.4%)	60(74.1%)	22(27.2%)
	Separated	5(83.3%)	4(66.7%)	4(66.7%)
Children	Children	$\chi^2(3)=2.809, p=0.422$	$\chi^2(3)=22.893, p<0.001$	$\chi^2(3)=2.952, p=0.399$
	None	5(15.2%)	9(27.3%)	4(12.1%)
	1-3 Children	19(30.2%)	37(58.7%)	17(27%)
	4-6 Children	4(21.1%)	16(84.2%)	4(21.1%)
	> 7 Children	3 (27.3%)	10 (90.9%)	3(27.3%)
Region of Residence	Region of Residence	$\chi^2(3)=1.672, p=0.643$	$\chi^2(3)=2.791, p=0.425$	$\chi^2(3)=1.601, p=0.659$
	Rented House	13(28.3%)	29(63%)	12(26.1%)
	Ancestral Home	18(23.7%)	42(55.3%)	16(21.1%)

HIV&HSV-2 Co-Infection Risk Factors

	Relative/Friend House	0(0%)	1(33.3%)	0(0%)
	Own House	0(0%)	0(0%)	0(0%)
<b>Socioeconomic</b>				
	Education	$\chi^2(1)=2.752, p=0.097$	$\chi^2(1)=0.596, p=0.440$	$\chi^2(1)=1.201, p=0.273$
	5-8 yrs	20(30.8%)	35(53.8%)	17(26.2%)
	9-14yrs	11(18%)	37(60.7%)	11(18%)
	Wealth index	$\chi^2(2)=2.509, p=0.285$	$\chi^2(2)=0.723, p=0.697$	$\chi^2(2)=1.898, p=0.387$
	Low	9(34.6%)	13(50%)	8(30.8%)
	Medium	9(27.3%)	19(57.6%)	8(24.2%)
	High	13(19.4%)	40(59.7%)	12(17.9%)
	Fishing net	$\chi^2(1)=0.090, p=0.764$	$\chi^2(1)=6.129, p=0.013$	$\chi^2(1)=0.028, p=0.867$
	Yes	8(26.7%)	23(76.7%)	7(23.3%)
	No	23(24%)	49(51%)	21(21.9%)
	Fishing Boat	$\chi^2(1)=0.468, p=0.494$	$\chi^2(1)=8.312, p=0.004$	$\chi^2(1)=0.273, p=0.602$
	Yes	8(29.6%)	22(81.5%)	7(25.9%)
	No	23(23.2%)	50(50.5%)	21(21.2%)
	Mobile Phone	$\chi^2(1)=0.011, p=0.916$	$\chi^2(1)=0.857, p=0.354$	$\chi^2(1)=0.009, p=0.924$
	Yes	15(24.2%)	38(61.3%)	14(22.6%)
	No	16(25%)	34(53.1%)	14(21.9%)
<b>Behavioral</b>				
	Lifetime Sexual Partners	$\chi^2(4)=3.206, p=0.524$	$\chi^2(4)=6.675, p=0.154$	$\chi^2(4)=2.268, p=0.687$
	≤ 5 Partners	9(20.9%)	19(44.2%)	8(18.6%)
	6-10 Partners	1(33.3%)	18(54.5%)	9(27.3%)
	11-15 Partners	4(30.8%)	9(69.2%)	4(30.8%)
	16-20 Partners	(12.5%)	12(75%)	2(12.5%)
	≥ 20 Partners	5(23.8%)	14(66.7%)	5(23.8%)
	Last 1 Month Sexual partners	$\chi^2(2)=6.262, p=0.044$	$\chi^2(2)=2.329, p=0.312$	$\chi^2(2)=4.897, p=0.086$
	≤ 1 partner	23(30.7%)	46(61.3%)	20(26.7%)
	2Partners	3(8.8%)	19(55.9%)	3(8.8%)
	≥ 3 Partners	5(29.4%)	7(41.2%)	5(29.4%)
	Paid someone for sex	$\chi^2(1)=2.781, p=0.095$	$\chi^2(1)=0.646, p=0.421$	$\chi^2(1)=1.766, p=0.184$
	Yes	26(28.6%)	54(54.3%)	23(25.3%)
	No	5(14.3%)	18(51.4%)	5(14.3%)
	Condom Use	$\chi^2(1)=0.134, p=0.715$	$\chi^2(1)=0.000, p=1.000$	$\chi^2(1)=0.006, p=0.940$
	Yes	4(28.6%)	8(57.1%)	3(21.4%)

HIV&HSV-2 Co-Infection Risk Factors

	No	27(24.1%)	64(57.1%)	25(22.3%)
Alcohol use Before Sex		$\chi^2(1)=2.829, p=0.093$	$\chi^2(1)=1.312, p=0.252$	$\chi^2(1)=3.880, p=0.049$
Yes		6(42.9%)	10(71.4%)	6(42.9%)
No		25(22.3%)	62(55.4%)	22(19.6%)
Age at first sex		$\chi^2(2)=1.134, p=0.567$	$\chi^2(2)=0.738, p=0.692$	$\chi^2(2)=0.908, p=0.635$
< 15 yrs		12(23.5%)	29(56.9%)	12(23.5%)
15-20 yrs		19(26.4%)	42(58.3%)	16(22.2%)
> 20 yrs		0(0%)	1(33.3%)	0(0%)
Proximate/Biological				
Diagnosed with an STI Before		$\chi^2(1)=14.961, p<0.001$	$\chi^2(1)=20.181, p<0.001$	$\chi^2(1)=16.898, p<0.001$
Yes		22(42.3%)	42(80.8%)	21(40.4%)
No		9(12.2%)	30(40.5%)	7(9.5%)
Personal hygiene				
Bath Before sex		$\chi^2(1)=1.331, p=0.249$	$\chi^2(1)=0.000, p=1.000$	$\chi^2(1)=0.173, p=0.678$
Yes		3(42.9%)	4(57.1%)	2(28.6%)
No		28(23.5%)	68(57.1%)	26(21.8%)
Bath After Sex		$\chi^2(1)=0.425, p=0.514$	$\chi^2(1)=0.000, p=1.000$	$\chi^2(1)=2.118, p=0.146$
Yes		1(14.3%)	4(57.1%)	0(0%)
No		30(25.2%)	68(57.1%)	28(23.5%)

Table 4. univariate logistic regression with HIV-1, HSV-2 and co-infection with HIV-1 and HSV-2 as response variables

		<b>HIV/HSV-2 Co-infection</b>			
<b>Variable</b>		<b>AOR</b>	<b>Lower CI</b>	<b>Upper CI</b>	<b>p-value</b>
Marital Status	Intercept	2.00	0.37	10.92	0.42
	Married	0.19	0.03	1.09	0.06
	Single	0.03	0.003	0.25	0.001
Previously Infected by Gonorrhoea	Intercept	0.18	0.11	0.32	<0.001
	Yes	4.44	1.78	11.10	<0.001
Previously infected by Chlamydia	Intercept	0.23	0.15	0.37	<0.001
	Yes	6.41	1.67	24.66	<0.001
Previously Infected with Syphilis	Intercept	0.18	0.102	0.31	<0.001
	Yes	5.67	2.203	14.57	<0.001
Sexual Partners for the last 1 Month	Intercept	0.36	0.22	0.61	<0.001
	≥ 3 partners	1.15	0.36	3.66	0.82
	2 partners	0.27	0.07	0.97	0.04
Alcohol Use Before Sex	Intercept	0.24	0.15	0.39	<0.001
	Yes	3.07	0.97	9.75	0.057
<b>HIV Infection</b>					
Marital Status	Intercept	2.0	0.37	10.92	0.42
	Married	0.19	0.03	1.09	0.06
	Single	0.03	0.003	0.25	0.001
Previously Infected with Gonorrhoea	Intercept	0.23	0.14	0.38	<0.001
	Yes	3.57	1.46	8.71	0.005
Previously Infected with Chlamydia	Intercept	0.27	0.18	0.43	<0.001
	Yes	5.46	1.43	20.86	0.013
Previously Infected with Syphilis	Intercept	0.205	0.12	0.35	<0.001
	Yes	5.70	2.25	14.45	<0.001
Sexual Partners for the last 1 Month	Intercept	0.44	0.27	0.72	0.001
	≥3 Partners	0.94	0.30	2.98	0.92
	2 Partners	0.22	0.06	0.79	0.02
<b>HSV-2 Infection</b>					
Age	Intercept	4.00	1.64	9.79	0.0024
	30-34 yrs	0.500	0.12	2.02	0.33



	35+yrs	1.88	0.33	10.53	0.48
	<25 yrs	0.14	0.05	0.39	<0.001
<b>Marital Status</b>					
	Intercept	2.00	0.37	10.92	0.42
	Married	1.43	0.24	8.38	0.69
	Single	0.13	0.02	0.83	0.03
<b>Previously Infected with Gonorrhea</b>					
	Intercept	0.98	0.66	1.46	0.92
	Yes	4.90	1.73	13.90	0.003
<b>Previously Infected with Chlamydia</b>					
	Intercept	1.19	0.82	1.71	0.35
	Yes	7.57	0.93	61.70	0.06
<b>Previously Infected with Syphilis</b>					
	Intercept	0.96	0.65	1.42	0.84
	Yes	7.98	2.25	28.28	0.001
<b>Likely infection with HIV/AIDS</b>					
	Intercept	1.21	0.74	2.00	0.45
	Impossible	4.67	1.24	17.56	0.02
	Likely	0.82	0.32	2.12	0.69
	Unlikely	0.67	0.24	1.86	0.44

Table 5. Multivariate logistic regression with the predictors found to be significant in Table 3

Best Fitting Model	HIV/HSV-2 Co-infection					
	Variable	AIC	OR	Lower CI	Upper CI	p-value
	Intercept		0.72	0.10	5.12	0.75
	Married		0.24	0.03	1.73	0.16
	Single		0.04	0.003	0.47	0.01
	gonorrhoeas	113.7	3.11	1.10	8.74	0.03
	syphilis		5.16	1.78	14.93	0.002
	≥3 partners		2.31	0.56	9.52	0.23
	2 partners		0.36	0.08	1.54	0.17
			HIV Infection			
	Intercept		4.02	0.45	35.66	0.21
	Married		0.10	0.01	0.95	0.04
	Single	126.3	0.02	0.002	0.29	0.003
	Chlamydia		4.75	1.10	20.51	0.03
	≥ 3 partners		1.37	0.37	5.15	0.64
	2 partners		0.29	0.08	1.13	0.07
			HSV-2 Infection			
	Intercept		2.00	0.13	6.20	0.92
	Married		0.19	0.26	13.24	0.54
	Single	136.01	0.03	0.02	1.58	0.13
	Gcys		2.00	0.83	8.13	0.10
	Syphes		0.19	1.57	25.21	0.01

## **CHAPTER 4.0            DISCUSSION AND CONCLUSION**

In this study of 126 fishermen, prevalence of HSV-2 is high (57%) compared to other African cohorts [44,45]. The prevalence of co-infection is also high (22.2%). This study aims to provide a clearer picture regarding risk factors occurring amongst HIV-1 and HSV-2 co-infected individuals so that the scientific findings could be translated into sustainable prevention programmes and improved public health policies.

We see associations between age and HSV-2 but no association with HIV or coinfection. These differ slightly with existing literature as HIV is usually associated with age [23]. This could be caused by a larger number of younger fishermen than older fishermen and this explains the duration of sexual exposure to be less. For HSV-2 infection association with age is consistent with other studies. [46-49].

The risk of co-infection with HIV-1 and HSV-2 is lower among the married and single men as compared to those who are divorced/separated/widowed. The explanation of this could be, the divorced/separated/widowed are used to having a sexual partner and the culture allows them to inherit. As they try to fill this gap they become more vulnerable to acquiring STIs.

The risk of infection with HIV-1 or HSV-2 or both is higher among those who were previously diagnosed with gonorrhea or syphilis or chlamydia as compared to those not. Acquiring of any STI is associated with unsafe sex and therefore increasing the chances of co-infection. We also find that the number of sexual partners explains co-infection with HIV-1 and HSV-2. Having multiple sexual partners exposes one to acquiring STIs.

A limitation of this cross-sectional study is the inability to determine the effect of male circumcision on HIV-1 and HSV-2 co-infection among this group since all men are not circumcised, thus limiting the analysis of behavioral correlates with co-infection.

The high prevalence of both HIV-1 and HSV-2 among fishermen underlines the need for education and counselling among this group.

With HSV-2, we really may see reduced HIV transmission rates with the treatment and suppression of HSV-2 infection, even in patients without symptoms of disease.

Approaches to reducing the transmission of HSV-2 and HIV-1 are needed. The increasing number of HSV-2- and HIV-1 seropositive persons is of concern regarding the influence HSV-2 will continue to play in promulgating the HIV-1 epidemic among Fishermen in Kisumu. Recent consideration by the World Health Organization to include antiviral therapy for the syndromic treatment of GUD will lead toward providing more appropriate therapy for genital ulcers in the developing world. The tight association between HSV-2 and HIV-1 also provide an impetus to speed development of a vaccine to prevent HSV-2 acquisition or reactivation. From both clinical and public health perspectives, there is a clear imperative to test different approaches to interrupting the synergistic link between HSV-2 and HIV-1.

*List Of References*

1. Looker KJ, Garnett GP, Schmid GP. An estimate of the global prevalence and incidence of herpes simplex virus type 2 infection. *Bull World Health Organ.* 2008;86(10):805-12, A.
2. Celum C, Wald A, Hughes J, et al. Effect of aciclovir on HIV-1 acquisition in herpes simplex virus 2 seropositive women and men who have sex with men: a randomised, double-blind, placebo-controlled trial. *Lancet.* 2008;371(9630):2109-19.
3. Watson-Jones D, Weiss HA, Rusizoka M, et al. Effect of herpes simplex suppression on incidence of HIV among women in Tanzania. *N Engl J Med.* 2008;358(15):1560-71.
4. Corey L, Wald A, Celum CL, Quinn TC. The effects of herpes simplex virus-2 on HIV-1 acquisition and transmission: a review of two overlapping epidemics. *J Acquir Immune Defic Syndr.* 2004;35(5):435-45.
5. Schacker T, Zeh J, Hu HL, Hill E, Corey L. Frequency of symptomatic and asymptomatic herpes simplex virus type 2 reactivations among human immunodeficiency virus-infected men. *J Infect Dis.* 1998;178(6):1616-22.
6. Siegal FP, Lopez C, Hammer GS, et al. Severe acquired immunodeficiency in male homosexuals, manifested by chronic perianal ulcerative herpes simplex lesions. *N Engl J Med.* 1981;305(24):1439-44.
7. Ng'ayo MO et al. Sexual and demographic determinants for herpes simplex virus type 2 among fishermen along Lake Victoria, Kenya. *Sex Transm Infect* 2008;84:140-142 doi:10.1136/sti.2007.028795
8. Cunningham AL, Dwyer DE (2004) The pathogenesis underlying the interaction of HIV and herpes simplex virus after co-infection. *J HIV Ther* 9: 9–13.
9. Bosnjak L, Miranda-Saksena M, Koelle DM, Boadle RA, Jones CA, et al. (2005) Herpes simplex virus infection of human dendritic cells induces apoptosis and allows cross-presentation via uninfected dendritic cells. *J Immunol* 174: 2220–2227.
10. Jones CA, Fernandez M, Herc K, Bosnjak L, Miranda-Saksena M, et al. (2003) Herpes simplex virus type 2 induces rapid cell death and functional impairment of murine dendritic cells in vitro. *J Virol* 77: 11139–11149.
11. Mikloska Z, Bosnjak L, Cunningham AL (2001) Immature monocyte-derived dendritic cells are productively infected with herpes simplex virus type 1. *J Virol* 75: 5958–5964.

12. Smith JS, Robinson NJ (2002) Age-specific prevalence of infection with herpes simplex virus types 2 and 1: a global review. *J Infect Dis* 186: Suppl IS3–28.
13. Barbour JD, Sauer MM, Sharp ER, Garrison KE, Long BR, et al. (2007) HIV-1/HSV-2 Co-Infected Adults in Early HIV-1 Infection Have Elevated CD4+ T Cell Counts. *PLoS ONE* 2(10): e1080. doi:10.1371/journal.pone.0001080.
14. Crostarosa F, Aravantinou M, Akpogheneta OJ, Jasny E, Shaw A, et al. (2009) A Macaque Model to Study Vaginal HSV-2/Immunodeficiency Virus Co-Infection and the Impact of HSV-2 on Microbicide Efficacy. *PLoS ONE* 4(11): e8060. doi:10.1371/journal.pone.0008060.
15. Mark KE, Corey L, Meng TC, Margaret AS, Huang ML, et al. (2007) Topical resiquimod 0.01% gel decreases herpes simplex virus type 2 genital shedding: a randomized, controlled trial. *J Infect Dis* 195: 1324–1331.
16. Mary Beth Nierengarten. Treatment of HSV-2 Coinfection Reduces HIV Shedding.. 13th CROI: Abstract 33LB. Presented February 7, 2006.
17. Kallol Saha et al. Recent pattern of Co-infection amongst HIV seropositive individuals in tertiary care hospital, kolkata. *Virol J.* 2011; 8: 116.
18. Looker KJ, Garnett GP, Schmid GP: An estimate of the global prevalence and incidence of herpes simplex virus type 2 infection. *Bull World Health Organ* 2008, 86(10):805-12.
19. Corey L, Wald A, Celum CL, et al.: The Effects of Herpes Simplex Virus-2 on HIV-1 Acquisition and Transmission: A Review of Two Overlapping Epidemics. *J Acquir Immune Defic Syndr* 2004, 35(5):435-445.
20. Robinson NJ: Age-specific prevalence of infection with herpes simplex virus types 2 and 1: a global review. *J Infect Dis* 2002, 186(Suppl 1):S3-28.
21. Freeman EE, Weiss HA, Glynn JR, et al.: Herpes simplex virus 2 infection increases HIV acquisition in men and women: systematic review and meta-analysis of longitudinal studies. *AIDS* 2006, 20(1):73-83.
22. Van de Perre P, Segondy M, Foulongne V, et al.: Herpes simplex virus and HIV-1: deciphering viral synergy. *Lancet Infect Dis* 2008, 8(8):490-7.
23. Andrew Mujugira et al, Risk Factors for HSV-2 Infection among Sexual Partners of HSV-2/HIV-1 Co-Infected Persons. *BMC Research Notes* 2011, 4:64.

24. Bollen LJM, Whitehead SJ, Mock PA, et al. Maternal herpes simplex virus type 2 coinfection increases the risk of perinatal HIV transmission: possibility to further decrease transmission? *AIDS* 2008;22:1169-76.
25. Emma Hitt. Understanding the Role of HSV Coinfection. Published: 10/07/2002; Updated: 10/04/2002.
26. Mark Mascolini ,HSV-2 Coinfection Seems to Slow Early CD4 Drop With HIV. Second International Workshop on HIV Transmission. August 26-28, 2007. Washington, DC
27. Monica A. Magadi. Cross-national analysis of the risk factors of child malnutrition among children made vulnerable by HIV/AIDS in sub-Saharan Africa: evidence from the DHS. DOI: 10.1111/j.1365-3156.2011.02733.x
28. John S. Preisser et al, Detecting Patterns of Occupational Illness Clustering with Alternating Logistic Regressions Applied to Longitudinal Data. Vol. 158. No. 5, DOI: 10.1093/aje/kwg169. *Am J Epidemiol* 2003; 158:495-501.
29. Edwin Michael et al, Ecological Meta-Analysis of Density-Dependent Processes in the Transmission of Lymphatic Filariasis: Survival of Infected Vectors. *J Med Entomol.* 2009 July ; 46(4): 873–880.
30. Gruder, C.L., Mermelstein, R.J., Kirkendol, S., Hedeker, D., Wong, S.C., Schreckengost, J., Warnecke, R.B., Burzette, R. & Miller, T.Q. (1993). Effects of social support and relapse prevention training as adjuncts to a televised smoking cessation intervention, *Journal of Consulting and Clinical Psychology* 61, 113–120.
31. Cowan FM, Hargrove J, Langhaug LF, Jaffar S, Mhuriyengwe L, Swarouth TD, et al. The appropriateness of core group interventions using presumptive periodic treatment among rural Zimbabwean women who exchange sex for gifts or money. *J Acquir Immune Defic Syndr* 2005; 38:202–207.
32. Schacker T, Zeh J, Hu H, et al. Changes in plasma human immunodeficiency virus type 1 RNA associated with herpes simplex virus reactivation and suppression. *J Infect Dis.* 2002;186:1718–1725.
33. McFarland W, Gwanzura L, Bassett MT, et al. Prevalence and incidence of herpes simplex virus type 2 infection among male Zimbabwean factory workers. *J Infect Dis.* 1999;180:1459–1465.
34. Kamali A, Nunn AJ, Mulder DW, et al. Seroprevalence and incidence of genital ulcer infections in a rural Ugandan population. *Sex Transm Infect.* 1999;75:98–102.

35. Gray RH, Mawer MJ, Brookmeyer R. et al. Probability of HIV-1 transmission per coital act in monogamous, heterosexual, HIV-1 discordant couples in Rakai, Uganda. *Lancet*. 2001;357:1149–1153.
36. Grosskurth H, Gray RH, Hayes R. et al. Control of sexually transmitted diseases for HIV-1 prevention: understanding the implications of the Mwanza and Rakai trials. *Lancet*. 2000;355:1981–1987.
37. H.A. Weiss et al. The epidemiology of HSV-2 infection and its association with HIV infection in four urban African populations. *AIDS* 2001, Vol 15 (suppl 4).
38. Freedman E, Mindel A. Epidemiology of herpes and HIV co-infection. *J HIV Ther*. 2004 Feb;9(1):4-8.
39. Ellen Setsuko Hendriksen et al. Predictors of Condom Use Among Young Adults in South Africa: The Reproductive Health and HIV Research Unit National Youth Survey. *Am J Public Health*, 2007 July; 97(7): 1241-1248.
40. LI, MINGI et al. Factors Associated with Adolescents' Physical Inactivity in Xi'an City, China. December 2006 - Volume 38 - Issue 12.
41. Mark Spencer and Kevin O'Hara. A Multiple Logistic Regression Model for Predicting the Development of *Phytophthora ramorum* symptoms in Tanoak (*Lithocarpus densiflorus*). USDA Forest Service Gen. Tech. Rep. PSW-GTR-194. 2007.
42. Anita Nirenberg et al. Oncology Nurses' Use of National Comprehensive Cancer Network Clinical Practice Guidelines for Chemotherapy-Induced and Febrile Neutropenia. Vol. 37, No.6, November 2010.
43. JAMES LEE. An Insight on the Use of Multiple Logistic Regression Analysis to Estimate Association between Risk Factor and Disease Occurrence. *Int. J. Epidemiol*. (1986) 15 (1): 22-29. doi: 10.1093/ije/15.1.22.
44. Tobian AA, Charvat B, Ssempijja V, et al: Factors Associated with the Prevalence and Incidence of Herpes Simplex Virus Type 2 Infection among Men in Rakai, Uganda. *J Infect Dis* 2009.
45. Freeman EE, Glynn JR: Factors affecting HIV concordancy in married couples in four African cities. *AIDS* 2004, 18(12):1715-21.
46. Kebede Y, Dorigo-Zetsma W, Mengistu Y, et al: Transmission of Herpes Simplex Virus Type 2 among factory workers in Ethiopia. *J Infect Dis* 2004, 190(2):365-72.
47. Mertz GJ, Benedetti J, Ashley R, et al: Risk factors for the sexual transmission of genital herpes. *Ann Intern Med* 1992, 116:197-202.

48. Weiss H: Epidemiology of herpes simplex virus type 2 infection in the developing world. *Herpes* 2004, 11(Suppl 1):24A-35A.
49. Wald A: Herpes simplex virus type 2 transmission: risk factors and virus shedding. *Herpes* 2004, 11(Suppl 3):130A-137A.
50. Reynolds SJ, Risbud AR, Shepard ME, et al. Recent herpes simplex virus type 2 infection and the risk of human immunodeficiency virus type 1 acquisition in India. *J Infect Dis* 187:1513-21, 2003.
51. Julie M Schneider et al. The use of multiple logistic regression to identify risk factors associated with anemia and iron deficiency in a convenience sample of 12–36-mo-old children from low-income families. *Am J Clin Nutr* March 2008 vol. 87 no. 3 614-620.
52. Diego M Diaz et al. Antireflux Surgery Outcomes in Pediatric Gastroesophageal Reflux Disease. *The American Journal of Gastroenterology* (2005) **100**, 1844–1852; doi:10.1111/j.1572-0241.2005.41763.x
53. M. P. Daugherty and R. P. P. Almeida. Estimating *Xylella fastidiosa* transmission parameters: decoupling sharpshooter number and feeding period. DOI: 10.1111/j.1570-7458.2009.00868.x



**ASSESSMENT OF ALTERNATIVE FOOD RESOURCES OF THE LESSER  
FLAMINGO (*Phoeniconaias minor*) IN SOME RIFT VALLEY SALINE LAKES IN  
KENYA AND TANZANIA**

**Margaret Nduku Kyalo**  
**B.Sc. Hon. University of Nairobi,**  
**Reg. No. I56/70585/2007**



Photo: Lesser flamingos (*Phoeniconaias minor*) engaged in different activities at Lake Bogoria. A large group in the courtship dance at the back while a few are feeding in the front (Photograph by Margaret Kyalo)

**A THESIS SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF  
MASTER OF SCIENCE IN HYDROBIOLOGY OF THE UNIVERSITY OF  
NAIROBI.**

**FEBRUARY 2012**

## DECLARATION

This thesis is my original work and has not been presented for a degree in any other University or Institution.

Margaret Nduku Kyalo

Reg. No. I56/70585/2007

Signature 

Date 13<sup>th</sup> February, 2012

This thesis has been submitted for examination with our approval as the university supervisors;

Prof. Kenneth Mavuti

School of Biological Sciences

University of Nairobi

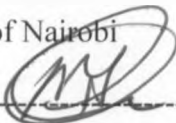
Signature 

Date 14/02/2012

Dr. Nathan Gichuki

School of Biological Sciences

University of Nairobi

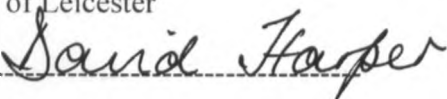
Signature 

Date 13/02/2012

Dr David Harper

Department of Biology

University of Leicester

Signature 

Date 14/02/2012

## **DEDICATION**

To my parents Dr. Richard (Late) and Sarah Luti, and to my special grandmother Susu Afia who have colored my life with inexpressible meaning.

## ACKNOWLEDGEMENTS

It is a pleasure to thank the many people who made this thesis possible. First and foremost I would like to express my deepest respect and most sincere gratitude to my supervisor, Prof. Kenneth Mavuti for his guidance and encouragement at all stages of my work. His constructive criticism and comments from the initial conception to the end of this work is highly appreciated. I must express my sincere gratitude to Dr. Nathan Gichuki as being my second supervisor. I want to thank him for his assistance, patient guidance and encouraging advises which were helpful during the thesis writing. With the deepest gratitude I wish to thank Dr. David Harper. His efforts, and advises to me were extremely great. I appreciate his funding for the fieldwork and his help is unforgettable. My thanks and gratitude goes to Dr. Ann Muohi, my initial supervisor, for her guidance during the proposal writing. I am grateful to Dr. John Githaiga for his kind assistance at the initial stages of this work.

I would like to thank the Kenya Wildlife Service for issuing me the permit to carry out my work at Lake Nakuru and Lake Bogoria also the manager of Crater Lake Lodge, Evans Kipturgo, for allowing me access to Lake Sonachi. I am grateful to Dr. Roberta Bettinetti for her help in identifying diatoms and Geoffrey Ongondo who helped me with the identification of protozoa. I also thank Timothy Mwinami and John Musina from the National Museum Ornithology Section, for providing me with National lesser flamingo population estimates.

I would like to thank Victoria Robinson and Emma Tebbs for their help with my field work and the participants of British Council Field IT East Africa Project who assisted me with data collection at Lake Natron, Lake Sonachi and Lake Oloidien. Special thanks to Velia Carn the camp manager and all her camp staff and to Reuben Ngeete and James Njoroge who made my field work bearable. I always enjoyed their company. I owe my deepest gratitude to Sally, May, Prudence, McKenzie, Muriithi, Mackonya and the Mavuno Mashariki family for their friendship, emotional support and for the many good moments. I am most indebted to God without whose word I would not have accomplished this work.

## TABLE OF CONTENTS

DECLARATION .....	i
DEDICATION .....	ii
ACKNOWLEDGEMENTS .....	iii
ABBREVIATIONS.....	xii
ABSTRACT.....	xiii
<b>CHAPTER ONE</b>	
INTRODUCTION .....	1
1.1. Background .....	1
1.2. Objectives .....	2
1.3. Research questions.....	2
1.4. Hypothesis.....	2
<b>CHAPTER TWO</b>	
LITERATURE REVIEW .....	3
2.1. Classification and geographical distribution of lesser flamingos .....	3
2.2. Geology, climate and hydrology in shaping the East African alkaline-saline lakes .....	3
2.3. Biological characteristics of alkaline-saline lakes .....	5
2.4. Main food resources of <i>Phoeniconaias minor</i> .....	5
2.5. Filter feeding adaptation of lesser flamingo .....	6
2.6. Lesser flamingo feeding behaviours .....	7
2.7. Other food resources of the lesser flamingo .....	8
2.8. Lesser flamingo population trends.....	9
2.9. Productivity of saline lakes.....	12
2.10. Justification of the study .....	13
<b>CHAPTER THREE</b>	
STUDY AREA, MATERIALS AND METHODS .....	15
3.1. Study area .....	15
3.1.1. Description of study sites.....	16

3.2.	Materials and Methods.....	23
3.2.1.	Determination of the composition of lesser flamingo food .....	23
3.2.1.1.	Sample collection.....	23
3.2.1.2.	Species identification .....	24
3.2.2.	Assessment of available food .....	24
3.2.2.1.	Photosynthetic pigment analysis.....	24
3.2.2.2.	Calculation of the standing crop of available food .....	25
3.2.3.	Assessment of primary productivity .....	26
3.2.3.1.	Exposure experiment method .....	26
3.2.3.2.	Exposure experiment for measurement of planktonic and epipellic primary productivity .....	28
3.2.3.3.	Exposure experiment for measurement of planktonic primary productivity .....	29
3.2.4.	Statistical analysis.....	29

## CHAPTER FOUR

RESULTS .....	30	
4.1.	General observations and explanation of terms .....	30
4.2.	Taxonomic composition .....	30
4.3.	Standing crop of the various categories of <i>Phoeniconaias minor</i> food resources .....	36
4.3.1.	Standing crop of the planktonic and suspended epipellic food resource .....	36
4.3.2.	Standing crop of the sedimented <i>Arthrospira fusiformis</i> and the epipellic food resource. ....	38
4.3.3.	Standing crop of the lake shore mud food resource.....	40
4.4.	Effect of water depth on the standing crop of the various food resources.....	41
4.5.	<i>Phoeniconaias minor</i> population estimates and standing crop of the different food resources .....	43
4.6.	Primary productivity of the various communities .....	44
4.6.1.	Primary productivity of the planktonic and suspended epipellic community.....	44
4.6.2.	Primary productivity of the sedimented <i>Arthrospira fusiformis</i> and epipellic community .....	46

**CHAPTER FIVE**

**DISCUSSION, CONCLUSION AND RECOMMENDATIONS** ..... 48

5.1. Energy budget by *Phoeniconaias minor*..... 48

5.2. Primary producer community categories ..... 49

5.3. Production rate of lesser flamingo food..... 50

5.4. Foraging options ..... 51

5.5. Conclusion ..... 52

5.6. Recommendations..... 53

5.6.1. Further research actions ..... 53

5.6.2. Conservation and management actions..... 53

**REFERENCES**..... 54

**APPENDIX**..... 61

## LIST OF TABLES

Table 4.1: Species composition of the microbial community at the study lakes during the study period. ( (3) dominant, (2) abundant, (1) present , (-) absent (p) present).....	32
Table 4.2: Mean ( $\pm$ se) standing crop of the planktonic community .....	36
Table 4.3: Mean ( $\pm$ se) standing crop of the suspended epipellic community .....	37
Table 4.4: Mean ( $\pm$ se) standing crop of the sedimented <i>Arthrospira fusiformis</i> . .....	38
Table 4.5: Mean ( $\pm$ se) standing crop of the epipellic community .....	39
Table 4.6: Mean ( $\pm$ se) standing crop of the lake shore mud algae. ....	40
Table 4.7 : Comparison of primary productivity values measured during this study with others from previously published work from different locations. ....	43
Table 4.8 : Mean ( $\pm$ se) primary productivity and respiration of the plankton community. ....	45
Table 4.9 : Mean ( $\pm$ se) primary productivity and respiration of the suspended epipellic community. ....	45
Table 4.10 : Mean ( $\pm$ se) primary productivity and respiration of the sedimented <i>Arthrospira fusiformis</i> . ....	46
Table 4.11: Mean ( $\pm$ se) primary productivity and respiration of the epipellic community. ....	47
Table 5.1: Comparison of primary productivity values measured during this study with others from previously published work from different locations. ....	51



## LIST OF FIGURES

Figure 2.1: Lesser flamingo population trends for the East African population .....	11
Figure 2.2: Population trend at Lake Natron, the only viable breeding site in East Africa .....	11
Figure 2.3: Lesser flamingo population trend at Lake Bogoria, a major feeding site.. .....	11
Figure 2.4: Population trend of lesser flamingo at Lake Nakuru, a major feeding site .....	11
Figure 3.1: The location of study lakes in the Eastern Rift Valley.....	15
Figure 3.2: The location and image of Lake Bogoria in the Eastern Rift Valley. ....	17
Figure 3.3: The location and image of Lake Nakuru in the Eastern Rift Valley. ....	18
Figure 3.4: The location and image of Lake Elementeita in the Eastern Rift Valley.....	19
Figure 3.5: The location and image of Lake Oloidien and Lake Naivasha from which it was separated. ....	20
Figure 3.6: The location and image of Lake Sonachi in the Eastern Rift Valley .....	21
Figure 3.7: The location and image of Lake Natron in the Eastern Rift Valley.....	22
Figure 4.1: Mean ( $\pm$ se) standing crop of planktonic food resource in the study lakes. ....	36
Figure 4.2: Mean ( $\pm$ se) standing crop of suspended epipellic food resource in the study lakes.....	37
Figure 4.3: Mean ( $\pm$ se) standing crop of sedimented <i>A. fusiformis</i> at Lake Bogoria. ....	38
Figure 4.4: Mean ( $\pm$ se) standing crop of epipellic food resource in the study lakes. ....	39
Figure 4.5: Mean ( $\pm$ se) standing crop of lake shore mud algae in the study lakes. ....	40
Figure 4.6: Mean ( $\pm$ se) planktonic standing crop at varying water depths of all the study lakes .....	41
Figure 4.7: Mean ( $\pm$ se) epipellic standing crop at varying water depth of the study lakes .....	41
Figure 4.8: The comparison between the mean planktonic and sedimented <i>A.</i> <i>fusiformis</i> standing crop at Lake Bogoria at different water depths.....	42
Figure 4.9: Comparison between the mean suspended epipellic and epipellic biomass of Lake Natron and Lake Elementeita at different water depths.....	42
Figure 4.10: Mean ( $\pm$ se) net and gross primary productivity and respiration of the planktonic community of the lakes studied. ....	45

Figure 4.11: Mean ( $\pm$  se) net and gross primary productivity and respiration of the suspended epipellic community of the lakes studied..... 46

Figure 4.12: Mean ( $\pm$  se) net and gross primary productivity and respiration of the sedimented *A. fusiformis* at Lake Bogoria. .... 47

Figure 4.13: Mean ( $\pm$  se) net and gross primary productivity and respiration of the epipellic community of the studied lakes..... 47

## LIST OF PLATES

Plate 2.1: The head of <i>P. minor</i> showing (a) the upper mandible (UM) and the lower mandible (LM), (b) the transverse section of the bill and (c) the lamellated area of the bill. From Jenkin (1957) .....	6
Plate 2.2: Lesser flamingo feeding behaviours (a) filter feeding while swimming in deep water and (b & c) filter feeding while standing on shallow water (modified from Ridley <i>et al.</i> , 1955). .....	7
Plate 3.1: Gilson's corer showing the position of (a) the planktonic community and (b) the epipelagic community in the water column, and (c) sediment. ....	23
Plate 3.2: (a) Oxygen and temperature meter and (a <sub>1</sub> ) probe, (b) metabolism chamber top, (c) metabolism chamber base plate, (d) motor and (d <sub>1</sub> ) stirrer, (e) metabolism chamber dark cover. ....	26
Plate 3.3: (a) Set up of productivity exposure experiment, (a <sub>1</sub> ) dark metabolism chamber, (a <sub>2</sub> ) light metabolism chamber and (a <sub>3</sub> ) oxygen meter. (b) exposure experiment set up <i>in situ</i> .....	27
Plate 4.1: Cyanobacteria taxa identified. A. <i>Arthrospira fusiformis</i> (i) extended and (ii) compressed. B. <i>Arthrospira</i> spp. C. <i>Spirulina subsalsa</i> . D. <i>Anabaena</i> spp. E. <i>Anabaenopsis abijatae</i> . F. <i>Lyngbya pseudospirulina</i> . G. <i>Oscillatoria</i> spp. (i) <i>O. limosa</i> (moving),(ii) <i>O. limnetica</i> (stationary). ....	33
Plate 4.2: Some of the Bacillariophyta species identified. A-D. <i>Navicula</i> sp. E. & F. <i>Melosira</i> spp. G. <i>Cyclotella</i> spp. H. <i>Amphora</i> spp. I. <i>Sellaphora</i> spp. J. <i>Cymbella</i> spp. K. <i>Pleurosigma</i> spp. ....	34
Plate 4.3: Protozoan species (A-H) and rotifer species (I & J). A. <i>Frontonia</i> spp. B. <i>Euplotes</i> spp. C. an Euglenophyte species present at lake Nakuru only. D. <i>Condylostoma</i> spp. E. <i>Campanella</i> spp. was seen only at Lake Oloidien. F. Unidentified species 1. G. Unidentified species 2 (Tumbler). H. <i>Amoeba</i> sp. only seen in Lake Oloidien. I. <i>Brachionus plicatilis</i> J. <i>Hexarthra jenkinae</i> . ....	35

## LIST OF APPENDICES

Appendix 1.a: Mann-Whitney U Test results for planktonic standing crop.....	61
Appendix 1.b: Mann-Whitney U Test results for the standing crop of the suspended epipellic community. ....	62
Appendix 1.c: Mann-Whitney U Test results for standing crop of the sedimented <i>Arthrospira fusiformis</i> .....	62
Appendix 1.d: Mann-Whitney U Test results for standing crop of the epipellic community.	63
Appendix 1.e: Mann-Whitney U Test results for standing crop of the wet mud community.	64
Appendix 2.a: Mann-Whitney U Test results for primary productivity of planktonic community. ....	65
Appendix 2.b: Mann-Whitney U Test results for primary productivity of the suspended epipellic community.....	66
Appendix 2.c: Mann-Whitney U Test results for primary productivity of the sedimented <i>Arthrospira fusiformis</i> .....	67
Appendix 2.d: Mann-Whitney U Test results for primary productivity of the epipellic community. ....	68
Appendix 3.a: Energy calculations for standing crop and net primary productivity of the planktonic food resource.....	69
Appendix 3.b: Energy calculations for standing crop and net primary productivity of the suspended epipellic food resource .....	69
Appendix 3.c: Energy calculations for standing crop and net primary productivity of the sedimented <i>Arthrospira fusiformis</i> food resource.....	69
Appendix 3.d: Energy calculations for standing crop and net primary productivity of the epipellic food resource. ....	70
Appendix 3.e: Energy calculations for standing crop and net primary productivity of the wet mud food resource.....	70

## ABBREVIATIONS

Chl <i>a</i>	Chlorophyll <i>a</i>
C	Carbon
O	Oxygen
DW	Dry weight
mg	Milligram
cm	Centimeter
M	Meter
hr	Hour (Time)
L	Litre
μS	Micro-Siemens
SC	Standing crop
SCmax	Maximum standing crop
SCmin	Minimum standing crop
NPP	Net primary productivity
GPP	Gross primary productivity
Resp.	Respiration
Sd	Standard deviation
r.p.m.	Revolutions per minute
kcal	Kilo calories

## ABSTRACT

The saline lakes of Kenya and Tanzania are of high economic value and of great conservation and scientific value. They host >75% of the world's lesser flamingo population, which are a major income earner for these countries. The lesser flamingos (*Phoeniconaias minor*) are of scientific concern as they are near threatened. The analysis of the food resources that sustain the lesser flamingo populations and the ecology of alkaline saline ecosystems are useful in developing conservation strategies for these lakes and the flamingos. The study was carried out at saline lakes of the eastern rift valley within Kenya and Tanzania. The lakes were Bogoria, Nakuru, Elementeita, Oloidien, Sonachi and Natron. The study explored the diversity of the lesser flamingo food resources and their significance to the lake's ecology.

Four categories of the food resources that are utilized by lesser flamingos are presented here. These include planktonic, sedimented, epipellic food resources and algae growing on wet mud. The planktonic food resource which occurred within the water column was mainly composed of *Arthrospira* spp. Sedimented *A. fusiformis* was observed at Lake Bogoria where it formed a film on the sediment in the shallow water. The epipellic food resource was mainly composed of benthic diatoms which grow on the water-sediment interface in shallow water where light penetrated to the sediment. The wet mud resource was also composed of diatoms growing on mud along the lake edges. The study confirmed that lesser flamingos are indeed the main primary consumers on the saline lakes. At a daily energy intake of 314 kcal for body maintenance, the species can consume > 92% of the cyanobacteria at Lake Bogoria, Lake Oloidien and Lake Sonachi. It was found that epipellic and wet mud diatoms significantly contribute to the lesser flamingo diet. These food resources supported >98% of the lesser flamingo's food requirement at Lake Natron's southern lagoon. This was also true for Lake Elementeita in August 2009 when the lake level was very low and the maximum wadeable depth was not more than 3 cm.

Lesser flamingos cannot be sustained by cyanobacterial and algal food alone. Their diet is naturally enriched with 'animal' protein provided by protozoa and rotifer species. At least eight protozoan species were found with the commonest being *Frontonia* spp. which was found in all the lakes and dominant in some of the lakes. Some protozoan species were restricted to certain lakes, such as *Amoeba* spp. and *Campanella* spp. which only occurred in

Lake Oloidien. Two rotifer species, *Brachionus spp.* and *Hexarthra spp.*, were present in all lakes except in Lake Natron, where none was recorded.

The lakes exhibited very high primary productivity for both planktonic and epipellic measurements. The highest net primary productivities were recorded at Lake Bogoria with  $204.6 \text{ mg C m}^{-2} \text{ hr}^{-1}$  for planktonic cyanobacterial and  $103.01 \text{ mg C m}^{-2} \text{ hr}^{-1}$  for sedimented *Arthrospira*. This is the first study to describe the primary productivity of the epipellic community of saline lakes and its contribution to the ecology of the rift valley lakes studied. At the shallow lakes, Lake Elementeita and Lake Natron, it contributed 100% to the primary production. The highest epipellic net primary productivity was recorded at Lake Elementeita with  $73 \text{ mg C m}^{-2} \text{ hr}^{-1}$ , while negative values of up to  $-33 \text{ mg C m}^{-2} \text{ hr}^{-1}$  were recorded for the suspended epipellic community at Lake Natron. The epipellic and wet mud resources contribute greatly to the lesser flamingo diet than earlier thought and more so to the maintenance of the food chains on the saline lakes. The epipellic community is the main primary producer in the shallow lakes such as Natron.

# CHAPTER ONE

## INTRODUCTION

### 1.1. Background

The Eastern Rift Valley, which bisects Kenya from North to South, contains a series of shallow alkaline saline lakes. These lakes are in closed basins without outlets, and have high concentrations of sodium carbonate and bicarbonate (Melack and Kilham, 1974) as high evaporation leaves the salts behind. For this reason, Brown (1973) described the Kenyan saline lakes as “harsh wastes” since they are extremely saline with lakes Bogoria and Nakuru having conductivities of 40,000 – 80,000  $\mu\text{S cm}^{-1}$  (Harper *et al.*, 2003) and 14,000 – 26,000  $\mu\text{S cm}^{-1}$  (Vareschi, 1978) respectively.

Brown (1973) acknowledged that flamingos are one of the world’s greatest ornithological spectacles. The shores of these saline lakes are occasionally lined with enormous congregations of lesser flamingos (*Phoeniconaias minor* Geoffroy). Indeed, the saline lakes of the Kenyan Rift Valley are recognized worldwide as theatres for the magnificent pink aggregations of lesser flamingos, which are a major tourist attraction and foreign income earner for Kenya. These lakes are of significance to flamingos because they can support very high densities of their primary food organism that can sustain for example, up to a 1.5 million *P. minor* in Lake Nakuru (Vareschi, 1978).

The birds are known to feed predominantly on colonies of the microscopic planktonic cyanobacterium *Arthrospira fusiformis* (synonym *Spirulina platensis*). They feed by filtering *A. fusiformis* from few centimetres of the surface water. In addition, *P. minor* have been recorded to glean for food on mud along the shores of saline lakes, where the food consists mainly of diatoms (Tuite, 1981). The food resources vary from time to time in quantity and flamingos have evolved a nomadic behaviour to cope with this variability. The lesser flamingo population at a lake has been associated with varying food quantity.

Recently, *P. minor* have been observed displaying unusual feeding behaviour that has never been documented in detail before - they immerse their heads, sometimes also including the neck, into the water to feed at the water-mud interface in the littoral zone. These different feeding behaviours suggest that the birds were utilizing the epipelagic microbial community, which grows on the water-mud interface where light could penetrate to the sediment surface. This could potentially be an important food resource for lesser flamingos. This study is built on this premise and



aimed to investigate and quantify the alternative food resources of lesser flamingos in various saline lakes in Kenya and Lake Natron in Tanzania.

## 1.2. Objectives

The broad objective was;

To investigate the alternative food resources of lesser flamingos and quantify their contribution to the algal primary productivity of the Rift Valley saline lakes.

The specific objectives were;

1. To determine the composition of the planktonic, epipelagic and the lake shore mud food resources.
2. To establish the relative importance of these communities as possible food resource for the *P. minor* by measuring respective *in situ* biomass of the planktonic, epipelagic and the lake shore mud microbial communities.
3. To measure the productivity of the planktonic and epipelagic microbial communities.

## 1.3. Research questions

1. Is the blue-green algae *Arthrospira fusiformis* from the Rift Valley saline lakes the only food resource for lesser flamingo population in eastern Africa?
2. What other food resources support the lesser flamingo population?
3. What is the diversity of these food resources?
4. What is the production rate of the food resources?
5. How do the various food resources contribute to the primary productivity of these lakes?

## 1.4. Hypotheses

1. The food resources utilized by the lesser flamingos are variable and the eastern Africa population of lesser flamingos could not be maintained on planktonic *A. fusiformis* alone.
2. Epipelagic and lake shore mud microbial communities contribute significantly to the algal primary productivity of the saline lakes of East Africa and consequently to the food resources of lesser flamingos.

## CHAPTER TWO

### LITERATURE REVIEW

#### 2.1. Classification and geographical distribution of lesser flamingos

Flamingos belong to the family Phoenicopteridae which is made up of birds with remarkably long thin legs and extremely long flexible necks. They are descendants of an ancient lineage of microphagous, colonial wading birds that inhabited hypersaline lakes of tropical and subtropical regions (Bildstein *et al.*, 1993) but their relationship to other birds is unclear (Jenkins, 1957). Today, there are six living flamingo species in the world. Two species occur in East Africa, that is, the greater flamingo (*Phoenicopterus ruber roseus*) and the lesser flamingo (*Phoeniconaias minor* Geoffroy Saint-Hilaire, 1798). Lesser flamingo stand out distinctively with their pink plumage and dark-red bills (Stevenson and Fanshawe, 2002)

*P. minor* is the most numerous of all flamingo species (Childress *et al.*, 2008) with four recognized populations that are probably separate. The global population is estimated to be between 2,220,000 to 3,240,000 with the largest population which is more than 75% of the total population in the East African eco-region of approximately 1.5 to 2.5 million individuals. Other small populations of *P. minor* occur in West India/Pakistan with 650,000 individuals, 55,000-65,000 in southern Africa and 15,000- 25,000 in West Africa. Lesser flamingos inhabit alkaline saline lakes and pans in Africa and Asia. In Eastern Africa they are found within the Eastern Rift valley from Ethiopia through Kenya down to the saline lakes of Tanzania.

#### 2.2. Geology, climate and hydrology in shaping the East African alkaline-saline lakes

The East African Rift Valley was formed in the Cenozoic approximately 40 million years ago (Ma) as a result of volcanism and tectonic activity. In the Pliopleistocene, 2.5 Ma the African climate became drier due to the onset of the glacial cycles in the Northern Hemisphere (deMenocal 1995). From this time the East African region started experiencing periods of alternating humidity and aridity between 2.7-0.9 Ma (Trauth, 2005).

Diatomite evidence depicts the presence of very large lakes within the central Rift with deep lakes of more than 250m depth between 1.9-0.9 Ma, receiving 2000 mm yr<sup>-1</sup> of rainfall. The lakes have been hypothesized to have contributed to the

evolution of early hominids by forming barriers both when full of water and when dry (Trauth *et al.*, 2010). A large lake still existed 9,200 years ago within the central Rift but by 3,000 years it had become smaller (Richardson and Richardson, 1972). The present small lakes within the central Rift are all remnants of this large lake.

Recent historical sediment studies have shown the presence of lake level fluctuations accompanied by conductivity changes at Lake Sonachi (Verschuren *et al.*, 1999) and Lake Oloidien (Verschuren *et al.*, 2004) during the 19<sup>th</sup> century. The two lakes experienced dry spells in the mid 19<sup>th</sup> century with conductivity of 7,000 – 13,000  $\mu\text{S cm}^{-1}$  at Sonachi and 12,700 – 13,600  $\mu\text{S cm}^{-1}$  at Oloidien in 1883. In 1940-1960 the lakes went through another dry spell. The low lake level and incidents of major environmental episodes also coincide with ecological succession events of chironomids, diatoms (Verschuren *et al.*, 1999; Verschuren *et al.*, 2004) and rotifers (Epp *et al.*, 2010). Similar changes can be inferred for the other lakes at the same time.

The lakes are characterized by high alkalinity with pH as high as 10 or more due to high levels of  $\text{Na}^+$  trachyte lava from the surrounding volcanic highlands which is washed into the rivers that drain into these lakes. Low  $\text{Mg}^{2+}$  and  $\text{Ca}^{2+}$  ions lead to the formation of an alkaline salt  $\text{Na}_2\text{CO}_3 \cdot \text{NaHCO}_3 \cdot 2\text{H}_2\text{O}$  (Grant, 2004), known as trona or soda ash and locally referred to as 'magadi'. The lakes within the rift valley are endorheic without outlets and over time trona has accumulated as it is left behind after evaporation. It has been industrially mined to make glass, toothpaste and has been traditionally used in its raw form as a food tenderizer and additive (Nielsen and Dahi, 1995).

The lakes are renowned for their high salinity with high but stable conductivity of around 70,000  $\mu\text{S cm}^{-1}$  at deeper lakes such as Lake Bogoria (Harper *et al.*, 2003; Schagerl and Oduor, 2008) and high variability at shallow lakes such as Lake Nakuru 11,000 – 160,000  $\mu\text{S cm}^{-1}$  due to lake level fluctuations (Verschuren *et al.*, 2004). As a result of the high alkalinity and salinity the lakes are referred to as alkaline saline lakes.

### 2.3. Biological characteristics of alkaline-saline lakes

The phytoplankton is often dominated by the cyanobacteria *Arthrospira fusiformis*, *Anabaenopsis* sp. and the single celled *Synechococcus* sp. Common diatom species include *Anomoeoneis* sp. *Navicula* sp. and *Nitzschia* sp. (Oduor and Schagerl, 2007). Phytoplanktonic primary productivity is very high, with the main primary consumer of the cyanobacteria being the lesser flamingo.

Rotifer species of *Brachionus dimidiatus*, *B. plicatilis* and *Hexartha jenkiniae* characterized the zooplankton. A copepod *Lovenula africana* is sometimes a dominant primary consumer at Lake Nakuru. An alkaline-water cichlid fish *Alcolapia alkalicus graham* (formerly *Sarotherodon alkalicus graham*) (Vareschi and Jacobs, 1984) is present in lakes Natron, Magadi, Nakuru and Elementeita, where it has a population refuge in the hot springs. The fish sustains the large numbers of pelicans of more than 22,000 individuals as recorded in 1969 by Bartholomew and Pennycuik (1973).

### 2.4. Main food resources of *Phoeniconaias minor*

*Phoeniconaias minor* are able to congregate in large numbers on these Kenyan saline lakes particularly because of the availability of their primary food resource, the microscopic cyanobacteria, mainly *Arthrospira fusiformis*. The major lakes in Kenya utilized by lesser flamingos include Lake Bogoria, Lake Nakuru and Lake Elementeita and most recently Lake Oloidien, which was separated from Lake Naivasha, a freshwater lake, in the 1980s due to declining water level (Harper and Mavuti, in McClanahan and Young, 1996) and has been slowly salinizing progressing towards an alkaline-saline lake.

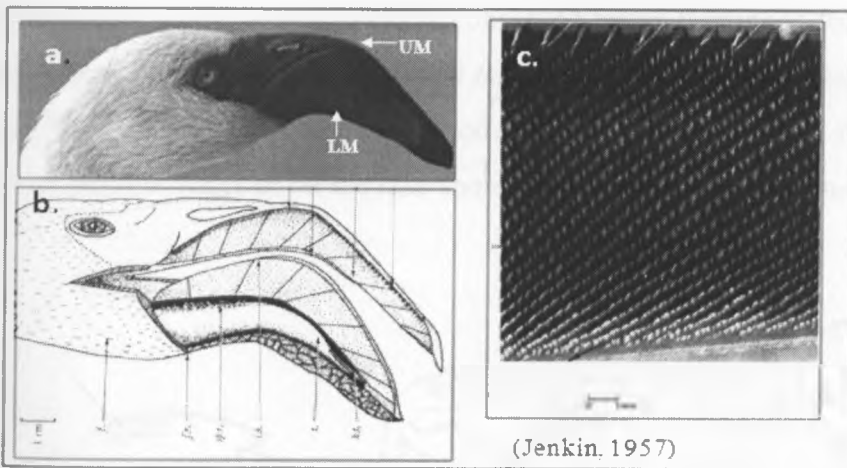
In these lakes, *A. fusiformis* can sustain a continuous steady algal bloom but from time to time, unpredictably, there is a crash (Tuite, 1981). For this reason *P. minor* are nomads of saline lakes because of their random movements from lake to lake in search of suitable ecological conditions and food. According to Tuite (1979), the saline lakes often have high densities of flamingos, which coincide with high densities of *A. fusiformis*. Diatoms are also a food resource for *P. minor* and seem to have more stable densities in shallow lakes where light can reach the sediment (Tuite, 1981) with a mean of  $\sim 45 \text{ mg Chl } a \text{ m}^{-2}$  in shallow lakes. Not much research has been conducted on assessing how much diatoms contribute to the food of lesser flamingo or to the ecology of the saline lakes. The most notable studies relating to the

food of lesser flamingo are those carried out in the 1970s at Lake Nakuru by Vareschi (1978 & 1982) and Vareschi and Jacobs, (1984, 1985); however they focused more on *A. fusiformis* because it occurred in higher densities in most lakes than the other cyanophytes and algal species.

## 2.5. Filter feeding adaptation of lesser flamingos

Flamingos are filter feeders; they are adapted to this mode of feeding by being equipped with specialized and unique filtering equipment (shown on Plate 2.1). They have lamellae with platelets that form a filter that acts as a sieve within the bills. The six species of flamingos have different sizes of filters that enable them to sieve organisms of different sizes. *P. minor* has the finest filter that enables it to sieve organisms in the size range 20-100 $\mu$  (Jenkin, 1957).

The tongue functions as a pump pressing water through the bill while the platelets strain out the cyanobacteria and algae. These are the main food for *P. minor* in the alkaline-saline lakes of Kenya. Such an adaptation allows *P. minor* not only to acquire food in such harsh conditions but also to avoid consuming the lethal waters of alkaline saline lakes that may cause physiological damage.



**Plate 2.1:** The head of *P. minor* showing - (a) the upper mandible (UM) and the lower mandible (LM), (b) the transverse section of the bill and (c) the lamellated area of the bill. From Jenkin (1957)

According to the study by Pennycuick and Bartholomew (1973), flamingos have a constant pumping rate, but the concentration of cyanobacteria and algae in the water is what determines how much time *P. minor* would devote to feeding. To make a positive energy surplus, a minimum *A. fusiformis* concentration of 120 g dry weight (DW)  $\text{m}^{-3}\text{day}^{-1}$  is required by a non-breeding lesser flamingo and a breeding one would require 250 g DW  $\text{m}^{-3}\text{day}^{-1}$  to produce an egg. This could be achieved if the flamingo spends 80% of its time feeding. Feeding rates decrease with decreasing algal densities and feeding stops at algal concentrations of 100 g DW  $\text{m}^{-3}$  and below (Pennycuick and Bartholomew, 1973; Vareschi, 1978).

## 2.6. Lesser flamingo feeding behaviours

Different patterns of feeding behaviours displayed by *Phoeniconaias minor* have been documented. For instance, Ridley *et al* (1955) explained that they feed by sweeping the surface of the water with their beaks while swinging their heads from side to side to collect floating phytoplankton from the upper two inches (Brown, 1973) of water. They can feed in this way while swimming or standing in shallow water (Plate 2.2). Jenkin (1957) mentioned an additional feeding behaviour of *P. minor*, the stamping of their webbed feet to stir up food from the bottom in shallow waters. In another study, Vareschi (1978) observed lesser flamingos feeding on mud flats during times of very low *Arthrospira fusiformis* density. He suggested that they may have been feeding on diatoms, a food resource found growing mostly on the sediment in shallow water or on the lake shore mud (Hecky and Kilham, 1973).

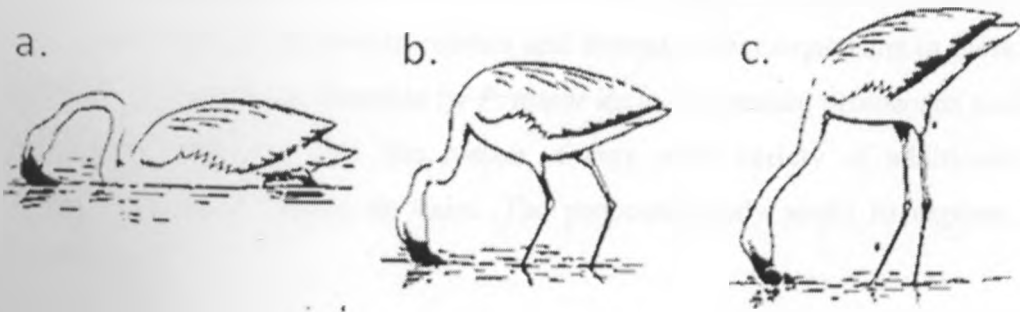


Plate 2.2: Lesser flamingo feeding behaviours (a) filter feeding while swimming in deep water and (b& c.) filter feeding while standing on shallow water (modified from Ridley *et al.*, 1955)

## 2.7. Other food resources for the lesser flamingos

Whilst *P. minor* are very specialized in the unique way they feed, they are generalist feeders in that they can ingest anything of the size range of 20-100 $\mu$  (Jenkins, 1957). Although Vareschi (1978) recognized *A. fusiformis* as the main food resource for *P. minor*, he expounded on other possible food resources for *P. minor*. He demonstrated that, during seasons of high rotifer densities- of 19 mg DW L<sup>-1</sup> at Lake Nakuru - the rotifers significantly supplemented the *P. minor* diet by up to 20%.

Other minor food resources include cyanobacterial species of the genera *Anabaenopsis* and *Anabaena*, and diatoms. *Anabaenopsis* and *Anabaena* flourish at Lake Nakuru and Lake Elementeita when there is low *Arthrospira fusiformis* biomass (Vareschi, 1978). Single-celled cyanobacteria of the genus *Synechocystis*, *Synechococcus* and *Monoraphidium* dominate the phytoplankton of saline lakes such as Lake Nakuru from time to time. They are too small (2-6 $\mu$  in diameter) to be retained by the lamellae. Between 1974- 1975 single celled cyanobacteria dominated Lake Nakuru, which coincided with low *P. minor* population.

In research carried out in 1974 at Lake Nakuru, Vareschi (1978) also observed that at times of low densities of *A. fusiformis*, *P. minor* tended to feed on the mud flats at the lake edges where there was a film rich in diatoms that seemed to be a sustainable food source for them. He noted a remarkable increase in flamingo numbers at a time when the cyanobacterial density was low and suggested that this film of diatoms on the mud flats could significantly substitute for *A. fusiformis* as food for lesser flamingos. It is reasonable to imagine that benthic diatoms could sufficiently meet the nutritional needs of *P. minor* and may be of equal or even higher importance as flamingo food than *A. fusiformis* when they occur in high densities.

Apart from *A. fusiformis*, rotifers and diatoms, other organisms in these lakes that fall in the size range filterable by *P. minor* include desmids, protozoans and other cyanobacteria species. For this reason, a very wide variety of additional food resources for the *P. minor* do exist. The proposed study seeks to explore these alternatives.

## 2.8. Lesser flamingo population trends

A National Waterbird census is carried out in Kenya every January and July over the last two decades. Flamingo census is done at lakes Nakuru, Bogoria, Elementeita, Sonachi and most recently Lake Oloidien. Due to its recent salinisation and colonization by *A. fusiformis* (Ballot *et al.*, 2009) lesser flamingos recently started to be utilized it as a feeding site.

Different methods are used in estimating flamingo populations, the most affordable being ground census where the lake is divided into sections and teams cover the shore by foot counting birds encountered (Owino *et al.*, 2001). The teams are led by an experienced ornithologist and the members estimate the numbers of flamingos within a specified crowd of flamingos. The numbers that are within close range among the group members are averaged while those that are far out of the group's range are excluded. This is the method used for the national water birds census to estimate flamingo numbers. The method's main limitation is its inability to accurately estimate numbers in large groups of flamingo but it has been proved reliable (Morales-Roldan *et al.*, 2011).

Aerial photography surveys have also been used to estimate flamingo population (Tuite, 1979), where photographs are taken from airplane flying over a lake. Other than the cost implications it seems to be the most reliable. New methods are being developed such as use of photographs taken using mobile phones (Iliffe *et al.*, 2011) to estimate flamingo populations.

The earliest available reports from observations in the 1950s by Brown (1973) reported very high estimates of 4 million lesser flamingos within Kenya and Northern Tanzania (Fig. 2.1.). In October 1958 he observed more than 500,000 breeding pairs that may have resulted in at least 460,000 young chicks by February of 1959 at Lake Natron. Thereafter, between 1968 and 1969 Bartholomew and Pennycuick (1973) reported a total population of 1,043,000 from an areal survey in 24 lakes in Kenya and Tanzania. The National Waterbirds Census conducted each January since 1991 has recorded numbers that ranged from 337,000-1,470,000 from 1991-1999 within the Kenyan saline lakes (Owino *et al.*, 2001). In 1994, Woodworth *et al.* (1997) conducted an aerial photographic survey of flamingo population at 9 lakes in Tanzania and recorded a total of 907,000 birds. An estimated 2,800,000 birds were reported in 13 lakes with 1% of the population being greater flamingos (Bartholomew



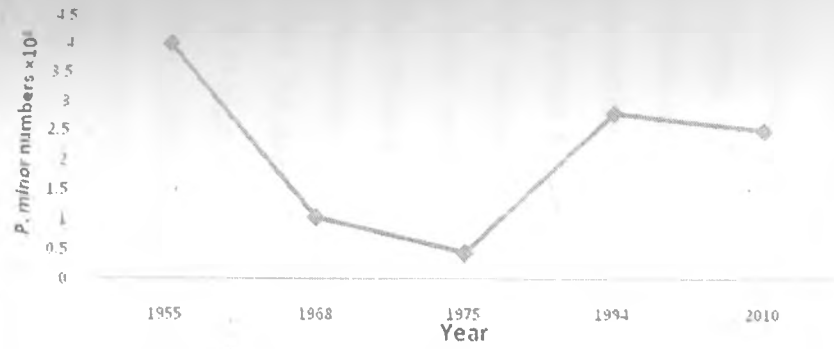
and Pennycuik, 1973). Between 1974 and 1976 estimates by Tuite (1979) at 22 lakes in Kenya and Northern Tanzania gave a lower estimate of 350,000 – 550,000 birds with some lakes having no flamingos. In 2002, 634,440 birds were recorded in 13 lakes in Tanzania (Raini and Ngowe, 2009).

A survey of nests at Natron for three years from 1965-1967 gave a total of ~100,000 nests whereas no breeding was reported in 1968 while in 1969 a minimum of 100,000 nests were recorded (Fig. 2.2). Most recently in January 2011, Baker (2011) recorded 35,000 hatchlings at Natron. At Lake Bogoria, lesser flamingo numbers varied from 40,000 in December 2000; 297,000 in January 2001; 510,000 in August 2001 to 30,000 birds in 2003 (Fig. 2.3.). In the 1970's high population of approximately one million flamingos was recorded at Lake Nakuru (Fig. 2.4.) by Vareschi (1978). Periods of high mortality have also been witnessed such as at Lake Bogoria in August 2001 where 114 birds died daily (Harper *et al.*, 2003).

There is certainly a big disparity between earlier estimates by Brown (1973) and more recent reports that suggests a serious decline in numbers. In fact there are reports of declining population in different locations; Etosha Pan in Southern Africa (Simmons, 1995), Sambhar Salt Lake in India (Kulshreshtha *et al.*, 2011) and Lake Natron in Tanzania (Clamsen *et al.*, 2011). These declines can be explained by the fact that flamingos often disperse outside the main feeding lakes, which are mostly surveyed. Counts seem to only concentrate on the easily accessible lakes and those seen to occasionally harbour large flamingo populations.

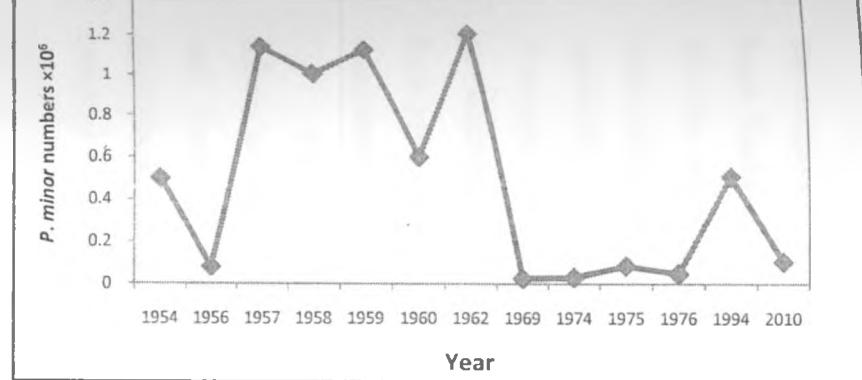
Nonetheless, these reports of declining population (Simmons, 1996; Clamsen *et al.*, 2011) raise concern. In 1988 the IUCN Red List had classified *P. minor* as a species of 'Low Risk/Least Concern'. Although it is the most numerous flamingo species its status was raised in 2000 to 'Low Risk/Near Threatened' and again in 2004 to the current 'Near Threatened' (IUCN, 2010) status. This is due to increasing concern owing to reports that imply decreasing population and threats posed to the only breeding site, Lake Natron, for the East Africa population by the development of a soda mining facility.

**Total lesser flamingo population estimate in east africa**



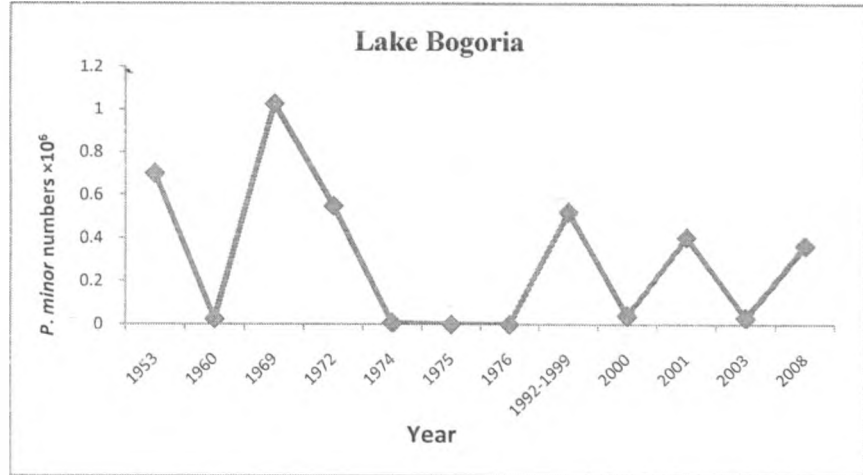
**Fig. 2.1:** Lesser flamingo population trends for the East African population.

**Lake Natron**



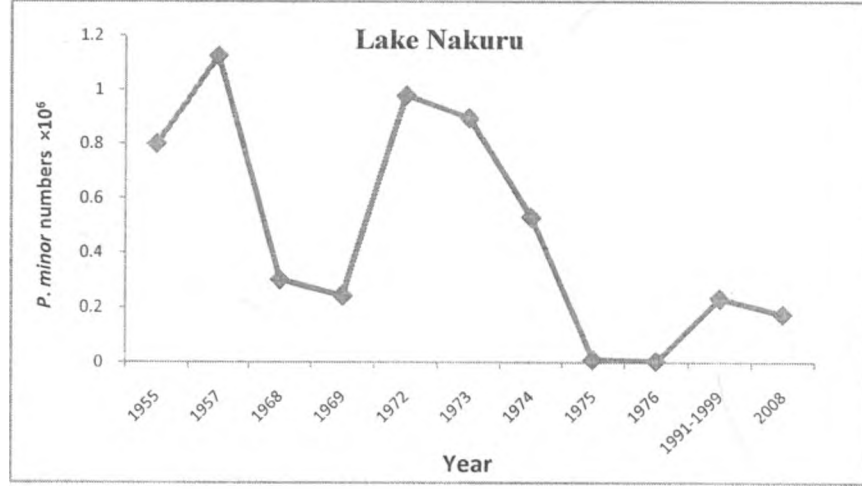
**Fig. 2.2:** Population trend at Lake Natron, the only viable breeding site in East Africa.

**Lake Bogoria**



**Fig. 2.3:** Lesser flamingo population trend at Lake Bogoria, a major feeding site.

**Lake Nakuru**



**Fig. 2.4:** Population trend of lesser flamingo at Lake Nakuru, a major feeding site.

References: Brown, 1973; Bartholomew and Pennycuik, 1973; Tuite, 1979; Vareschi, 1978; Woodworth *et al.*, 1997; Owino *et al.*, 2001; Harper *et al.*, 2003; IUCN, 2010; Clamsen *et al.*, 2011.

## 2.9. Productivity of saline lakes

Apart from their rich bird life, tropical saline lakes are fascinating because of the magnitude of the rate of primary production of the algal blooms that support the huge flocks of *P. minor*. An understanding of the productivity of algae is essential because primary producers are the principal source of energy for saline lake ecosystems (Vareschi and Jacobs, 1985). Hammer (1981) gave a global review of the variability of primary production rates of saline lakes in three continents, Africa, Australia and North America. The highest reported phytoplankton primary productivity was  $10,000 \text{ mg C m}^{-3} \text{ h}^{-1}$  at Lake Aranguadi in Ethiopia. Indeed, such a high primary productivity clearly shows that saline lake ecosystems are important ecosystems.

Most of what is known about the primary production of the saline lakes in Africa has come from studies by Talling *et al.*, (1973), Melack and Kilham, (1974) and Vareschi, (1982). They mainly studied the primary productivity of the phytoplankton, mainly dominated by *A. fusiformis* of lakes in East Africa and reported high primary productivity. These researchers measured phytoplanktonic primary production but did not lay much emphasis on the importance of the epipellic algae in these lakes and its contribution to the lakes' overall primary production.

According to Hammer (1981) the importance of the epipellic microbial community to the flora of tropical saline lakes has been previously underestimated because of sampling difficulties. Nevertheless, studies carried out in temperate regions have demonstrated the importance of benthic algae production to total lake production. For instance, Stanley (1976) established that epipellic production is very important in aquatic ecosystems. He carried out this study between 1971 and 1973 and worked on tundra ponds and lakes with depths ranging from 20 cm to 2 m. Stanley recorded epipellic primary productivity that was nine times higher than phytoplanktonic production. In tundra ponds 20 cm deep, he recorded epipellic production that ranged from  $4 \text{ to } 10 \text{ g C m}^{-2} \text{ yr}^{-1}$  compared to  $1 \text{ g C m}^{-2} \text{ yr}^{-1}$  for the lakes 2m deep. Shallower ponds had a higher epipellic production than deeper lakes due to high light intensity at the sediment level. In another study, Björk-Ramberg and Ånell (1985) conducted experiments at Lake Stugsjön, a shallow Swedish subarctic lake. They found that epipellic algae constituted 70-83% of the total production in that lake.

Hammer (1981), referred to a study done by Wetzel (1964) at Borax Lake in North America on the primary production of blue-green algae that dominated the benthic littoral zone of the lake. The annual primary production was  $267\text{g C m}^{-2}$ , which was 69% of the total primary production of Lake Borax. This indeed, suggests that the epipelagic microbial community of saline lakes may be very important to the total primary production of saline lakes. Nevertheless, no studies in the tropical regions have investigated the importance of the epipelagic microbial communities and its response to the dynamics of change of lakes in the tropical region yet it seems to be very significant to the general ecology of saline lakes.

## 2.10. Justification of the study

The lesser flamingo is of great conservation and scientific value and also of great economic value as the species is a major tourist attraction and foreign income earner for Kenya and Tanzania. At present, it is classified as 'near threatened' on the IUCN (International Union for Conservation of Nature) Red List (2010) due to its specialized habitat requirements. There is increasing concern due to large-scale die-offs of *P. minor* which have occurred with greater frequencies since 1993 (Ndetei and Muhandiki, 2005) than previously within the Kenyan saline lakes.

Recent studies in the East African flamingo eco-region indicated that cyanobacterial toxins may be the cause of *P. minor* deaths in Kenya (Krienitz *et al.*, 2003, Ballot *et al.*, 2004, Ballot *et al.*, 2005) and in Tanzania (Lugomela *et al.*, 2006). Ballot *et al.*, 2004 and Ballot *et al.*, 2005 attributed these deaths to the ingestion of hepatotoxins and neurotoxins produced mainly by *Arthrospira fusiformis*. Flamingo deaths have also been reported to also occur during times of algal crash, such as at Lake Bogoria. *P. minor* were weakened by lack of food at a time of *A. fusiformis* crash and those that could not move to other lakes eventually died of starvation (Owino *et al.*, 2001). There is also potential DDT contamination (Bettinetti *et al.*, 2011).

Saline lakes are faced with mounting pressures from deforestation of their catchments, water over-abstraction from the rivers that feed the lakes, siltation and alteration of habitats (Mwinami *et al.*, 2010). All these factors may have resulted in habitat modification hence contributing to decrease in the flamingo population. The rate of population decline is, however, difficult to quantify due to the birds nomadic

nature (Childress *et al.*, (2008). They move from lake to lake unpredictably in search of suitable ecological conditions and abundant food resources.

As the lesser flamingo is of great conservation and scientific value, it therefore makes the protection of their habitat an issue of major concern. A better understanding of the food and feeding requirements of *P. minor* is necessary so as to comprehend how saline lakes can provide alternative food resources to sustain food requirements of lesser flamingos throughout the year. This will enable further investigations into other possible causes of deaths of the species. In the face of a changing climate it is important to know how these changes would affect the food resources that the species depends on.

# CHAPTER THREE

## STUDY AREA, MATERIALS AND METHODS

### 3.1. Study area

The primary focus of the study was lakes Bogoria, Nakuru and Elementeita that lay within the Central Rift Valley in Kenya. However other study sites were considered in the study in order to obtain some regional estimates for comparison. These were lakes Oloidien and Sonachi (Crater Lake) within the Naivasha basin in Kenya and Lake Natron within the South Rift Valley in Tanzania (Fig. 3.1).

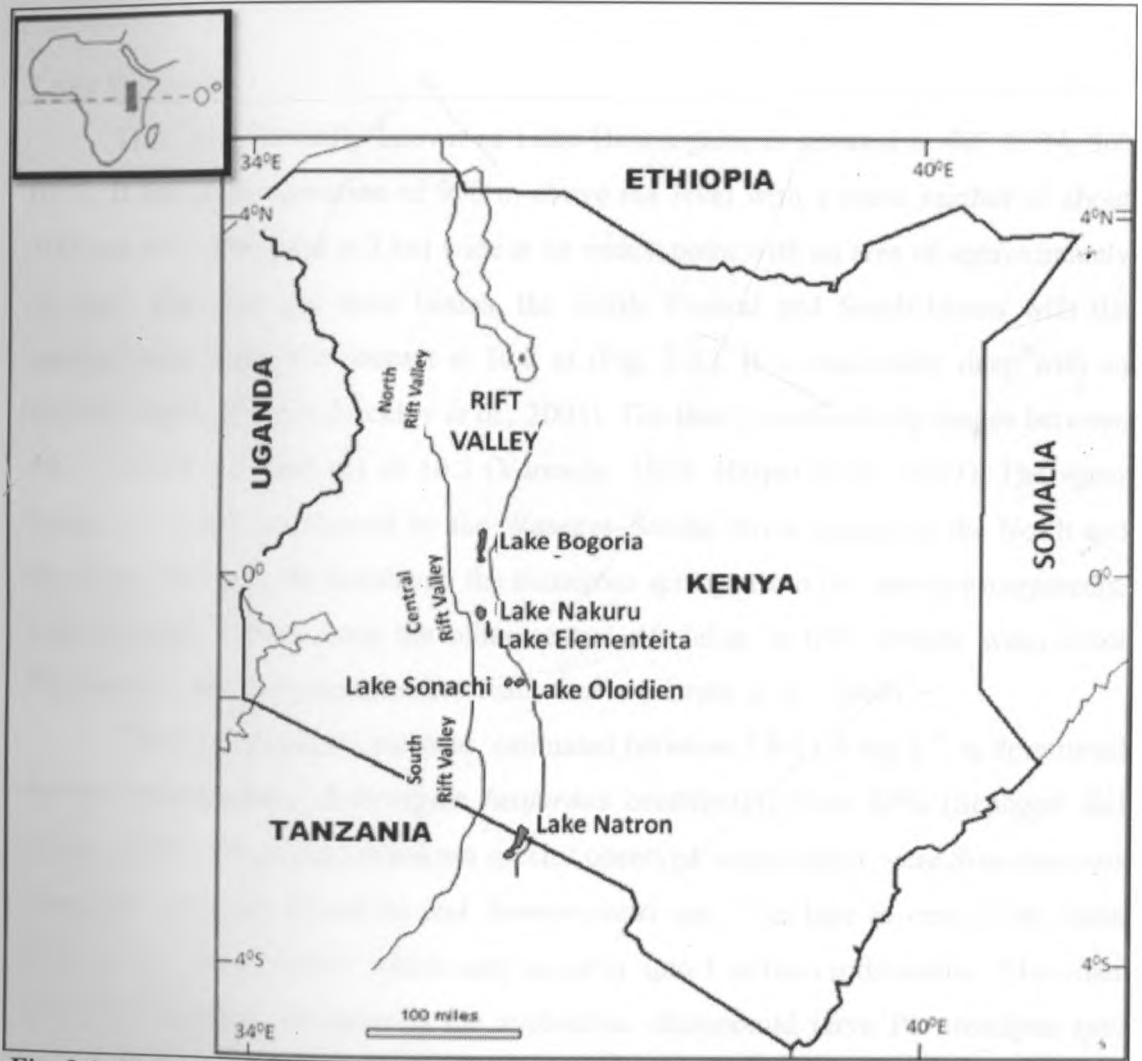


Fig. 3.1: The location of study lakes in the Eastern Rift Valley.

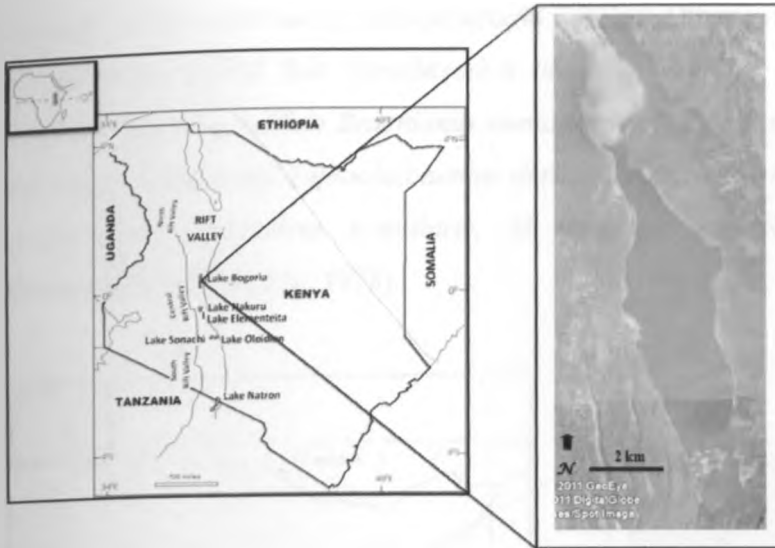
### 3.1.1. Description of study sites

The lakes of the East African Rift valley system have undergone periods of erratic extreme changes in water level and environmental conditions during the last hundreds of years (Verschuren, 2001) and on geological time scales (Trauth *et al.*, 2005). In spite of these lakes being in close proximity to each other they are stunningly different in hydrology, ecology and range in depth from less than 0.15 m at Lake Natron to a maximum depth of 10.2 m at Lake Bogoria. The phytoplankton and zooplankton communities of saline lakes are poor in species diversity and unpredictable in occurrence compared to freshwater and marine ecosystems.

#### Lake Bogoria.

This lake, formally known as Lake Hannington, is situated at 00° 15'N, 36° 05'E. It lies at an elevation of 990 m above sea level with a mean rainfall of about 500 mm yr<sup>-1</sup>. This lake is 3 km wide at its widest point with an area of approximately 34 km<sup>2</sup>. The lake has three basins, the North, Central and South basins with the central basin being the deepest at 10.2 m (Fig. 3.2.). It is reasonably deep with an average depth of 5.4m (Hickley *et al.*, 2003). The lake's conductivity ranges between 40-77,000µS cm<sup>-1</sup> and pH of 10.3 (Vareschi, 1978; Harper *et al.*, 2003). The water budget is mostly maintained by the Waseges–Sandai River system in the North and the Emsos River in the South and the numerous springs from the nearby escarpments. Lake Bogoria differs from the other central rift lakes in that, despite water level fluctuations, the lake remains chemically stable (Harper *et al.*, 2003).

The phytoplankton biomass, estimated between 5.8-51.4 mg L<sup>-1</sup>, is dominated by the cyanobacteria *Arthrospira fusiformis* contributing over 80% (Schagerl and Oduor, 2008). Other phytoplankton species observed occasionally were *Synechocystis spp.*, and *Navicula halophila* and *Keratococcus spp.* The lake is one of the main feeding sites for *P. minor*, which may occur in upto 1 million individuals. The other important primary consumer is the zoobenthic chironomid larve *Paratendipes spp.* with a density of 4×10<sup>4</sup> organisms m<sup>-2</sup> (Harper *et al.*, 2003).



**Fig. 3.2:** The location and image of Lake Bogoria in the Eastern Rift Valley (Image from Google Earth 2011).

### Lake Nakuru

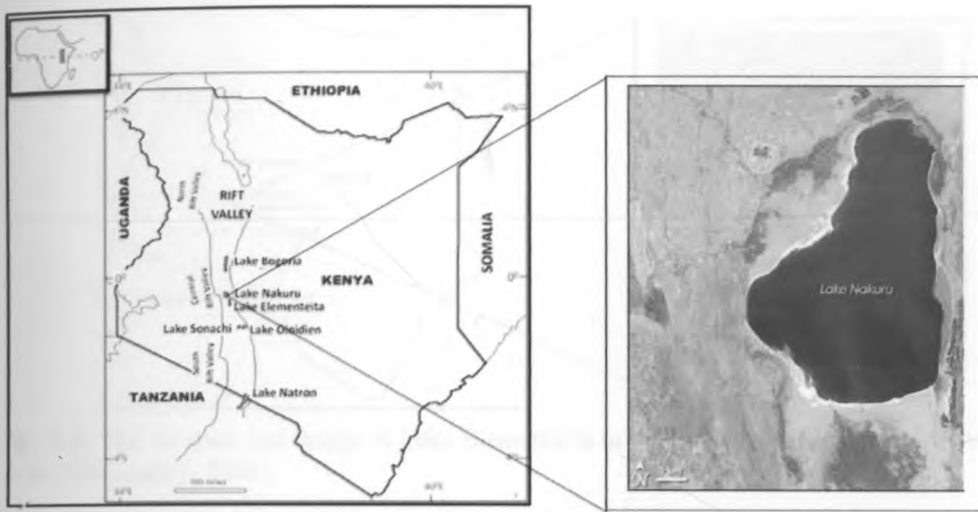
The lake is located at  $0^{\circ} 24'S$ ,  $36^{\circ}05'E$  and lies at an altitude of 1,759m above sea level with a mean rainfall is about  $800 \text{ mm yr}^{-1}$ . The lake's surface area varies from  $5 \text{ km}^2$  to  $45 \text{ km}^2$  with a mean depth of 2.3 m (Ndetei and Muhandiki, 2005). The conductivity is highly variable and ranges between 11–160,000  $\mu\text{S cm}^{-1}$  (Verschuren, 2004) with a pH of approximately 10.5 (Vareschi, 1978) (Fig. 3.3.). The water budget is maintained by recharge from four seasonal inflowing rivers namely, Rivers Njoro, Nderit, Makalia and Naishi, in addition to Baharini Springs and other springs along the eastern shoreline which are perennial. Treated wastewater from Nakuru Town also drains into the lake.

In 2003-2005 the net primary productivity was  $10.7 \text{ g O}_2 \text{ m}^{-2} \text{ day}^{-1}$  while the algal biomass ranged between  $40.3\text{-}77.9 \text{ mg L}^{-1}$  with *A. fusiformis* contributing 60% of the phytoplanktonic biomass (Oduor and Schagerl 2007). Other species of importance as flamingo food at the lake include cyanobacteria *Anabaenopsis arnoldii*, *Anabaenopsis abijatae* and *Anabaena spp.* and diatom species *Navicula halophila*, *Navicula elkab*, *Nitzschia frustulum*, *Nitzschia sigma* and *Anomoeoneis sphaerophora*. Other phytoplankton at the lake are *Synechococcus spp.* and the chlorophytes *Monoraphidium minutum* and *Chlorococcum spp.*

Lesser flamingo populations at the lake have been monitored for long and though erratic the lake can host 1 million flamingos which are the main primary



consumer of the dominating cyanoabacteria species. Other primary consumers are alkaline-water cichlid fish *Sarotherodon alkalicus graham*, a calanoid copepod *Lovenula Africana*, rotifers *Brachionus dimidiatus*, *B. plicatilis*, *Hexarthra jenkiniae* and larval chironomids *Leptochironomus deribae*, *Tanytarsus horni* and water bugs *Anisopsvaria*, *Micronecta scuteilaris*, *M. jenkiniae* and *Sigara hieroglyphica kilimandjaronis* (Vareschi, 1978).



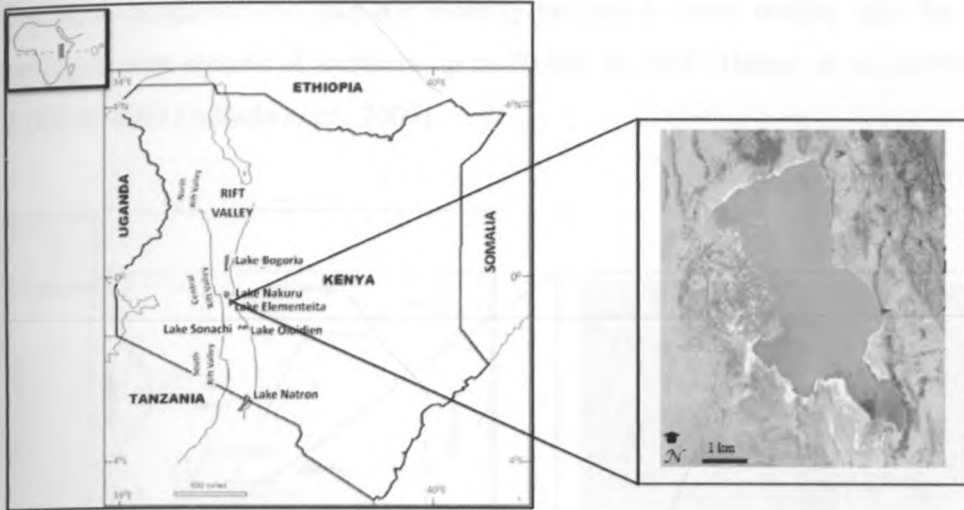
**Fig. 3.3:** The location and image of Lake Nakuru in the Eastern Rift Valley (NASA Earth Observatory, 2008).

### Lake Elementeita

It is located 0°27' S, 36°23' E. Most of the lake falls within an elevation of 1800m above sea level. The area of the lake is approximately 19 to 21 km<sup>2</sup>. The lake is quite shallow with a mean depth of about 1.2 m and a drainage basin of about 500 km<sup>2</sup> (Fig. 3.4). The lake's conductivity is between 11.9– 25,000 μS cm<sup>-1</sup> with a pH of 9.8 (Vareschi, 1978). Water supplies come primarily from three inflowing rivers, Chamuka, Mbaruk and Kariandusi Rivers and warm springs on the southern lakeshore.

Elementeita has a similar assortment of phytoplankton as Nakuru. *A. fusiformis* contributes < 50% of the phytoplanktonic biomass *Anabaenopsis arnoldii*, *A. abijatae*, *Synechococcus spp.*, and *Anabaena spp.* are also present. Common diatom species include *Navicula halophila*, *N. elkab*, *Nitzschia frustulum*, *N. sigma* and *Anomoeoneis sphaerophora* and the chlorophyte species *Monoraphidium*

*minutum*, *Chlorococcum spp.* but also *Keratococcus spp.* (Oduor and Schagerl, 2007). The lake is one of the major feeding sites of lesser flamingos as the lake can support a bloom of cyanobacteria during high lake levels while during low lake levels it supports benthic diatoms which are presumed to also nourish the birds.



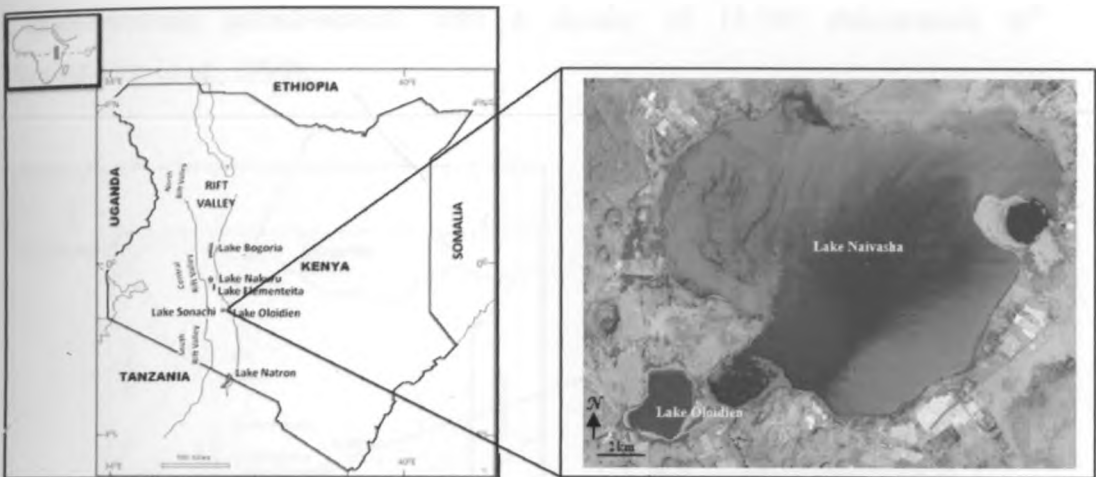
**Fig. 3.4:** The location and image of Lake Elementeita in the Eastern Rift Valley (NASA Earth Observatory, 2008).

### **Lake Oloiden**

This lake is located  $0^{\circ}50'S$   $36^{\circ}17'E$  at an altitude of 1887 m a.s.l. It is small with an area of  $5.5 \text{ km}^2$  and shallow with a mean depth of 5.6 m (Harper and Mavuti, in McClanahan and Young, 1996), while the main inflow is from precipitation. Formerly, Lake Oloiden was a bay of Lake Naivasha but by 1984 it had become completely separated from Lake Naivasha due to decreasing lake levels (Fig. 3.5). It gradually became eutrophic (Lyngs, 1996) and has recently progressed towards a hypereutrophic alkaline-saline state. Information obtained from the sediment study of the lake by Verschuren *et al.* (2004) shows that the lake's conductivity has been changing through out its recent history depending on lake levels. From  $12,700 - 13,600 \mu\text{S cm}^{-1}$  in the mid 19<sup>th</sup> century,  $320-431 \mu\text{S cm}^{-1}$  in 1929-1931 when Oloiden was connected to Lake Naivasha,  $6500 \mu\text{S cm}^{-1}$  between 1946- 1957 during a period of low lake level, and  $\sim 1200 \mu\text{S}\cdot\text{cm}^{-1}$  in 1991 after the separation.

The lake's ecology has also changed becoming more of an alkaline-saline lake. Between 2001 and 2005 there was a shift in the dominating phytoplankton from

dominance of coccoid Chlorophyceae to cyanobacteria *Arthrospira fusiformis* and *Anabaenopsis elenkini* (Ballot *et al.*, 2009). The chironomid community has been shifting with changing lake conductivity from fresh water species *Tanytarsus horni* and *Dicrotendipes septemmaculatus* to salt tolerant species *Microchironomus deribae*, *Kiefferulus disparilis* (Verschuren *et al.*, 2004). Due to colonization of the lake by *A. fusiformis* the lake has recently become a major feeding lake for lesser flamingos with reports of numbers up to 70,000 in 2006 (Harper *et al.*, 2006) and 25,000 in 2007 (Adhola *et al.*, 2009).



**Fig. 3.5:** The location and image of Lake Oloidien and Lake Naivasha from which it was separated (Image from Google Earth 2011).

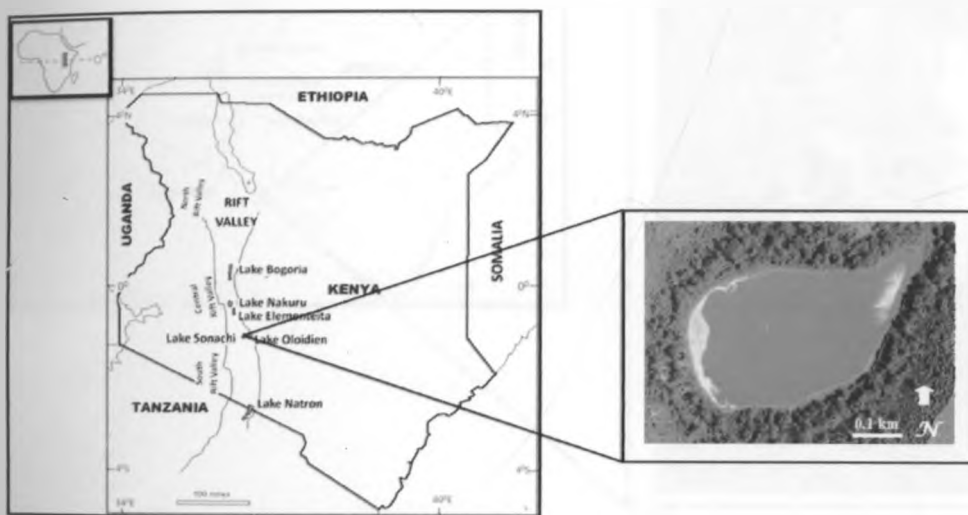
**Lake Sonachi**

It is a small crater lake located approximately 0°47'S 36°16'E at 1,884 m above sea level in the Rift Valley in Kenya. It is shallow with a depth of about 4.25 m recorded in 1993(Fig. 3.6). Being a crater lake there are no rivers flowing into the lake, the main inlets are rainfall and ground water through subsurface flow from the large freshwater Lake Naivasha which is situated 3 km away (Verschuren *et al.*, 1999). Since the beginning of the nineteenth century, lake levels have fluctuated frequently and reached maximum depths of 18 m during a high stand at the end of 19th century, and near desiccation in 2003.

Photosynthetic activity ranged between 150-870 mg O<sub>2</sub> m<sup>-2</sup> h<sup>-1</sup> in the 1970s (Melack, 1981) with the phytoplankton community dominated by cyanobacteria *Synechococcus bacillaris* which contributed 53% of algal biomass, *Lyngbya*

*limnetica*, *Synechocystis aquatilis*, *Spirulina laxissima* and *Spirulina platensis* (which is now *Arthrospira fusiformis*). Green algae of species *Chlorella spp.* and *Oocystis parva* and diatom species *Nitzschia spp.*, *Navicula cryptocephala*, *Anomoeoneis sphaerophora*, *Craticula eklab*, have been recorded at the lake (Verschuren *et al.*, 1999).

The zooplankton community is composed of rotifer species of *Brachionus dimidiatus*, *B. Plicatilis* that dominated at different times in recent history of the lake (Epp *et al.*, 2010). The copepod *Paradiaptomus africanus* has been recorded at the lake and chironomid larvae of *Kiefferulus disparilis*, *Microtendipes sp.* and *Cladotanytarsus pseudomancus* with a density of 13,500 chironomids m<sup>-2</sup> (Verschuren *et al.*, 1999).

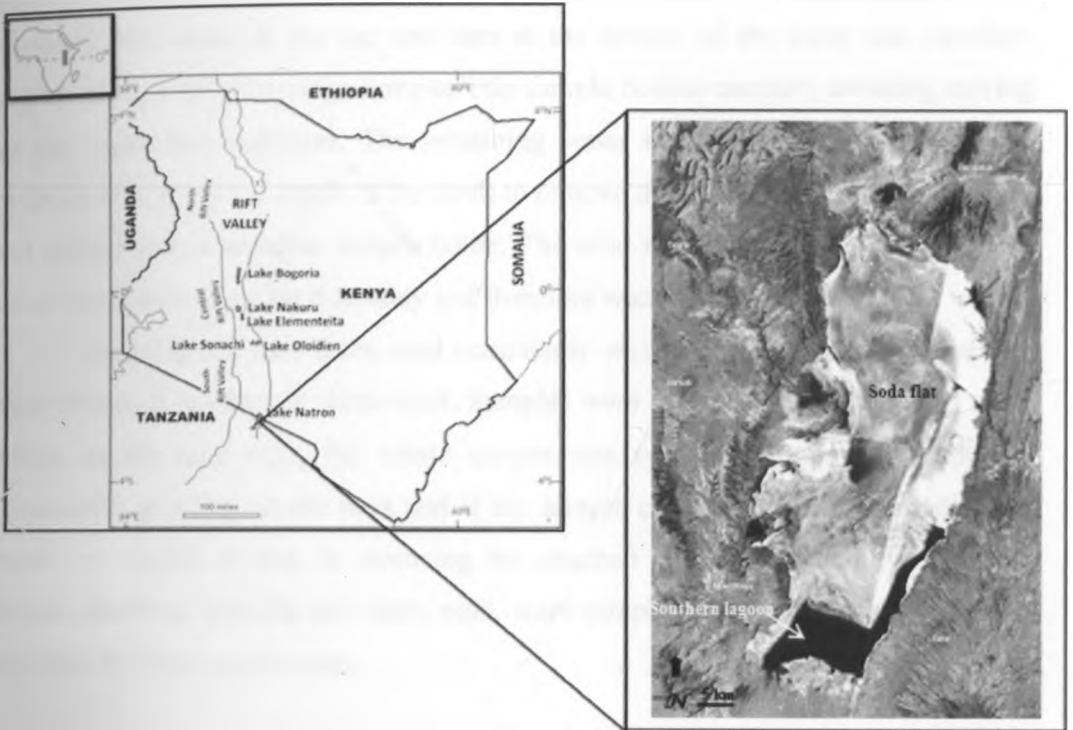


**Fig. 3.6:** The location and image of Lake Sonachi in the Eastern Rift Valley (Image from Google Earth, 2011).

### Lake Natron

The lake is located 2° 09' – 2° 36' S and 35° 54' - 36° 06' E at an altitude of 610 m a.s.l. It is a large basin with an area of 1,039 km<sup>2</sup> and shallow with depth ranging from a few centimeters to 2 m (Kasule *et al.*, 1993), the pH varies between 9 and 10. The lake is made up of several lagoons, the largest is the southern lagoon where the study was carried out (Fig. 3.7). Natron's water budget is mainly from the permanent Ewaso Ngiro River which rises from the Mau escarpment in Kenya. Seasonal rivers from the Loita hills, Loliondo Mountains, Ngorongoro highlands and Mt. Ngelai in Tanzania drain into the lake. Springs that occur all around the lake also

contribute about a quarter of the water budget and maintain the lagoons during times of low rainfall. There is insufficient information regarding the biological characteristics of the lake. Only one alkaline water tilapia species *Oreochromis alcalica* is present and it sustains the pelicans and storks. The lake is significant to flamingos as it is the only viable breeding site for the East African *P. minor* population (Brown, 1973).



**Fig. 3.7:** The location and image of Lake Natron in the Eastern Rift Valley. The red arrows show some of the other smaller lagoons at the lake. (Image from NASA Earth Observatory, 2011).

## 3.2. Materials and Methods

### 3.2.1. Determination of composition of lesser flamingo food

#### 3.2.1.1. Sample collection

A Gilson corer (Plate 3.1) was used to collect samples for assessment of planktonic standing crop, epipellic standing crop and the lake shore mud standing crop randomly on transects running from the lake shore to the maximum safely wadeable depth. The depth of the water in the corer was measured by ruler and the depth recorded. The water at the top and then at the bottom of the corer was carefully sucked using a 60 ml syringe sampler into sample bottles carefully avoiding mixing up the water and sediment. The remaining water in the corer was then swirled (without mixing up too much of the mud) to remove the attached epipellic community and poured into a separate sample bottle. The mud remaining at the bottom of the corer was unnecessary for this study and therefore was discarded.

Sampling for lake shore mud community was done at lakes where *P. minor* were observed feeding on shore mud. Samples were collected at water depths of 0-0.9cm on the lake edge, the whole sample was swirled to remove the epipellic community growing on the mud and if the sample collected had no water distilled water was added to help in removing the attached epipellic community. Different layers; plankton, epipellic and shore mud, were sampled for comparison of the food resource in these communities.



**Plate 3.1:** Gilson's corer showing the position of (a) the planktonic community, (b) the epipellic community in the water-mud interface and (c) sediment.

### 3.2.1.2. Species identification

The samples collected as described above were examined under the microscope for species present. A Sedgewick-Rafter cell counting chamber with a capacity of 1ml and evenly divided into 1000 squares was used to count the organisms present. Individuals of each species occurring in a minimum of 30 squares were counted. The number of organisms in 1 ml of the sample was calculated. For the cyanobacteria species *Arthrospira fusiformis* that is coiled, the number of coils was counted on ten colonies chosen at random.

Photographs of the species were taken using a Canon digital camera aligned with the eyepiece lens of the microscope to assist in identification. Diatom species were initially collectively counted as diatoms without counting per species later electron microscopy was used to identify species of diatoms. The species of microorganisms were identified at least to genus level using identification keys by Bellinger (1992); Barber and Haworth (1994) and Boney (1989) and with the help of experts and supervisors. The main output was a species check list.

### 3.2.2. Assessment of available food

#### 3.2.2.1. Photosynthetic pigment analysis

Samples collected as described above (refer to section 3.2.1.1.) were measured for chlorophyll *a* (Chl *a*) content. The analysis of the photosynthetic pigment was used to determine the standing crop of cyanobacteria and algae as a measure of the food available for the lesser flamingo.

Subsamples of 50 or 100 ml of the water samples and 10 ml of lake shore mud samples were filtered through Whatman GF/C filter papers with a pore size of 0.47  $\mu\text{m}$  to retain the algae. Each of the filter papers with the algae was ground in a mortar with a pestle in 5 ml of 90% acetone with  $\text{MgCO}_3$  and a pinch of sand. When the contents were finely ground they were put in 15 ml centrifuge tube and filled to the 15 ml mark with 90% acetone. The samples were centrifuged in a hand-centrifuge at maximum speed, which was about 5000 r.p.m. The supernatant was decanted cautiously into a 1 cm path-length glass cuvette and the absorbance read against an acetone blank in a spectrophotometer at wavelengths of 750, 663 and 665 nm. Two drops of 1 M HCl were added to the sample in the cuvette to break down Chl *a* to phaeophytin *a* and the absorbance was read again at the same wavelengths.

Absorbance was converted to concentrations of Chl *a* using the formulae below (Sartory and Grobbelaar, 1984).

### 3.2.2.2. Calculation of the standing crop of available food

- a) Chl *a* absorbance was corrected using the equation that is below based on unpublished data by Nic Pacini.

$$\text{Corrected absorbance} = 5.3 \times \text{absorbance} - 0.026$$

- b) The chlorophyll *a* was calculated in  $\mu\text{g cm}^{-2}$  using the following equation

$$\mu\text{g Chl } a \text{ cm}^{-2} = \left( \frac{26.7(E665_1 - E665_2) \times V_e}{A_s \times l} \right)$$

Where;

26.7 = absorbance correction factor

$E665_1$  = corrected absorbance at 665 nm before adding HCl = (abs 665 – abs 750)

$E665_2$  = corrected absorbance at 665 nm after adding HCl = (abs 665 – abs 750)

$V_e$  = represents volume of acetone used in the extraction (ml)

$A_s$  = area of sample ( $\text{cm}^2$ )

$l$  = path length of cuvette (cm)

c)  $\mu\text{g Chl } a \text{ m}^{-2} = \mu\text{g Chl } a \text{ cm}^{-2} \times 10,000$

The maximum and minimum carbon value calculated using Falkowski (1985) conversion value for Chl *a* to carbon value.

- d) Maximum standing crop (SC max)

$$\text{mg C m}^{-2} = \frac{\text{Chl } a \mu\text{g cm}^{-2}}{1000 \times 0.003}$$

Where;

L = volume filtered (litre)

0.003 = conversion of Chl *a* to maximum Carbon value

1000 = conversion from  $\mu\text{g}$  to  $\text{mg}$



e) Minimum standing crop (SC Min)

$$\text{mg C m}^{-2} = \frac{\text{Chl } a \text{ } \mu\text{g cm}^{-2}}{1000 \times 0.1}$$

Where;

L = volume filtered (litre)

0.003 = conversion of Chl *a* to maximum Carbon value

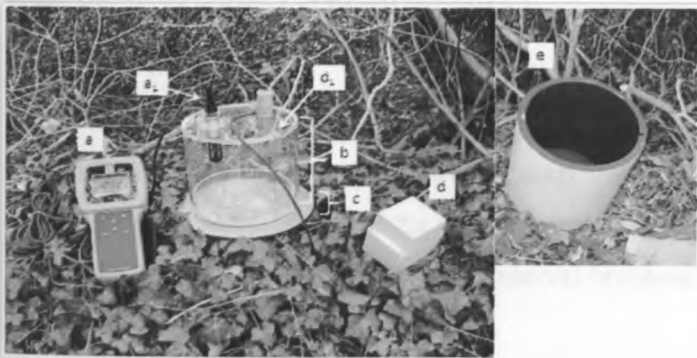
0.1 = conversion of Chl *a* to minimum Carbon value

1000 = conversion from  $\mu\text{g}$  to mg

### 3.2.3. Assessment of primary productivity

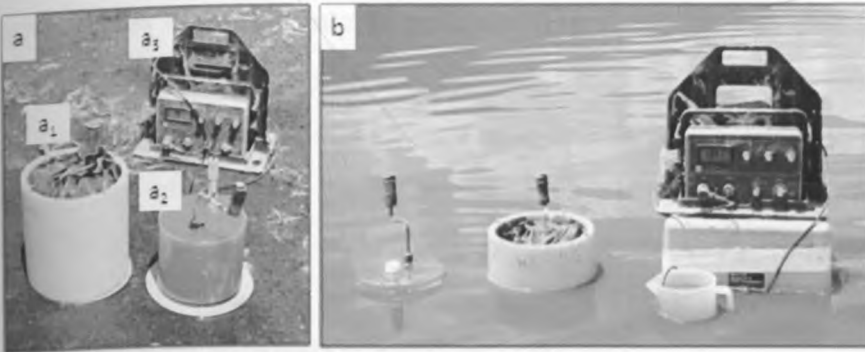
#### 3.2.3.1. Exposure experiment method

Photosynthetic activity was determined by measuring the concentration of Oxygen ( $\text{O}_2$ ) in the water using light and dark bottle technique. This technique measures the variations in the concentration of oxygen under different experimental conditions to infer net primary production (Vollenweider, 1974). The equipment that was used in measuring the rate of photosynthetic productivity is shown in Plate 3.2. It is composed of two transparent perspex metabolism chambers that fit tightly to a heavy bottom plate and one dark cover. Each chamber has an opening at the top that can fit an oxygen meter probe, and has a capacity of 4 litres and is equipped with a stirrer that (is run by a motor or hand) which ensures uniformity in the ambient conditions within the chamber.



**Plate 3.2:** (a) Oxygen and temperature meter and (a<sub>1</sub>) probe, (b) metabolism chamber top, (c) metabolism chamber base plate, (d) motor and (d<sub>1</sub>) stirrer, (e) metabolism chamber dark cover.

Two metabolism chambers were used for each exposure; one of the chambers was covered with the dark cover to become the dark chamber, while the other one was left uncovered to become the light chamber (Plate 3.3).



**Plate 3.3:** (a) Set up of productivity exposure experiment, (a<sub>1</sub>) dark metabolism chamber, (a<sub>2</sub>) light metabolism chamber and (a<sub>3</sub>) oxygen meter. (b) Exposure experiment set up *in situ*.

Due to exposure to sunlight in the light chamber, photosynthesis was expected to take place, releasing oxygen as a by-product of photosynthesis and dissolved oxygen (DO) in the light chamber was expected to increase. Respiration by the whole community in the dark chamber was expected to result in a decrease in the dissolved oxygen in the dark chamber. The change in DO concentration in the chambers was measured using a YSI Professional Plus portable dissolved oxygen meter. The net primary production (NPP) was arrived at by deducting oxygen consumed by the respiration of the whole community (Resp., dark chamber) in the dark chamber from the gross primary productivity (GPP).

$$(F_l - F_d) - (I_n - F_d) = (F_l - I_n)$$

$(F_l - F_d)$  = the GPP per unit volume over the time interval involved

$(I_n - F_d)$  = the Resp. per unit volume over the time interval involved

$(F_l - I_n)$  = the NPP per unit volume over the time interval involved

Where;

$I_n$  = the initial concentration of dissolved oxygen in the both the light and dark chambers

$F_l$  = final concentration of dissolved oxygen in the light chamber

$F_d$  = final concentration of dissolved oxygen in the dark chamber

The productivity which was measured as change in dissolved oxygen ( $\Delta \text{ mg O}_2 \text{ L}^{-1} \text{ hr}^{-1}$ ) was converted to change in organic Carbon ( $\Delta \text{ mg C L}^{-1} \text{ hr}^{-1}$ ) using the relationship 1.0 mg of oxygen is equivalent to 0.30 mg of carbon to enable easy comparison with what is found in other ecosystems. The volumetric rate of primary productivity was converted into an area rate ( $\text{mg C m}^{-2} \text{ hr}^{-1}$ ) by dividing it by the height of the metabolism chamber.

Two primary production exposure experiments were carried out *in situ* on the shallow shores of the lake under investigation. The first exposure experiment measured both the planktonic and epipellic primary productivity while the second one measured only the planktonic primary productivity. This was to enable estimation of epipellic productivity by simple subtraction of the second measurement from the first one. The exposure experiments carried out were incubated between 10:00 and 14:00 hours because photosynthetic activity was expected to be highest at these hours.

### 3.2.3.2. Exposure experiment for measurement of planktonic and epipellic primary productivity

The base plates of both the light and dark metabolism chambers were filled with lake sediment and left in the shallow area of the lake at a depth of not more than 20 cm for approximately 24 hours. This allowed time for the sediment and the associated epipellic microorganism to settle. After 24 hours, the tops of the metabolism chambers were fitted carefully without disturbing the sediment and cautiously filled with lake water and one of them covered with the dark cover. This set up was meant for the measurement of the primary productivity of both the planktonic and epipellic microbial communities, '**total primary productivity**'. Once the equipment was set, the initial oxygen concentration in both chambers was measured using a dissolved oxygen meter (YSI model 58). The exposures were allowed a maximum time of one hour. This caution was taken because if they stayed longer than 1 hour, the oxygen concentration in the chambers could reach super-saturation level, which would be too high to be measured. Oxygen concentration within the chambers was measured repeatedly every 10 minutes. At the same time the concentration of oxygen in the open water was recorded. After 1 hour, the experiment was ended and the chambers emptied of all sediment and water.

### 3.2.3.3. Exposure experiment for measurement of planktonic primary productivity

The second exposure experiment was of lake water only this was to measure the productivity of the plankton microbial community. The exposures ran for 1 hour and oxygen concentration measures were repeated every 10 minutes. Samples of water and/or sediment were collected from each chamber at the end of each exposure.

### Calculations for primary productivity and respiration of epipelagic community

$$a. \text{ Epipelagic primary productivity} = \text{Total primary productivity} - \text{planktonic primary productivity}$$

$$b. \text{ Epipelagic respiration} = \text{Total respiration} - \text{planktonic respiration}$$

Where:

Total primary productivity = primary productivity of both the planktonic and epipelagic community

Total respiration = respiration of both the planktonic and epipelagic community

### 3.2.4. Statistical analysis

The data were analyzed using the SPSS 10.0 programme. Most of the data were not normally distributed and therefore the non-parametric statistical tests were preferred. The Mann-Whitney U test was used together with Spearman's rank-order correlation, to test for significant relationships between the standing crop, water depth, and lesser flamingo population.

## CHAPTER FOUR

### RESULTS

#### 4.1. General observations and explanation of terms.

1. The cyanobacteria within the water column is referred to as the **planktonic community**. During the study it was observed that at Lake Bogoria planktonic *Arthrospira fusiformis* often sank and formed a film of concentrated algal mass just above the bottom in shallow sheltered bays, upon which lesser flamingos fed on by diving down and submerging most of their bodies into the water.
2. At Bogoria there was no true epipellic community but sunken *A. fusiformis*, which is referred to as **sedimented *A. fusiformis***.
3. In 2009, Lake Elementeita and Lake Natron's lagoon and marsh were very shallow with depth not exceeding 10cm. Light was able to penetrate to the bottom and benthic diatoms growing on the sediment surface formed a 'mat' which is referred to as the **epipellic community**.
4. Lakes Elementeita and Natron did not have a truly planktonic community for the study period. The epipellic community would become suspended in the water by the lesser flamingo trampling activities. The material suspended into the water column is referred to as the **suspended epipellic community**. This was also observed in some few cases after a storm at Lake Natron.
5. Lesser flamingos were occasionally seen feeding on the mud at the lake shore. The biomass on the shore mud was measured and is referred to as the **lake shore mud community**.

#### 4.2. Taxonomic composition

Four groups of taxa were encountered which were; Cyanobacteria, Bacillariophyceae, Protozoa and Rotifera. A taxon was considered dominant if it contributed 40% of the total count within its category or it was at least twice as numerous as the second most numerous species.

Table 4.1 summarizes the species that contribute to the flamingo food and Plate 4.1 shows some of these species. *Arthrospira spp.* was found in most lakes and achieved dominance at lakes Bogoria, Oloidien and Sonachi. Although *Anabaena spp.*, *Oscillatoria spp.*, *Lyngbya pseudospirulina* and *Spirulina spp.* were never dominant they were found in nearly all the lakes. *Anabaenopsis magna* and *Abijatae*

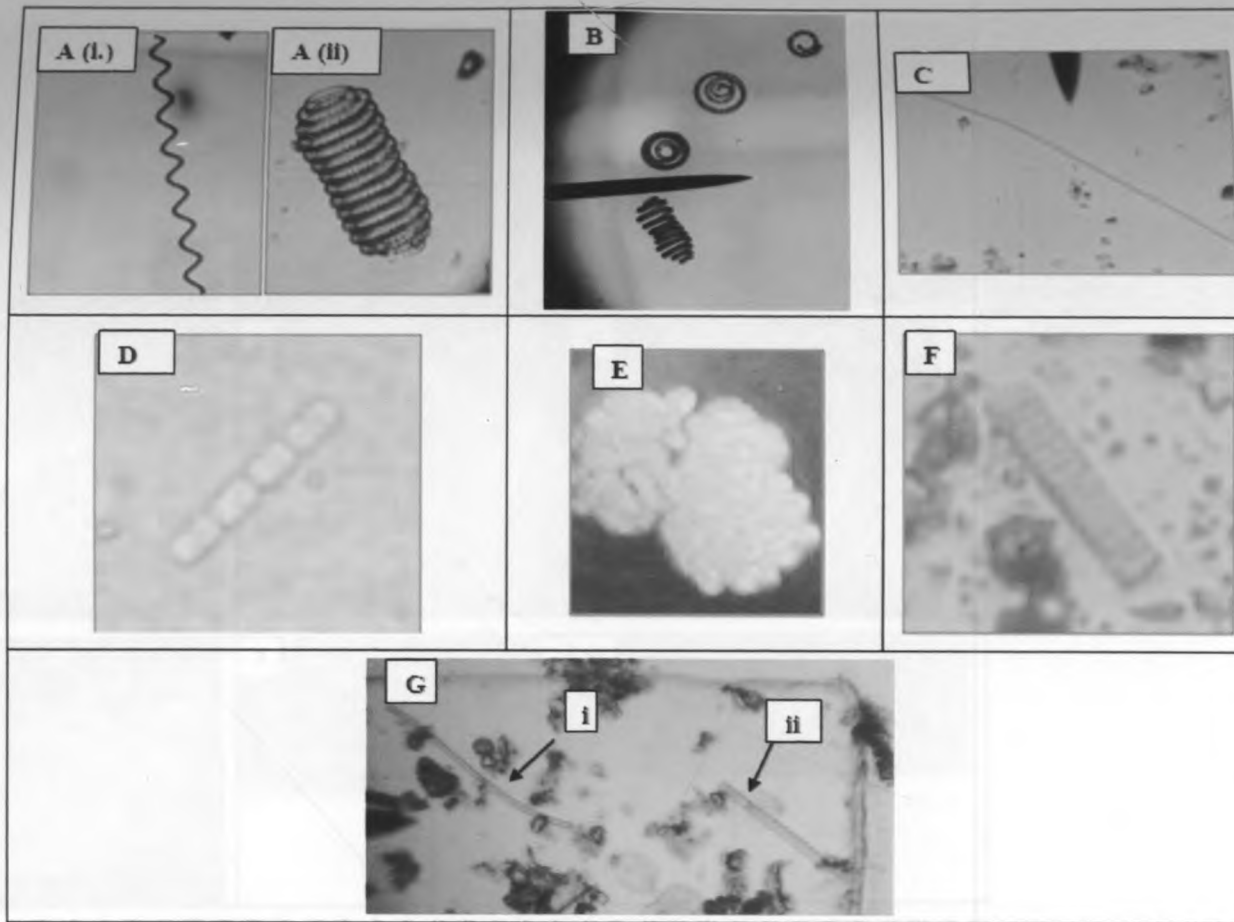
*spp* occurred at Nakuru. *Arthrospira spp.* and *Anabaenopsis spp.* were characteristically missing from Lake Natron.

Presence of diatom species at a lake is denoted as 'p' in the table because diatom species were initially counted collectively as diatoms without counting per species and later identified from electron microscopy therefore abundances for each species is lacking. Several *Navicula spp.*, *Amphora spp.* and *Melosira spp.* were observed in all the lakes. Interestingly, *Sellaphora spp.* for which there is no previous record for saline lakes in Kenya was found only at Bogoria.

Among the protozoa *Frontonia spp.* was the most common and was present in all the lakes but in some it achieved dominance such as at lakes Bogoria and Sonachi. *Condylostoma spp.* was present in most but never achieved dominance. *Amoeba spp.* and *Campanella spp.* were recorded at Lake Oloidien only. *Dileptis spp.* and an unidentified species were found in Bogoria only. Lake Natron was very poor in protozoa with only *Frontonia spp.* Recorded while no rotifer species were seen. Plates 4.1, 4.2 and 4.3 show some of the species encountered during the study.

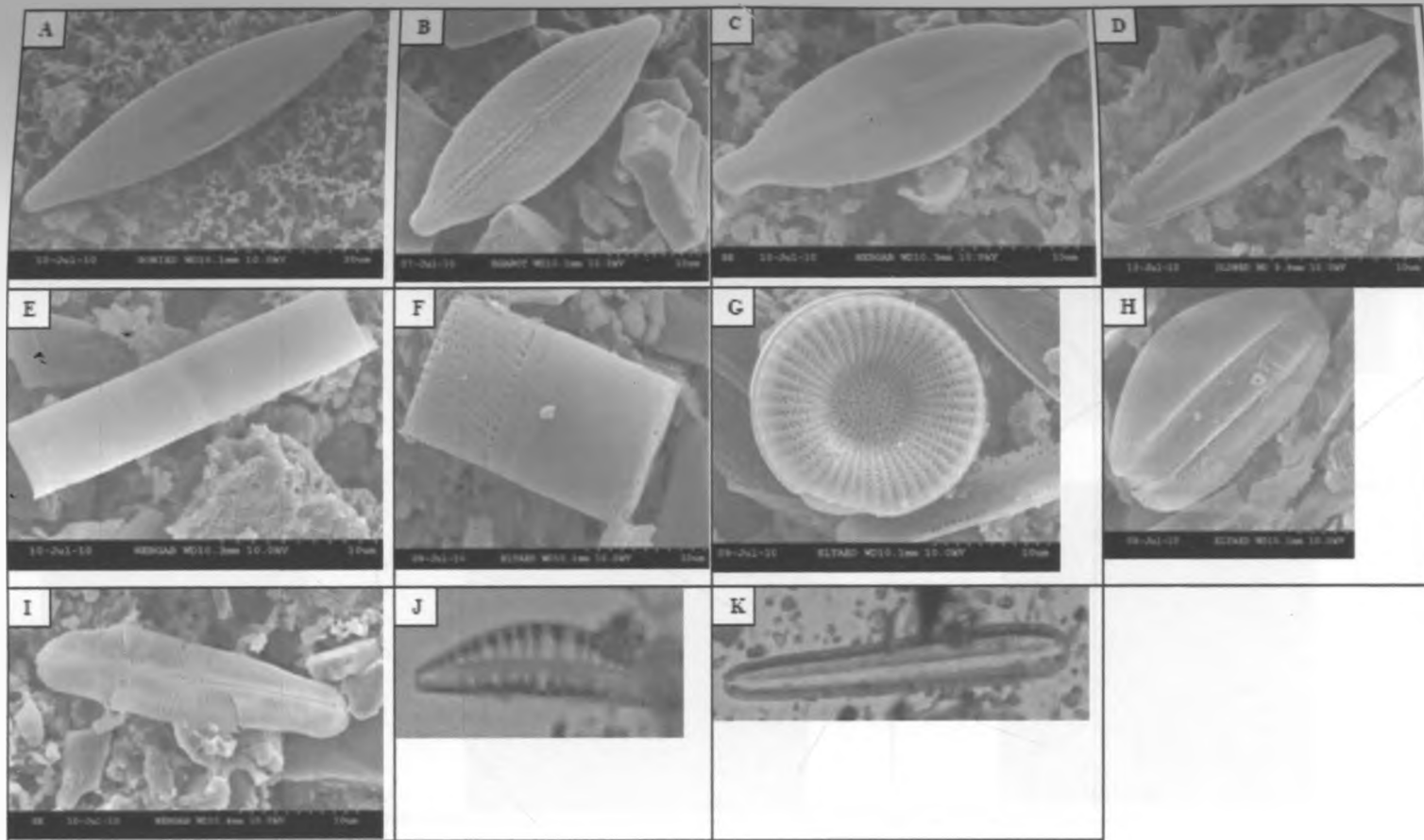
**Table 4.1:** Species composition of the microbial community at the study lakes during the study period ((3) dominant, (2) abundant, (1) present, (-) absent (p) present)

Taxon	Bogoria	Nakuru	Natron	Elementeita	Oloidien	Sonachi
<b>Cyanobacteria</b>						
<i>Arthrospira fusiformis</i>	3	2	-	1	3	3
<i>Arthrospira spp.</i>	-	1	-	-	3	-
<i>Anabaenopsis magna</i>	-	3	-	1	-	-
<i>Anabaenopsis abijatae</i>	-	3	-	1	-	-
<i>Anabaena spp.</i>	-	1	1	1	1	-
<i>Oscillatoria spp.</i>	1	2	1	1	1	1
<i>Lyngbya pseudospirulina</i>	-	1	1	1	1	-
<i>Spirulina subsalsa</i>	-	-	1	1	2	1
<b>Bacillariophyceae</b>	1	3	3	3	2	2
<i>Navicula spp.</i>	p	p	p	p	p	p
<i>Amphora spp.</i>	p	p	p	p	p	p
<i>Melosira spp.</i>	p	p	p	p	p	p
<i>Synedra/Flagillaria spp.</i>	p	p	p	p	-	-
<i>Cyclotella spp.</i>	-	p	p	p	-	-
<i>Cymbella spp.</i>	-	p	-	p	-	-
<i>Pleurosigma spp.</i>	p	-	-	-	-	-
<i>Sellaphora spp.</i>	p	-	-	-	-	-
<b>Protozoa</b>						
<i>Frontonia spp.</i>	3	2	1	1	2	3
<i>Condylostoma spp.</i>	2	2	-	1	1	1
<i>Euplotes spp.</i>	2	1	-	-	-	-
<i>Dileptis spp.</i>	1	-	-	-	-	-
<i>Campanella spp.</i>	-	-	-	-	2	-
<i>Amoeba spp.</i>	-	-	-	-	2	-
<i>Euglena spp.</i>	-	1	-	1	-	-
Unidentified species 1	2	-	-	-	-	-
Unidentified species 2	1	1	-	-	1	-
<b>Rotifera</b>						
<i>Brachionus spp.</i>	1	2	-	2	1	2
<i>Hexarthra jenkiniae</i>	1	2	-	1	-	-

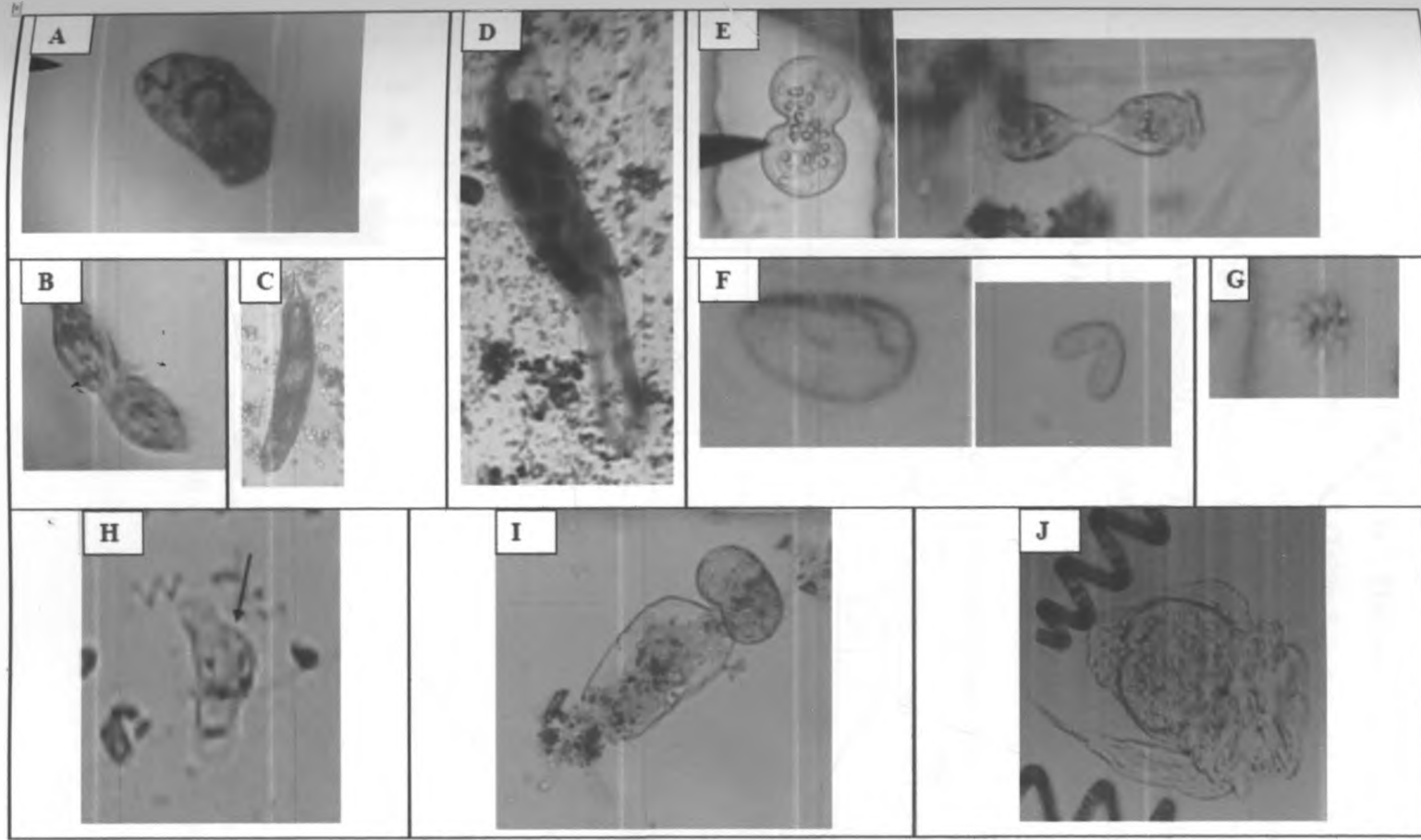


**Plate 4.1:** Cyanobacteria taxa identified (Bellinger, 1992 and Boney, 1989). A. *Arthrospira fusiformis* (i) extended and (ii) compressed. B. *Arthrospira* spp. C. *Spirulina subsalsa*. D. *Anabaena* spp. E. *Anabaenopsis abijatae*. F. *Lyngbya pseudospirulina*. G. *Oscillatoria* spp. (i) *O. limosa* (moving), (ii) *O. limnetica* (stationary).





**Plate 4.2:** Some of the Bacillariophyta species identified (Bellinger, 1992 and Barber and Haworth, 1994) . A-D. *Navicula* spp. E. & F. *Melosira* spp. G. *Cyclotella* spp. H. *Amphora* spp. I. *Sellaphora* spp. J. *Cymbella* spp. K. *Pleurosigma* spp.



**Plate 4.3:** Protozoan species (A – H) and rotifer species (I & J) (Hall, 1953). A. *Frontonia* spp. B. *Euplotes* spp. C. An Euglenophyte species present at lake Nakuru only. D. *Condylostoma* spp. E. *Campanella* spp. was seen only at Lake Oloidien. F. Unidentified species 1. G. Actinosphaeridium. H. *Amoeba* spp. only seen in Lake Oloidien. I. *Brachionus plicatilis* J *Hexarthra jenkinae*.

### 4.3. Standing crop of the various categories of *Phoeniconaias minor* food resources

#### 4.3.1. Standing crop of the planktonic and suspended epipellic food resource

The highest planktonic standing crop was recorded at Lake Bogoria in April 2009 with a mean of  $37.74 \times 10^3 \text{ mg C m}^{-2}$ . The lowest was recorded at Lake Sonachi with a mean of  $0.56 \times 10^3 \text{ mg C m}^{-2}$  (Table 4.2, Fig. 4.1). The results from Lake Bogoria show a significant reduction ( $p < 0.01$ ) in planktonic standing crop between April 2009 and August 2009 from  $37.74 \times 10^3$  to  $5.17 \times 10^3 \text{ mg C m}^{-2}$ . The suspended epipellic biomass was generally low. The highest mean value of  $4.29 \times 10^3 \text{ mg C m}^{-2}$  was significantly lower ( $p < 0.01$ ) than the highest mean planktonic standing crop (Table 4.3, Fig. 4.2).

Table 4.2: Mean ( $\pm$ se) standing crop of the planktonic community.

Site	Month/year	Max $\text{mg C m}^{-2}(\times 10^3)$	Min $\text{mg C m}^{-2}(\times 10^3)$	se( $\times 10^3$ )	n
Bogoria	Apr-09	37.74	1.13	26.85	11
	Aug-09	3.24	0.10	5.17	40
	Aug-10	5.17	0.16	6.19	18
Nakuru Lake	Apr-09	11.97	0.36	2.51	3
	Aug-09	0.90	0.03	0.41	5
Nakuru Njoro river mouth	Aug-09	1.75	0.05	1.76	18
Sonachi	Sep-09	0.56	0.02	0.23	4
Oloidien	Sep-09	3.63	0.11	1.22	9

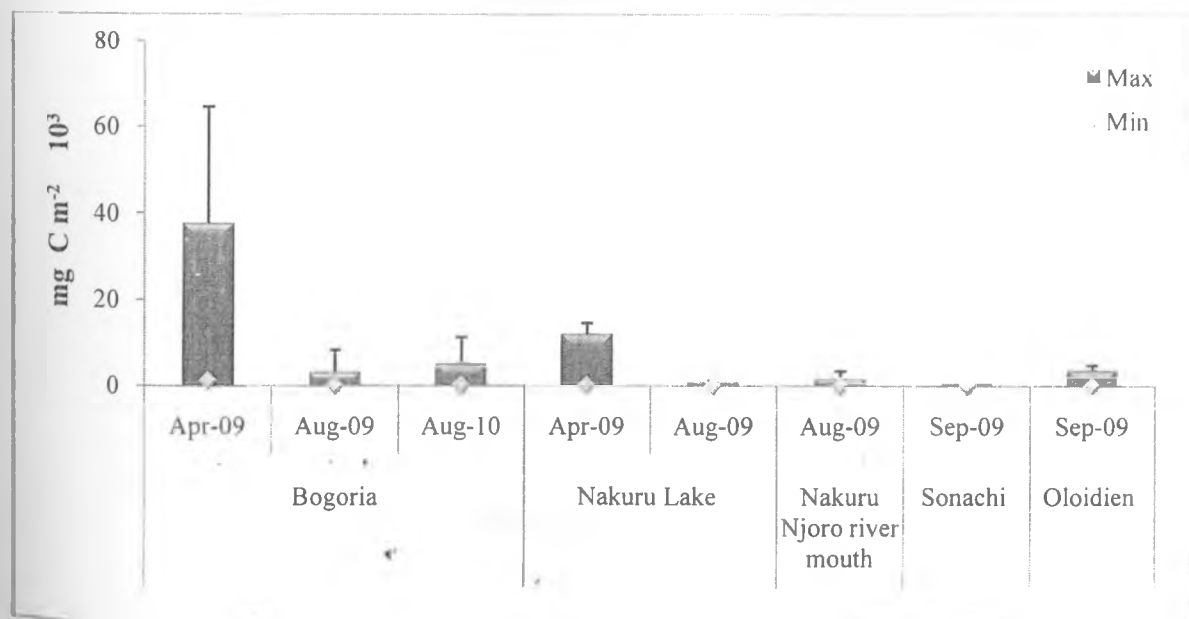
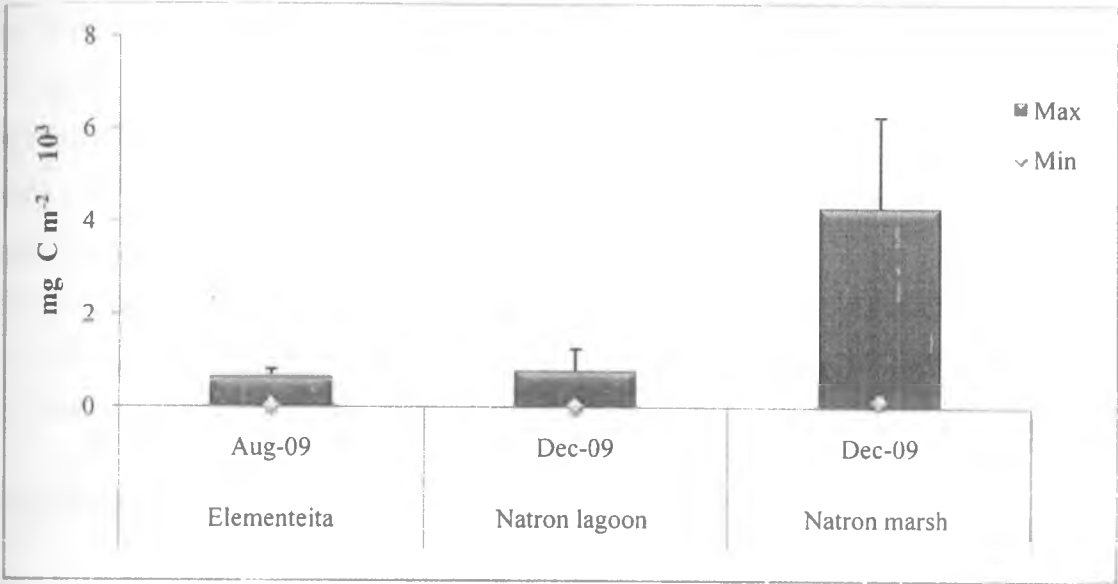


Fig. 4.1: Mean ( $\pm$  se) standing crop of planktonic food resource in the study lakes.

**Table 4.3:** Mean ( $\pm$ se) standing crop of the suspended epipellic community.

Site	Month/year	Max mg C m <sup>-2</sup> ( $\times 10^3$ )	Min mg C m <sup>-2</sup> ( $\times 10^3$ )	se( $\times 10^3$ )	n
Elementeita	Aug-09	0.66	0.02	0.16	6
Natron lagoon	Dec-09	0.79	0.02	0.48	7
Natron marsh	Dec-09	4.29	0.13	1.97	11



**Fig. 4.2:** Mean ( $\pm$  se) standing crop of suspended epipellic food resource in the study lakes.

### 4.3.2. Standing crop of sedimented *Arthrospira fusiformis* and the epipellic food resource

A decline in the sedimented *A. fusiformis* was recorded at Bogoria (Table 4.4; Fig. 4.3) and the pattern corresponded to the decline in planktonic standing crop at Bogoria observed between April 2009 and August 2010 (Fig. 4.1). Within this period, sedimented *Arthrospira* declined from  $17.24 \times 10^3$  to  $3.85 \times 10^3$  mg C m<sup>-2</sup>. The highest mean epipellic standing crop was recorded at Lake Nakuru in April 2009, ranging between  $67.67 - 30.36 \times 10^3$  mg C m<sup>-2</sup> with a mean of  $46.7 \times 10^3$  mg C m<sup>-2</sup>. The lowest recorded in Lake Nakuru was at the Njoro River mouth with a mean of  $2 \times 10^3$  mg C m<sup>-2</sup>.

Lake Natron recorded high epipellic standing crop compared to Lake Elementeita which was significantly lower than all measurements at Lake Natron's Lagoon and Marsh ( $p < 0.01$ ). Some of the standing crop measurements of the sedimented *A. fusiformis* at Lake Bogoria (April 2009), epipellic and lake shore mud value at Lake Natron's Marsh in April 2009 and December 2009 varied highly. The results suggest that values of at least  $55.9 - 95.6 \times 10^3$  mg C m<sup>-2</sup> were achieved for some measurements resulting in high mean standing crop (Table 4.5; Fig. 4.4).

Table 4.4: Mean ( $\pm$ se) standing crop of the sedimented *Arthrospira fusiformis*.

Site	Month/year	Max mg C m <sup>-2</sup> ( $\times 10^3$ )	Min mg C m <sup>-2</sup> ( $\times 10^3$ )	se ( $\times 10^3$ )	n
Bogoria	Apr-09	17.24	0.52	11.25	12
	Aug-09	4.25	0.13	5.79	40
	Aug-10	3.85	0.12	2.45	18

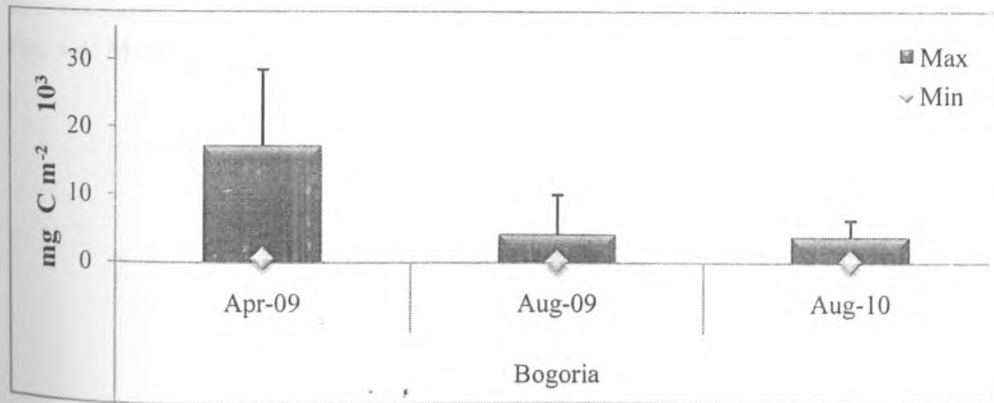


Fig. 4.3: Mean ( $\pm$  se) standing crop of sedimented *A. fusiformis* at Lake Bogoria.

Table 4.5: Mean ( $\pm$ se) standing crop of the epipellic community.

Site	Month/year	Max mg C m <sup>-2</sup> ( $\times 10^3$ )	Min mg C m <sup>-2</sup> ( $\times 10^3$ )	se ( $\times 10^3$ )	n
Elementeita	Aug-09	4.92	0.15	3.23	9
Nakuru Lake	Apr-09	46.67	1.40	13.66	7
	Aug-09	5.81	0.17	7.31	6
Nakuru Njoro river mouth	Aug-09	2.41	0.07	2.50	18
Natron lagoon	Apr-09	12.74	0.38	10.28	40
	Dec-09	20.69	0.62	6.66	27
Natron marsh	Apr-09	32.93	0.99	14.91	8
	Dec-09	26.68	0.80	22.03	12
Sonachi	Sep-09	3.12	0.09	2.72	6
Oloidien	Sep-09	3.12	0.09	2.63	8

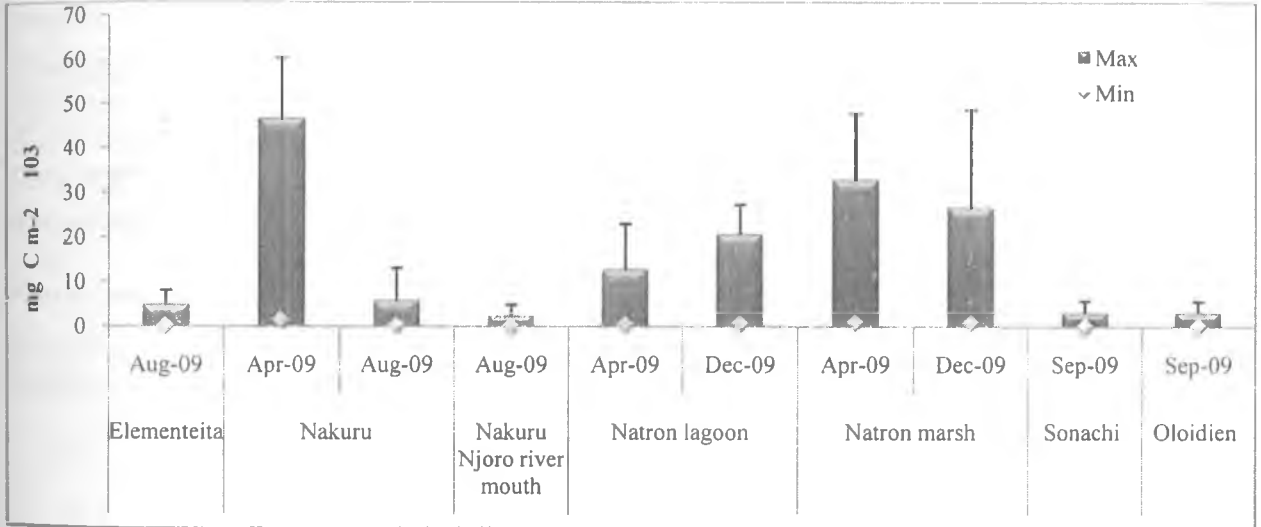


Fig. 4.4: Mean ( $\pm$  se) standing crop of epipellic food resource in the study lake.

### 4.3.3. Standing crop of lake shore mud food resource

The highest lake shore mud standing crop was recorded at Lake Bogoria in April 2009 with a mean of  $55.46 \times 10^3 \text{ mg C m}^{-2}$  and the lowest at Lake Bogoria in August 2010 with a mean of  $0.53 \times 10^3 \text{ mg C m}^{-2}$  (Fig. 4.5). Lake Elementeita's standing crop was significantly lower than that of Lake Natron's lagoon and marsh with  $p < 0.01$ .

Considering only the month of August 2009, the lake shore mud algae standing crop was highest at Lake Nakuru with a mean of  $14 \times 10^3 \text{ mg C m}^{-2}$  compared to Lakes Elementeita and Bogoria. The standing crop at Lake Natron lagoon declined slightly between April and December 2009 from  $22 \times 10^3 \text{ mg C m}^{-2}$  to  $16 \times 10^3 \text{ mg C m}^{-2}$  while the Natron Marsh maintained its standing crop at  $44 \times 10^3 \text{ mg C m}^{-2}$ .

Table 4.6: Mean ( $\pm$ se) standing crop of the lake shore mud algae.

Site	Month/year	Max $\text{mg C m}^{-2} (\times 10^3)$	Min $\text{mg C m}^{-2} (\times 10^3)$	se ( $\times 10^3$ )	n
Bogoria	Apr-09	55.46	1.66	6.86	2
	Aug-10	0.53	0.02	-	1
Elementeita	Aug-09	2.83	0.09	2.86	6
Nakuru Lake	Aug-09	14.27	0.43	17.07	2
Nakuru Njoro river mouth	Aug-09	0.62	0.02	1.48	4
Natron lagoon	Apr-09	22.97	0.69	14.65	14
	Dec-09	16.73	0.50	5.50	8
Natron marsh	Apr-09	44.08	1.32	26.40	7
	Dec-09	44.90	1.35	24.44	12
Sonachi	Sep-09	6.08	0.18	0.08	2
Oloidien	Sep-09	3.91	0.12	0.25	3

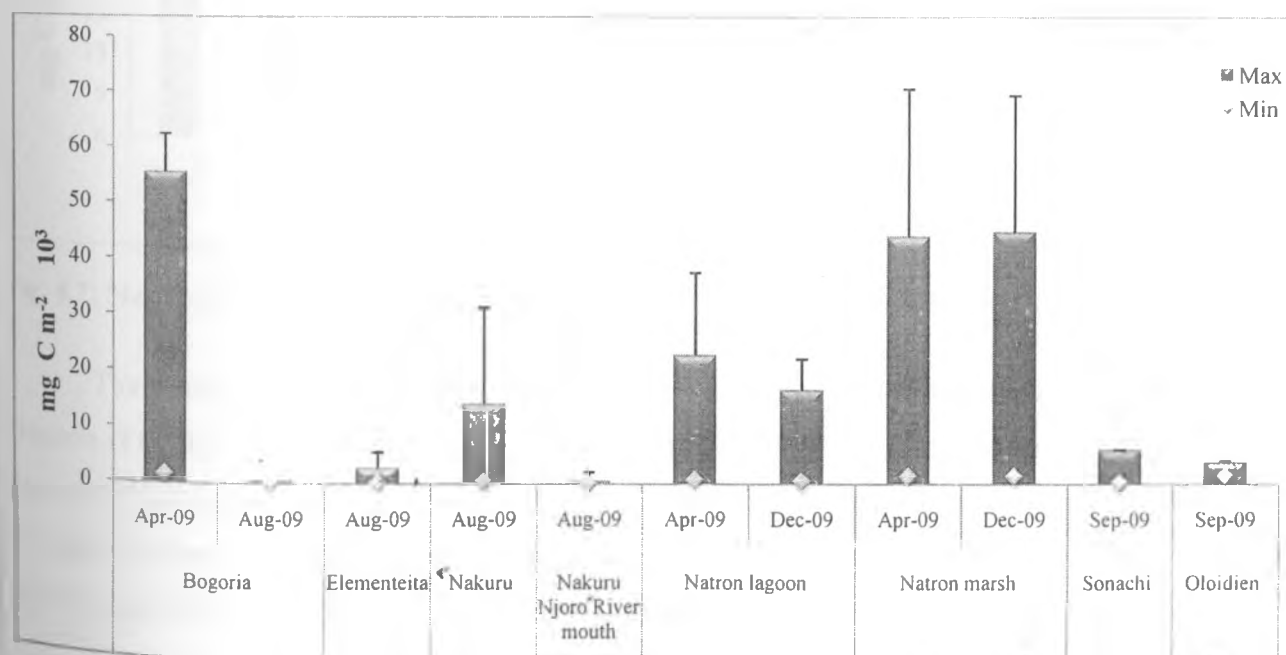


Fig. 4.5: Mean ( $\pm$  se) standing crop of lake shore mud algae in the study lakes.

#### 4.4. Effect of water depth on standing crop of the various food resources

The Spearman's rank-order correlation indicated a weak negative association between the standing crop of the planktonic algae and water depth (Fig. 4.6). There a general decrease in was a significant negative correlation between water depth and the epipellic standing crop of  $r_s = -0.297$  ( $p < 0.01$ ) (Fig. 4.7).

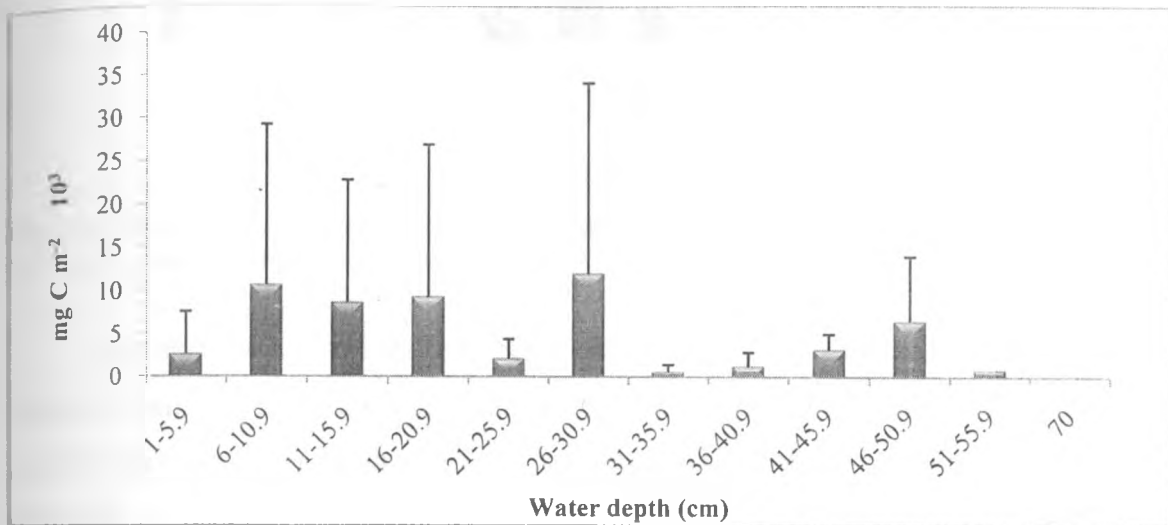


Fig. 4.6: Mean ( $\pm$ se) planktonic standing crop at varying water depths of all the study lakes.

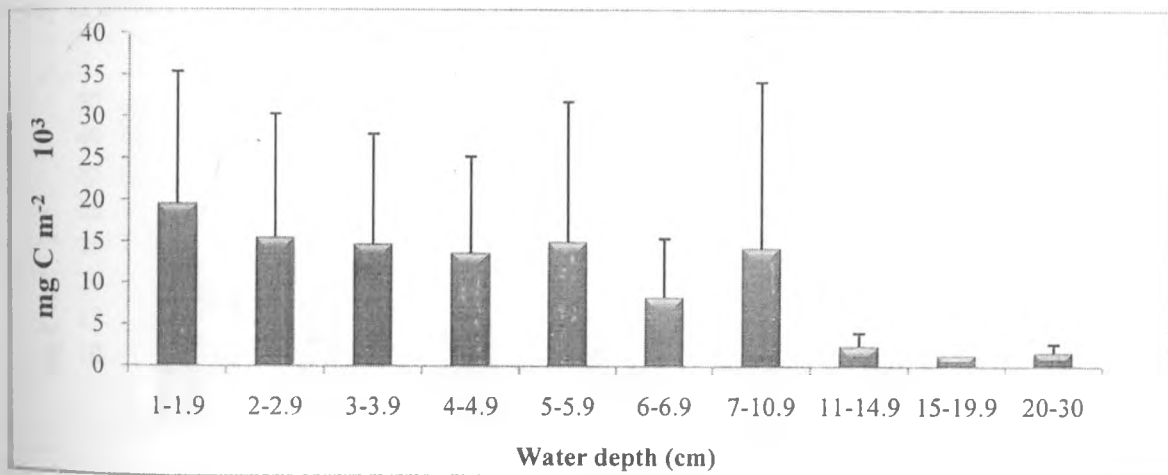


Fig. 4.7: Mean ( $\pm$ se) epipellic standing crop at varying water depth of the study lakes.

The standing crop of both the planktonic and sedimented *A. fusiformis* at Lake Bogoria (Fig. 4.8) increased slightly from a depth of 1- 10.9 cm of water depth and decreased to a water depth of 25.9 cm. It was highest at 26-30.9 cm water depth and thereafter decreased up to a depth of 50.9 cm water depth where the sedimented became higher than planktonic from 51-70 cm water depth.



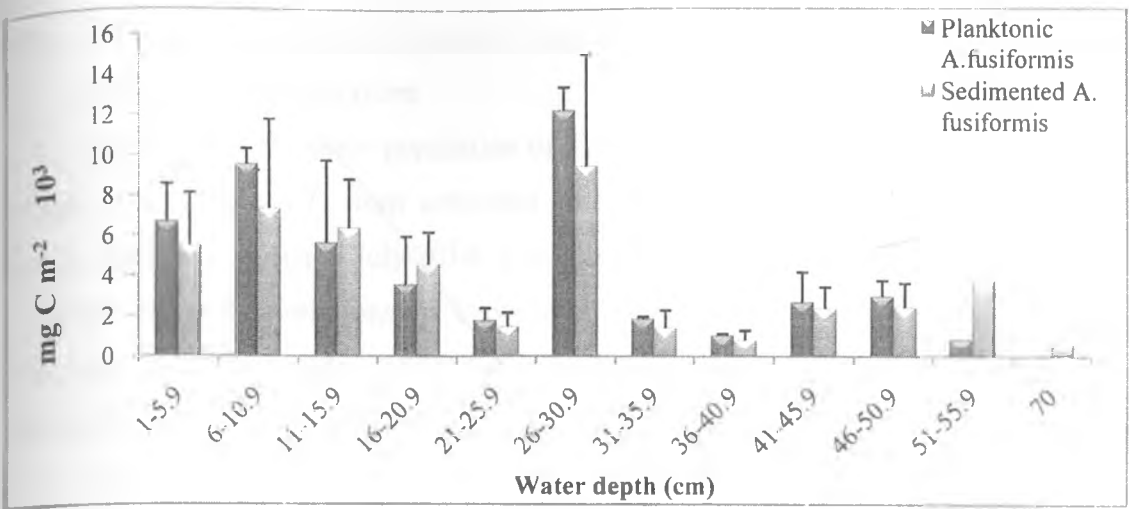


Fig. 4.8: The comparison between the mean planktonic and sedimented *A. fusiformis* biomass at Lake Bogoria at different water depths.

The shallow lakes; Lake Natron and Lake Elementeita, whose depth did not exceed 10cm, were characterized by high biomass of benthic diatoms. There was a general decrease in both the epipelagic and suspended epipelagic standing crop as observed in figure 4.9 with a weak positive Spearman's rank-order correlation ( $r_s=0.176$ ). The suspended epipelagic biomass was higher than the epipelagic biomass only at a water depth of 1-1.9 cm and thereafter the epipelagic biomass was higher.

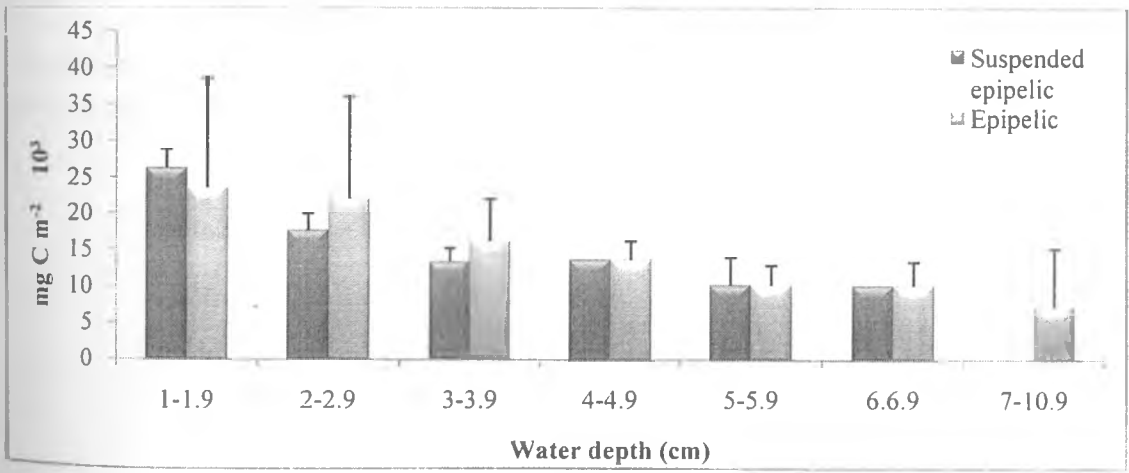


Fig. 4.9: Comparison between the mean suspended epipelagic and epipelagic biomass of Lake Natron and Lake Elementeita at different water depths.

#### 4.5. *Phoeniconaias minor* population estimates and standing crop of the different food resources

The greatest *P. minor* population of 516,979 was recorded at Lake Bogoria in August 2010 (Table 4.7) from estimates carried out by the National Museums of Kenya waterfowl census in July 2010. The Spearman's rank-order correlation showed a positive but weak relationship between *P. minor* population estimates and the planktonic standing crop ( $r_s=0.571$ ) but indicated a negative correlation with epipellic standing crop ( $r_s=-0.190$ ) and lake shore mud standing crop ( $r_s=-0.286$ ).

**Table 4.7:** Mean ( $\pm$ se) standing crop of different food resources and *P. minor* population estimates. (NMK= National Museums of Kenya national waterfowl census).

SCmax	Month-year	Planktonic mg C m <sup>-2</sup> × 10 <sup>3</sup>	Epipellic mg C m <sup>-2</sup> × 10 <sup>3</sup>	Lake shore mud mg C m <sup>-2</sup> × 10 <sup>3</sup>	<i>P. minor</i> population	<i>P. minor</i> population estimated by
Bogoria	April-2009	37.74	17.24	55.46	128,515	NMK January-2009
	August-2009	3.24	4.25	0.53	12,929	NMK July-2009
	August-2010	5.17	3.85	0.66	516,979	NMK July-2009
Elementeita	August-2009	0.66	4.92	2.83	6,325	NMK July-2009
Nakuru	April-2009	11.97	46.66	–	250,000	This study
	August-2009	0.90	5.81	14.27	255,294	NMK July-2009
Natron lagoon	April-2009	–	12.74	22.97	6,038	This study
Natron lagoon	Dec-2009	0.79	20.69	16.73	11,268	This study

#### **4.6. Primary productivity of the various communities.**

High primary productivity was indicated by extremely high dissolved oxygen concentration reaching super-saturation levels of over 300% with concentrations of 25 mg O<sub>2</sub> L<sup>-1</sup> at Lake Bogoria. Tables 4.8 – 4.11 and Figures 4.10 – 4.13 below summarize the primary productivity and respiration values of the study lakes during the periods of study.

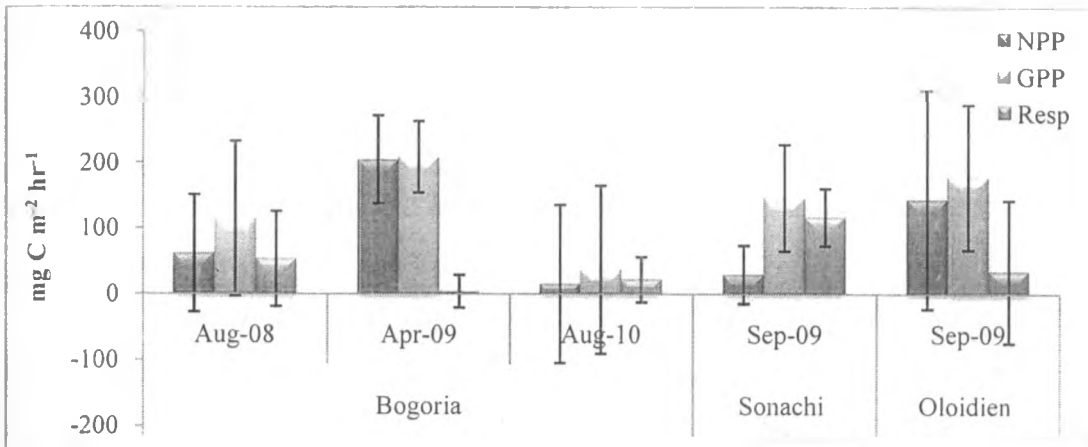
##### **4.6.1. Primary productivity of the planktonic and the suspended epipelagic community**

The highest hourly planktonic net primary productivity (NPP) was recorded at Lake Bogoria in April 2009 with a mean value of 204.6 mg C m<sup>-2</sup> hr<sup>-1</sup> (Table 4.8, Fig. 4.10). This corresponded with the highest planktonic standing crop of 37.7 × 10<sup>3</sup> mg C m<sup>-2</sup>. High primary productivity values were recorded in April with a significant decline (p<0.01) recorded at Lake Bogoria between April 2009 and August 2010. Lake Oloidien, which had the second highest NPP of planktonic standing crop 143.81 mg C m<sup>-2</sup> hr<sup>-1</sup> and a standing crop of 3.63 × 10<sup>3</sup> mg C m<sup>-2</sup>. The lowest planktonic standing crop of 0.56 × 10<sup>3</sup> mg C m<sup>-2</sup> recorded at Sonachi was comparable with low planktonic NPP of 29.67 mg C m<sup>-2</sup> hr<sup>-1</sup>. Sonachi recorded the highest community respiration of 117 mg C m<sup>-2</sup> hr<sup>-1</sup>. Lake Nakuru recorded high standing crop of 1.41 × 10<sup>3</sup> mg C m<sup>-2</sup>, but the productivity was not measured due to failure of the oxygen meter.

Lake Elementeita's suspended epipelagic community recorded high NPP of 60.82 mg C m<sup>-2</sup> hr<sup>-1</sup> although it was only 30% in comparison to the productivity of the planktonic community at Lake Bogoria. The suspended epipelagic community at Lake Natron's lagoon recorded negative net primary productivity values with a mean of -33 mg C m<sup>-2</sup> hr<sup>-1</sup> in April 2009. Lake Natron marsh which had high standing crop of 4.29 × 10<sup>3</sup> mg C m<sup>-2</sup> recorded negative NPP of -20.27 mg C m<sup>-2</sup> hr<sup>-1</sup> (Table 4.9, Fig. 4.11).

**Table 4.8:** Mean ( $\pm$ se) primary productivity and respiration of the plankton community. (NPP= Net primary productivity, GPP= Gross primary productivity, Resp = Respiration).

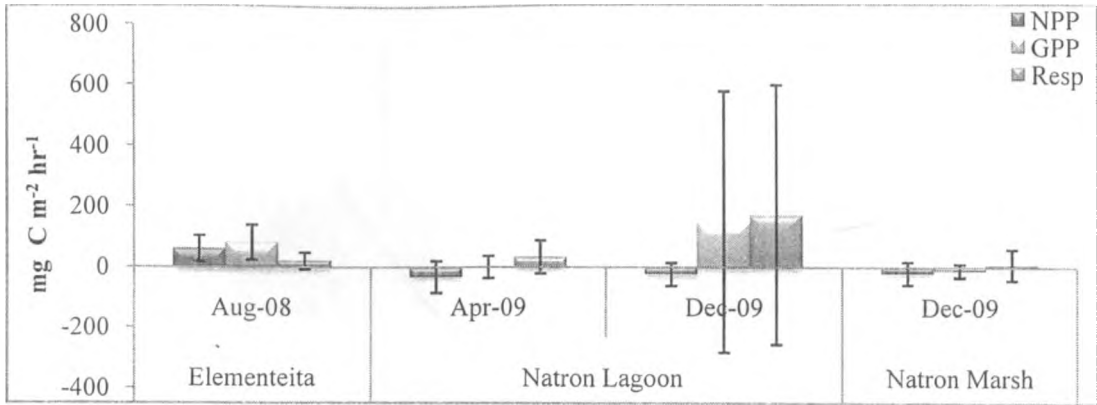
Lake	Month-Year	NPP mg C m <sup>-2</sup> hr <sup>-1</sup>			GPP mg C m <sup>-2</sup> hr <sup>-1</sup>			Resp mg C m <sup>-2</sup> hr <sup>-1</sup>		
		Mean	sd	n	Mean	sd	n	Mean	sd	n
Bogoria	Aug-08	61.73	89.25	68	115.18	117.86	68	53.43	72.42	68
	Apr-09	204.60	67.24	10	208.67	54.68	10	4.07	24.95	10
	Aug-10	15.49	119.90	22	37.55	127.66	22	22.06	34.65	22
Sonachi	Sep-09	29.67	44.43	6	146.67	81.32	6	117.00	43.91	6
Oloidien	Sep-09	143.81	166.71	7	178.10	110.80	7	34.29	108.56	7



**Fig. 4.10:** Mean ( $\pm$  se) net and gross primary productivity and respiration of the planktonic community of the lakes studied. (NPP= Net primary productivity, GPP= Gross primary productivity, Resp = Respiration).

**Table 4.9:** Mean ( $\pm$ se) primary productivity and respiration of the suspended epipelagic community. (NPP= Net primary productivity, GPP= Gross primary productivity, Resp = Respiration).

Lake	Month-Year	NPP mg C m <sup>-2</sup> hr <sup>-1</sup>			GPP mg C m <sup>-2</sup> hr <sup>-1</sup>			Resp mg C m <sup>-2</sup> hr <sup>-1</sup>		
		Mean	sd	n	Mean	sd	n	Mean	sd	n
Elementeita	Aug-08	60.82	43.19	13	81.74	57.62	13	20.97	27.81	13
Natron Lagoon	Apr-09	-33.00	53.39	8	0.00	37.03	8	33.00	54.89	8
	Dec-09	-23.88	38.52	11	146.42	430.11	11	170.36	426.74	11
Natron Marsh	Dec-09	-20.27	38.62	10	-11.60	22.60	10	8.67	51.48	10



**Fig. 4.11:** Mean ( $\pm$  se) net and gross primary productivity and respiration of the suspended epipelagic community of the lakes studied. (NPP= Net primary productivity, GPP= Gross primary productivity, Resp = Respiration).

#### 4.6.2. Primary productivity of the sedimented *A. fusiformis* and epipelagic community

The sedimented *Arthrospira fusiformis* of Lake Bogoria (August 2008) recorded high net productivity with a mean of  $103.01 \text{ mg C m}^{-2} \text{ hr}^{-1}$  (Table 4.10). By August 2010, Lake Bogoria's net productivity recorded a significant decline ( $p < 0.001$ ) with 60% less than the value recorded in August 2008 (Fig. 4.12).

The highest net primary productivity of the epipelagic community was recorded at Lake Elementeita with  $72.9 \text{ mg C m}^{-2} \text{ hr}^{-1}$  (Table 4.11, Fig. 4.13). This was only 70% in comparison to the sedimented *A. fusiformis* at Lake Bogoria. The lowest net primary productivity values were recorded at Lake Natron's marsh in December 2009 with mean of  $3.71 \text{ mg C m}^{-2} \text{ hr}^{-1}$ . High standing crop at Natron did not translate into high productivity. A standing crop of  $20.69 \times 10^3 \text{ mg C m}^{-2}$  at Natron Lagoon had a net productivity of  $21.62 \text{ mg C m}^{-2} \text{ hr}^{-1}$  compared to Bogoria 4 times less biomass had twice the productivity at  $41.38 \text{ mg C m}^{-2} \text{ hr}^{-1}$ .

**Table 4.10:** Mean ( $\pm$ se) primary productivity and respiration of the sedimented *Arthrospira fusiformis*. (NPP= Net primary productivity, GPP= Gross primary productivity, Resp = Respiration).

Lake	Month-Year	NPP $\text{mg C m}^{-2} \text{ hr}^{-1}$			GPP $\text{mg C m}^{-2} \text{ hr}^{-1}$			Resp $\text{mg C m}^{-2} \text{ hr}^{-1}$		
		Mean	se	n	Mean	se	n	Mean	se	n
Bogoria	Aug-08	103.01	35.10	24	126.61	37.40	24	23.60	20.43	24
	Aug-10	41.38	74.33	15	58.70	61.06	15	17.32	31.19	15

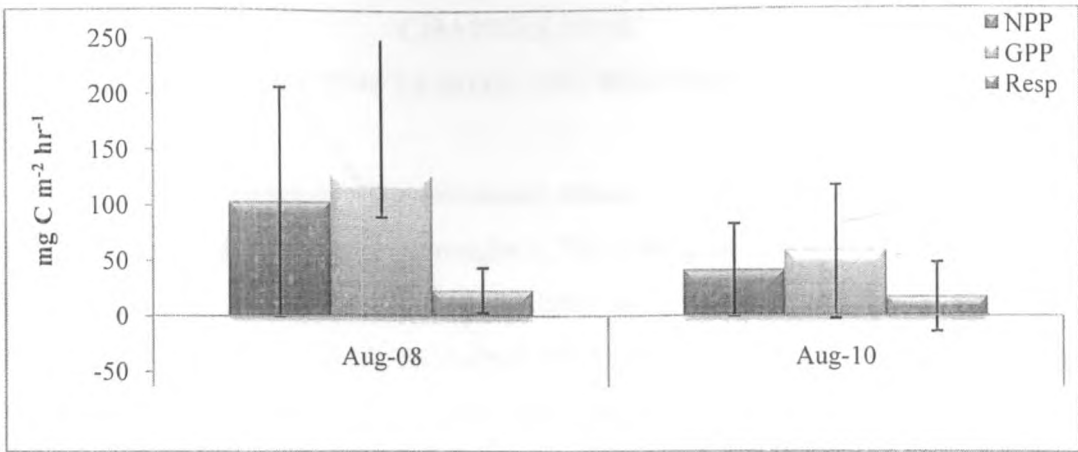


Fig. 4.12: Mean ( $\pm$  se) net and gross primary productivity and respiration of the sedimented *Arthrospira fusiformis* at Lake Bogoria. (NPP= Net primary productivity, GPP= Gross primary productivity, Resp = Respiration).

Table 4.11: Mean ( $\pm$ se) primary productivity and respiration of the epipelagic community. (NPP= Net primary productivity, GPP= Gross primary productivity, Resp = Respiration).

Lake	Month-Year	NPP mg C m <sup>-2</sup> hr <sup>-1</sup>			GPP mg C m <sup>-2</sup> hr <sup>-1</sup>			Resp mg C m <sup>-2</sup> hr <sup>-1</sup>		
		Mean	se	n	Mean	se	n	Mean	se	n
Elementeita	Aug-08	72.98	45.92	14	96.01	55.30	14	23.03	14.29	14
Natron lagoon	Apr-09	14.37	20.90	42	10.00	15.35	42	24.37	23.34	42
	Dec-09	21.62	10.89	11	55.97	27.81	11	34.35	28.46	11
Natron Marsh	Apr-09	28.32	15.13	11	40.33	21.23	11	12.01	10.95	11
	Dec-09	3.71	11.92	12	19.42	17.58	12	15.71	14.16	12

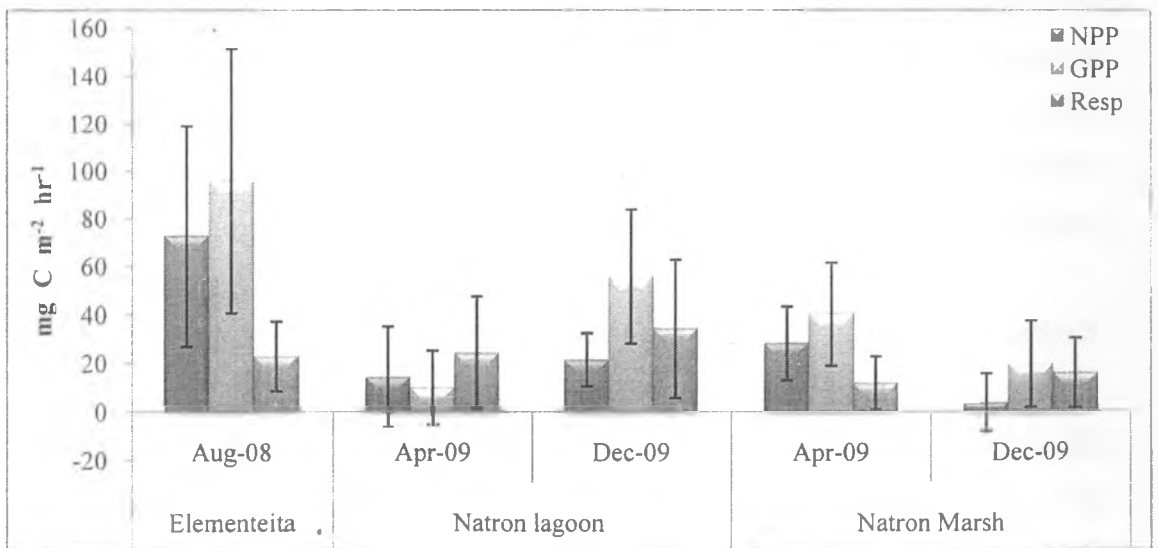


Fig. 4.13: Mean ( $\pm$  se) net and gross primary productivity and respiration of the epipelagic community of the studied lakes. (NPP= Net primary productivity, GPP= Gross primary productivity, Resp = Respiration).

## CHAPTER FIVE

### DISCUSSION CONCLUSION AND RECOMMENDATIONS

#### 5.1. Energy budget by *Phoeniconaias minor*

An adult lesser flamingo weighs  $1,730 \pm 40$  g and has to feed on  $72 \pm 6.5$  g of dry weight day<sup>-1</sup>, equivalent to  $314 \pm 28$  kcal day<sup>-1</sup> which is  $181.5$  kcal kg<sup>-1</sup> of body weight day<sup>-1</sup> (Vareschi, 1978). In August 2010 Lake Bogoria had the highest recorded lesser flamingo population of 516,979. The planktonic biomass on the shallow edges of the lake was  $5.17 \times 10^3$  mg C m<sup>-2</sup> ( $45.09$  kcal m<sup>-2</sup>). Assuming that the measured planktonic biomass was even over the whole lake, a lesser flamingo population of 516,979 would feed on 18.6 tons of C day<sup>-1</sup> (162,331 tons of calories day<sup>-1</sup>). This is only 92% of the total planktonic biomass of 20.15 tons C day<sup>-1</sup> (lake area 34 km<sup>2</sup>). This makes the lesser flamingo a major primary consumer on the saline lakes.

The remaining biomass is fed on by other important primary consumers such as the zoobenthic chironomid midge larvae, which occurs at a density of  $4 \times 10^4$  organisms m<sup>-2</sup> at Lake Bogoria (Harper *et al.*, 2003). Other consumers of planktonic biomass include protozoa, rotifers, and copepods. In lakes Nakuru, Elementeita and Natron the resident alkaline water fish *Alcolapia alcalicus grahami*, is an important primary consumer. A mean annual fish yield of 2,436 kg wet weight ha<sup>-1</sup> yr<sup>-1</sup> was recorded at Lake Nakuru in 1974 (Vareschi, 1978). With a production rate of 20.15 tons C day<sup>-1</sup> and a consumption rate of 18.6 tons of C day<sup>-1</sup>, it seems likely that the consumption may exceed production. During this period the lake regenerated at a rate of 2 tons C hr<sup>-1</sup> by considering both the planktonic and sedimented *A. fusiformis*. The high consumption actually facilitates in the regeneration of the cyanobacteria as 'cropping/pruning' by the consumers provides more space for growth.

At the shallow lakes, Lake Natron and Lake Elementeita, where the epipellic algae was suspended the lesser flamingo depended on the epipellic and lake shore mud food resources. Lake Natron lagoon had low suspended epipellic biomass which could only provide  $6.89$  kcal m<sup>-2</sup> ( $0.79 \times 10^3$  mg C m<sup>-2</sup>) of the required 314 kcal. The epipellic and lake shore mud recorded  $20.7 \times 10^3$  mg C m<sup>-2</sup> ( $180.68$  kcal m<sup>-2</sup>) and  $16.7 \times 10^3$  mg C m<sup>-2</sup> ( $145.9$  kcal m<sup>-2</sup>) respectively, could be able to meet this need. In this way, the different food resources, (i.e. the planktonic, epipellic and lake shore mud) are therefore able to complement each other to sustain the flamingo population in this system.

Though the food biomass calculation here takes into account the cyanobacteria (Plate 4.1) and algal (Plate 4.2) food components only, the diet of the lesser flamingo is also enriched with protein from several species of protozoa and rotifers (see Plate 4.3).

## 5.2. Primary producer community categories

Three primary producer communities can be recognized in saline lakes (Tuite, 1981). The three are categorized according to the dominant photo-synthesizers; 1. *Arthrospira* dominated phytoplankton, 2. Non-*Arthrospira* dominated phytoplankton and 3. benthic diatoms. From the results, the first type of primary producer community was represented in Lake Bogoria, Lake Oloidien and Lake Sonachi. Bogoria was characterised by high primary productivity by both the planktonic and sedimented *Arthrospira fusiformis* (204.60 and 103 mg C m<sup>-2</sup>hr<sup>-1</sup> respectively). The second state was represented by Lake Nakuru where *A. fusiformis* did not reach dominance level during the study period but other cyanobacteria species of *Anabaenopsis magna* and *A. abijatae*. However, it is important to note that the lake may transition between type 1 and 2. In fact *A. fusiformis* has been recorded by Vareschi (1978) to dominate Nakuru's phytoplankton at times when the lake level is high.

Primary productivity that is dominated by benthic diatoms system was represented by lakes Natron and Elementeita, where there was very low planktonic standing crop  $0.21 \times 10^3$  mg C. Primary production was totally supported by benthic diatom productivity. At Lake Natron benthic diatoms contributed 100% of the productivity considering that productivity measurements of the suspended epipellic community gave negative values of up to -33 mg C m<sup>-2</sup> hr<sup>-1</sup>, due probably due to high respiratory demands obtained. Diatoms did not achieve dominance at lakes where *Arthrospira spp.* dominated like lakes Bogoria, Nakuru, Oloidien and Sonachi, but dominated in the epipellic community of Lake Natron and Lake Elementeita. At the shallow lakes, Lake Natron and Lake Elementeita, protozoan and rotifer species were exceptionally scarce compared to the deeper lakes like lakes Bogoria, Nakuru and Oloidien.



### 5.3. Production rate of lesser flamingo food

High levels of primary productivity were recorded in the six lakes and Talling *et al.*, (1973) proposed that these high rates were due to a combination of high tropical temperatures, stability of the solar illumination in the tropics and high dissolved inorganic phosphorus resulting in high phytoplanktonic primary productivity. High net primary productivity of  $103 \text{ mg C m}^{-2} \text{ hr}^{-1}$  was measured for sedimented *A. fusiformis* at Lake Bogoria. Although it is expected that surface photo-inhibition would result in low production through self shading by the population. In some cases, the cyanobacteria species can adapt to relatively low light intensities (Wetzel, 2001) at the deeper layers resulting to higher photosynthetic rates than at the water surface as evidenced by the high productivity of the sedimented *A. fusiformis*.

Negative net primary productivity was recorded for the planktonic community of Lake Natron was due to high respiration by the whole planktonic community. Models estimating net primary productivity by Oduor and Schagerl (2007) showed negative values on certain days for lakes Bogoria and Nakuru. This situation was attributed to high respiration of the community or as a result of ecological conditions under cloud cover.

From a comparison of productivity values generated from this study and other previous measurements, showed that some of the previous measurements are at least 5 times higher (Table 5.1) than the values from this study but are within the expected range for tropical lakes. At Lake Bogoria, Melack (1981) measured a NPP of  $320 \text{ mg C m}^{-2} \text{ h}^{-1}$  while this study estimated a NPP of  $72 \text{ mg C m}^{-2} \text{ h}^{-1}$  similarly, at Lake Sonachi, Melack (1981) measured a NPP of  $146 \text{ mg C m}^{-2} \text{ h}^{-1}$  while this study estimated a NPP of  $29 \text{ mg C m}^{-2} \text{ h}^{-1}$ . These variations could be attributed mainly to the use of different methods in different studies. These results could also have been influenced by changes within the lakes ecology and chemistry human activities within the catchment

Diatoms contributed the highest biomass in shore mud samples for lakes where *P. minor* were observed feeding on mud at the lake shore. These lakes were Lake Elementeita and Lake Natron. The birds foraged in the shallow water where they would shuffle their legs in the water to suspend the epipellic food resource growing on the sediment. Occasionally they were observed feeding at the lake edge on diatoms growing on the shore mud. Epipellic and lake shore mud communities

contributed significantly to the food resources of lesser flamingos and the primary productivity of the lakes especially on shallow lakes.

**Table 5.1:** Comparison of primary productivity values measured during this study with others from previously published work from different locations.

Location	Lake	Planktonic NPP mg C m <sup>-2</sup> h <sup>-1</sup>	Epipellic NPP mg C m <sup>-2</sup> h <sup>-1</sup>	Reference
Kenya	Bogoria	93.94	72.2	<b>This study</b>
	Bogoria	320.00	-	Melack , 1981
	Bogoria	2.27	-	Odour and Schagerl, 2007
	Nakuru	216.67	-	Melack , 1981
	Nakuru	89.00	-	Melack and Kilham , 1974
	Nakuru	3.57	-	Odour and Schagerl, 2007
	Elementeita	60.82	72.98	<b>This study</b>
	Elementeita	166.67	-	Melack , 1981
	Elementeita	2.83	-	Odour and Schagerl, 2007
	Sonachi	29.67	-	<b>This study</b>
Sonachi	146.67	-	Melack , 1981	
Tanzania	Natron Lagoon	-28.44	18	<b>This study</b>
	Natron Marsh	-20.27	16.02	<b>This study</b>
	Reshitani	667	-	Melack , 1981
Ethiopia	Kilotes	266.67	-	Talling et al, 1973
	Aranguadi	665.00	-	Talling et al, 1973
United States of America	Borax	-	480	Hammer 1981

#### 5.4. Foraging options

All animals feed for growth, maintenance and reproduction. A lesser flamingo requires a daily maintenance food ration of  $72 \pm 6.5$  g of dry weight day<sup>-1</sup>, equivalent to  $314 \pm 28$  kcal day<sup>-1</sup> (Vareschi, 1978), this is the food needed by an adult bird to be in a state of constant body composition. The flamingo population estimates were weakly correlated with values of planktonic biomass. For most of the lakes the planktonic biomass would be unable to sustain the flamingo population without the flamingos depending on both the epipellic and lake shore mud resources. Correlation of flamingo population with epipellic and lake shore mud food biomass indicated a weak negative correlation.

As food resources become limited in a lake e.g. at Lake Elementeita (see Table 4.7.) where low food biomass is accompanied by low population, the energy gain is expected to gradually decrease as it takes more time to obtain the daily maintenance food ration. This also would require more energy to acquire the food. An ultimate point/peak point is reached where the energy cost in acquiring food is greater than energy gained from the food. This implies that the food available cannot sustain the population. At this point lesser flamingo will choose to stop feeding and start moving to the next site. They choose instead to conserve energy for travel to the next feeding site. In so doing they risking to loose energy in travelling as they do not know whether the next site will have adequate food supply.

The fundamental choice that a lesser flamingo has to make is how long to stay at a feeding site, when and where to move in search for the optimal feeding site among the widely scattered soda lakes of on the eastern rift valley. In a situation where food declines at Lake Nakuru and the flamingos would have to move. They have several choices of lakes that are within a radius of 70 km, Lake Bogoria, to the North and Lakes Elementeita, Oloidien and Sonachi to the south. When the time comes for them to move, a few flamingos start flying in a circle above the lake as the sun sets. The group becomes larger as more birds join and just as it becomes dark they fly off to the next lake. Though they move as a group it is the decisions of the individual birds that seem to coincide with the group's interests. The decision has to ensure the highest probability of finding a high quality and quantity food supply that justifies the energy spent travelling there. To make good decisions the assumption is that flamingos have acquired perfect knowledge and information about their environment, to make reliable decisions that have eventually ensured the survival of the species in the millions of years they have been on earth.

## 5.5. Conclusion

The study builds on the works of Tuite (1981) and Vareschi (1978) who were first to explore the importance of epipellic and lake shore mud food resources as flamingo food and their importance to the lake's ecology as they support the food chains. From this study, it can be appreciated that both the epipellic and lake shore mud algae play an important role in providing nutrition for the lesser flamingo. The lesser flamingo population is not just sustained on *Arthrospira fusiformis* but may

from time to time rely exclusively on the epipelagic algae especially when the birds are on the shallow lakes. The epipelagic algae is also important ecologically as it provides energy to the other trophic levels that includes consumers such as protozoa and rotifers which also enrich the *P. minor* diet. Time spent feeding would depend on the species type of the food resource and its biomass. *P. minor* displays a great ability to adapt to change in food availability by adapting their feeding behaviour to acquire the most food such as diving in the water at Lake Bogoria to feed on the sedimenting *A. fusiformis*.

## **5.6. Recommendations**

### **5.6.1. Further research actions**

I propose that further research should be carried out to determine the time lesser flamingos spend feeding on the different food resources and the nutritional value of the various food resources. Studies should be designed that can investigate any differentiation in the food requirements and intake between birds of different ages, the immature and the adult flamingos. The cause of sedimentation of *A. fusiformis* should be investigated. The hydrology of the lakes and the rivers and springs that sustain these lakes needs to be well established including the effects of human land use practices within the catchment. For the effective management and conservations of flamingos, the lesser flamingo population numbers and trends and the cause of mortality need to be further studied. There is a need to establish how local ecological changes contribute to the changes in saline lake ecosystems and the resilience of flamingos and saline lake to these changes.

### **5.6.2. Conservation and management actions**

Habitat connectivity of habitats used by flamingos should be ensured by protecting all the habitats that serve as main feeding or breeding sites including the other small lakes that are infrequently utilized by flamingos. Any management interventions should emphasise on the protection and ensuring the biological diversity and sustaining lake ecosystem functioning instead of focusing on single species.

## REFERENCES

Adhola, T., Ng'weno, F., Matiku, P., Mulwa, R., Ngari, A., Barasa, F., Madindou, I., Wanyiri, M., Mwango'mbe, J., Musyoki, C., Ndonye, P., Ogoma, M., Machekele, J. and Musila, S. (2009). Kenya's Important Bird Areas: Status and Trends 2008. Nature Kenya, Nairobi.

Baker, M. 2011. Lake Natron, Flamingos and the Proposed Soda Ash Factory. An overview. <http://www.tnrf.org/files/natronbriefing.pdf> (Accessed on 16/5/2011)

Ballot, A., Krienitz, L., Kotut, K., Wiegand, C., Metcalf, J. S., Codd, G. A. and Pflugmacher, S., 2004. Cyanobacteria and cyanobacterial toxins in three alkaline Rift Valley lakes of Kenya - Lakes Bogoria, Nakuru and Elementeita. *Journal of Plankton Research*. 26(8): 925-935.

Ballot, A., Krienitz, L., Kotut, K., Wiegand, C. and Pflugmacher, S., 2005. Cyanobacteria and cyanobacterial toxins in the alkaline crater lakes Sonachi and Simbi, Kenya. *Harmful Algae*. 4: 139-150.

Ballot, A., Kotut, K., Novelo, E. and Krienitz, L. 2009. Changes of phytoplankton communities in Lakes Naivasha and Oloidien, examples of degradation and salinization of lakes in the Kenyan Rift Valley. *Hydrobiologia*. 632: 359-363.

Barber, H.G. and Haworth, E.Y. 1994. A guide to the morphology of the diatom frustules with a key to the British freshwater genera. *Freshwater Biological Association*. Scientific Publication No. 44.

Bartholomew, G.A. and Pennycuik, C.J. 1973. The flamingo and pelican populations of the Rift Valley lakes in 1968-69. *East African Wildlife Journal*. 11:189-198.

Bellinger E.G. 1992. A key to common algae freshwater Estuarine and some coastal species. *The Institute of Water and Environmental Management*. 4<sup>th</sup> Ed.

Bettinetti, R., Quadroni, S., Crosa, G., Harper, D., Dickie, J., Kyalo, M., Mavuti, K., and Galassi, S. 2011. A Preliminary Evaluation of the DDT Contamination of Sediments in Lakes Natron and Bogoria (Eastern Rift Valley, Africa). *A Journal of the Human Environment*. 40(4): 341-350.

Bildstein, K. L., Golden, C. B., McCraith B. J., Bohmke, B. W. and Seibels R. E., 1993. Feeding behavior, aggression, and the conservation biology of flamingos: integrating studies of captive and free-ranging birds. *American Zoologist*. ISSN 0003-1569 CODEN AMZOAF.

Björk-Ramberg S. and Ånell C. 1985. Production and chlorophyll concentration of epipelagic epilithic algae in fertilized and nonfertilized subarctic lakes. *Hydrobiologia*. 126:213-219.

Boney, A.D. 1989. Phytoplankton 2<sup>nd</sup> Ed. *Routledge, Chapman and Hill*.

Brown, L., 1973. The mystery of the flamingos. East African Publishing House, Nairobi.

Childress, B., Nagy, S. and Hughes, B. (Compilers). 2008. International Single Species Action Plan for the Conservation of the Lesser Flamingo (*Phoeniconaias minor*). CMS Technical Series No. 18, AEWA Technical Series No. 34. Bonn, Germany.

Clamsen, T. E. M., Maliti, H. and Fyumagwa, R. 2011. Current population status, trend and distribution of Lesser Flamingo *Phoeniconaias minor* at Lake Natron, Tanzania. Flamingo, Bulletin of the IUCN-SSC/Wetlands International Flamingo Specialist Group, No. 18. Wildfowl & Wetlands Trust, Slimbridge, UK.

DeMenocal, P.B. 1995. Plio-Pleistocene African climate. *Science*. 270: 53-59.

Epp, S. L., Stoof, R. K., Trauth, H. M., and Tiedemann, R. 2010. Historical genetics on a sediment core from a Kenyan lake: intraspecific genotype turnover in a tropical rotifer is related to past environmental changes. *Journal of Paleolimnology*. 43: 939-954

Falkowski, P. G., Dubinsky, Z. and Wyman, K. 1985. Growth-irradiance relationships in phytoplankton. *Limnology and Oceanography*. 30(2): 311-321.

Grant, D. W. 2004. Half a lifetime in sodalakes. *Halophilic Microorganisms*. 17-21.

Hall, R. P. 1953. Protozoology. Prentice-Hall, Inc. New York.

Hammer, U. T., 1981. Primary production in saline lakes. *Hydrobiologia*. 81: 47-57.

- Harper, D.M., Childress, R.B., Harper, M.M., Boar, R.R., Hickley, p., Mills, S.C., Otieno, N., Drane, T., Vareschi, E., Nasirwa, O., Mwatha, W.E., Darlington. J.P. and Escute-Gasulla, X. (2003). Aquatic biodiversity and saline lakes: Lake Bogoria National Reserve, Kenya. *Hydrobiologia*. 500: 259-276.
- Harper, D. M., Muchane, M., Kimani, D. K. and Mwinami, T. 2006. Thousands of lesser flamingos at Lake Naivasha? *Scopus*. 26: 8-10.
- Hecky, R. E. and Kilham, P. 1973. Diatoms in alkaline, saline lakes: ecology and geochemical implications. *Limnology and Oceanography*. 18(1): 53-71.
- Hickley, P., Boar, R. R. and Mavuti, K. M., 2003. Bathymetry of Lake Bogoria, Kenya. *Journal of East African Natural History*. 92: 107-117.
- Hillebrand, H., Dürselen, C. D., Kirschtel, D., Pollinger, U. and Zohary. T. 1999. Biovolume calculation for pelagic and benthic microalgae. *Journal of Phycology*. 35: 403-424.
- Iliffe, M., Mwinami, T. and Harper, D. 2011. Counting flamingos with a mobile phone-connecting all the lakes? Flamingo, Bulletin of the IUCN-SSC/Wetlands International Flamingo Specialist Group, No. 18. Wildfowl & Wetlands Trust, Slimbridge, UK. Pp. 38-41.
- IUCN. 2010. IUCN Red List of Threatened Species. 2010. <http://www.iucnredlist.org/apps/redlist/details/144723/0>. (Accessed 11/05/2011).
- Jenkin, P. M. 1957. The Filter-Feeding and Food of Flamingoes (Phoenicopteri). *Philosophical Transactions of the Royal Society of London*. 240 (674): 401-493.
- Kasule, F. K., Mlingwa, C. O. F. and Mwasumbi, L. B. 1993. Environmental reconnaissance of the Lake natron area, July/August 1992. A report commissioned by MDPA.
- Krienitz, L., Ballot, A., Kotut, K., Wiegrand, C., Pütz, S., Metcalf, J.S., Codd, G. A. and Pflugmacher, S., 2003. Contribution of hot spring cyanobacteria to the mysterious deaths of Lesser Flamingos at Lake Bogoria, Kenya. *Microbiology Ecology*. 43: 141-148.

Kulshreshtha, S., Kulshreshtha, M and Sharma, B. K. 2011. Ecology and present status of flamingos at Sambhar Salt Lake, Rajasthan, India: a critical comparison with past records. *Flamingo*, Bulletin of the IUCN-SSC/Wetlands International Flamingo Specialist Group, No. 18. Wildfowl & Wetlands Trust, Slimbridge, UK. Pp. 24-27

Lugomela, C., Pratab, H. B. and Mgaya, Y. D., 2006. Cyanobacterial – A possible cause of mass mortality of lesser flamingos in Lake Manyara and Lake Big Momela, Tanzania. *Harmful Algae*. 5: 534 – 541.

Lyngs, P . 1996. Waterbirds at Lake Oloidien, Naivasha, Kenya, autumn 1987. *Wader Study Group Bull.* 79:91-102.

McClanahan, T. R. and Young, T. P. 1996. East African ecosystems and their conservation. Oxford University Press US. Melack J.M. Saline and freshwater lakes of the Kenyan Rift Valley. 171-185.

Melack, J. M., 1981. Photosynthetic activity of phytoplankton in tropical African soda lakes. *Hydrobiologia*. 81: 71-85.

Melack, J. M. and Kilham, P. 1974. Photosynthetic rates of phytoplankton in East African alkaline saline lakes. *Limnology and Oceanography*. 19 (5): 734-755.

Morales-Roldan, H. L., Mwinami, T., and Harper, D. M. 2011. Is ground census of flamingo lakes worthwhile? *Flamingo*, Bulletin of the IUCN-SSC/Wetlands International Flamingo Specialist Group, No. 18. Wildfowl & Wetlands Trust, Slimbridge, UK. Pp. 42-44.

Mwimani, T., Barasa, F., Ngari, A., Matiku, P., Ng'weno, F., Musina, J., Mwangombe, J. and Kanga, E. 2010. Kenya's Important Bird Areas: Status and trends 2009. Nature Kenya, Nairobi.

NASA Earth Observatory. 2008. Fire damages Kenya's Lake Nakuru National park: Image of the day 2008. [http://eoimages.gsfc.nasa.gov/images/imagerecords/8000/8514/kenya\\_ast\\_2008065\\_lrg.jpg](http://eoimages.gsfc.nasa.gov/images/imagerecords/8000/8514/kenya_ast_2008065_lrg.jpg) (Accessed 20/7/2011).

Ndetei, R. and Muhandiki, V.S. (2005). Mortalities of the lesser flamingos in Kenyan Rift Valley saline lakes and the implications for sustainable management of the lakes. *Lakes and Reservoirs: Research and Management*. 10:51-58.



- Nielsen, J. M. and Dahi, E. 1995. Occurrence of fluorine contaminated magadi(trona)in kenya and Tanzania. :Proceedings of 1st International Workshop on Fluorosis and Defluoridation of Water; ISFR for CDC. Pp 17-22.
- Oduor, S. O., and Schagerl, M. 2007. Phytoplankton primary productivity characteristics in response to photosynthetically active radiation in three Kenyan Rift Valley saline – alkaline lakes. *Journal of Plankton Research*. 29(12):1041-1050.
- Owino, A. O., Oyugi, J. O., Nasirwa,O. O. and Bennun, L. A. 2001. Patterns of variation in waterbird numbers on four Rift Valley lakes in Kenya, 1991–1999. *Hydrobiologia*. 458: 45–53.
- Pennycuick, C. J. and Bartholomew, G. A. 1973. Energy budget of the lesser flamingo (*Phoeniconaias minor*Geoffroy). *East African Wildlife Journal*.11: 199-207.
- Raini, J. and Ngowe, N. 2009. Aerial census of lesser flamingos (*phoeniconaias minor*) on the Rift Valley lakes of northern Tanzania, January 2002.*Flamingo*, Bulletin of the IUCNSSC/ Wetlands International Flamingo Specialist Group, No. 17, December 2009. Wildfowl & Wetlands Trust, Slimbridge, UK. Pp 30-34.
- Ramsar Advisory Mission: No 59, Tanzania. 2008. [http://www.ramsar.org/cda/en/ramsar-documents-rams-ram59/main/ramsar/1-31-112%5E23018\\_4000\\_0](http://www.ramsar.org/cda/en/ramsar-documents-rams-ram59/main/ramsar/1-31-112%5E23018_4000_0) (Accessed 11/5/2011)
- Ridley, M. W., Moss, B. L. and Percy R. C., 1955. The food of flamingoes in Kenya colony. *Journal of East Africa Natural History Society*.22 (5): 147 -158.
- Richardson, J. L. and Richardson, A. E. 1972. History of an African Rift lake and its climatic implications. *Ecological Monographs*. 42, 499-534.
- Sartory, D. P. and Grobbelaar, J. U. 1984. Extraction of chlorophyll a from freshwater phytoplankton for spectrophotometric analysis. *Hydrobiologia*. 114 (3): 177-187.
- Schagerl, M. and Oduor, S. O. 2008. Phytoplankton community relationship to environmental variables in three Kenyan Rift Valley saline-alkaline lakes. *Marine and Freshwater Research*.59: 125-136.
- Simmons, R.E. 1996. Population declines, viable breeding areas and management options for flamingos in Southern Africa. *Conservation Biology*. 10(2):504-514.

- Stanley, D. W. 1976. Productivity of epipelagic algae in tundra ponds and a lake near Barrow, Alaska. *Ecology*. 57 (5): 1015-1024.
- Stevenson, T. and Fanshawe, J. 2002. Field guide to the birds of East Africa. Kenya, Tanzania, Uganda, Rwanda, Burundi. T and A.D. Poyser, London. Pp 34-35.
- Talling, J. F., Wood, R. B., Prosser, M. V. and Baxter, R. M. 1973. The upper limit of photosynthetic productivity by phytoplankton: evidence from Ethiopian soda lakes. *Freshwater Biology*. 3: 53-76.
- Trauth, M.H., Maslin, M.A., Deino, A., and Strecker, M.R. 2005. Late Cenozoic moisture history of East Africa. *Science*. 309:2051–2053.
- Trauth, M.H., Maslin, M.A., Deino, A.I., Junginger, A., Lesoloyia, M., Odada, E.O., Olago, D.O., Olaka, L.A., Strecker, M.R., and Tiedemann, R. 2010. Human evolution in a variable environment: The amplifier lakes of Eastern Africa. *Quaternary Science Reviews*. In press.
- Tuite, C. H. 1979. Population size, distribution and biomass density of the lesser flamingo in the Eastern Rift valley, 1974-76. *Journal of Applied ecology*. 16 (3): 765-775.
- Tuite, C. H. 1981. Standing crop densities and distribution of *Spirulina* and benthic diatoms in East African saline lakes. *Freshwater Biology*. 11: 345-360.
- Vareschi, E., 1978. The ecology of Lake Nakuru (Kenya) I. Abundance and feeding of the Lesser Flamingo. *Oecologia* (Berl.). 32:11-35.
- Vareschi, E. 1982. The Ecology of Lake Nakuru (Kenya) III. Abiotic factors and primary production. *Oecologia* (Berl.). 55: 81-101.
- Vareschi, E. and Jacobs, J. 1984. The ecology of Lake Nakuru (Kenya) V. Production and consumption of consumer organisms. *Oecologia*. 61: 83-98.
- Vareschi, E. and Jacobs, J. 1985. The ecology of Lake Nakuru. VI. Synopsis of production and energy flow. *Oecologia*. 65: 412-424.

- Verschuren, D., Cocquyt, C., Tibby, J., Roberts, C.N. and Leavitt, P.R. 1999. Long-term dynamics of algal and invertebrate communities in a small, fluctuating tropical soda lake. *Limnology and Oceanography*. 44:1216–1231.
- Verschuren, D., 2001. Reconstructing fluctuations of a shallow East African lake during the past 1800 years from sediment stratigraphy in a submerged crater basin. *Journal of Paleolimnology*. 25:297-311.
- Verschuren, D., Cumming, B. F. and Laird, K. R. 2004. Quantitative reconstruction of past salinity variations in African lakes: assessment of chironomid-based inference models (Insecta: Diptera) in space and time. *Canadian Journal of Fisheries and Aquatic Sciences*. 61: 986–998.
- Vollenweider, R. A. 1974. IBP Handbook No. 12. A manual on methods of measuring primary production in aquatic environments. Second edition. Blackwell Scientific Publications, Oxford.
- Woodworth, B. L., Farm, B. P., Mufungo, C., Borner, M. and Kuwai, J.O. 1997. A photographic census of flamingos in the Rift Valley lakes of Tanzania. *African Journal of Ecology*. 35 (4): 326-334.

**APPENDICES**

**Appendix 1: Standing Crop**

**Appendix 1.a:** Mann-Whitney U Test results for planktonic standing crop.

	<b>Bogoria April 2009</b>	<b>Bogoria August 2009</b>	<b>Bogoria August 2010</b>	<b>Nakuru April 2009</b>	<b>Nakuru Njoro River August 2009</b>	<b>Nakuru August 2009</b>	<b>Sonachi September 2009</b>	<b>Oloidien September 2009</b>
<b>Bogoria April 2009</b>		0.000**	0.000**	0.243	0.002**	0.000**	0.004**	0.000**
<b>Bogoria August 2009</b>			0.019*	0.017*	0.338	0.781	0.206	0.016*
<b>Bogoria August 2010</b>				0.035*	0.011*	0.009**	0.011*	0.758
<b>Nakuru April 2009</b>					0.025*	0.007**	0.034*	0.012*
<b>Nakuru Njoro River August 2009</b>						0.681	0.142	0.019*
<b>Nakuru August 2009</b>							0.443	0.021*
<b>Sonachi September 2009</b>								0.020*
<b>Oloidien September 2009</b>								

\* = significant at  $p < 0.05$

\*\*= significant at  $p < 0.01$

**Appendix 1.b:** Mann-Whitney U Test results for the standing crop of the suspended epipelagic community.

	<b>Elementeita August 2009</b>	<b>Natron Lagoon December 2009</b>	<b>Natron Marsh December 2009</b>
<b>Elementeita August 2009</b>		0.668	0.001**
<b>Natron Lagoon December 2009</b>			0.001**
<b>Natron Marsh December 2009</b>			

\* = significant at  $p < 0.05$

\*\* = significant at  $p < 0.01$

**Appendix 1.c:** Mann-Whitney U Test results for standing crop of the sedimented *Arthrospira fusiformis*.

	<b>Bogoria April 2009</b>	<b>Bogoria August 2009</b>	<b>Bogoria August 2010</b>
<b>Bogoria April 2009</b>		0.000**	0.005**
<b>Bogoria August 2009</b>			0.151
<b>Bogoria August 2010</b>			

\* = significant at  $p < 0.05$

\*\* = significant at  $p < 0.01$

**Appendix 1.d:** Mann-Whitney U Test results for standing crop of the epipelagic community.

	Elementeita August 2009	Nakuru April 2009	Nakuru. Njoro River August 2009	Nakuru August 2009	Natron Lagoon April 2009	Natron Lagoon December 2009	Natron Marsh April 2009	Natron Marsh December 2009
Elementeita August 2009		0.001**	0.637	0.027*	0.010*	0.000**	0.001**	0.001**
Nakuru April 2009			0.003*	0.000**	0.000**	0.000**	0.132	0.022*
Nakuru Njoro River August 2009				0.257	0.040*	0.002**	0.007**	0.009**
Nakuru August 2009					0.000**	0.000**	0.000**	0.000**
Natron Lagoon April 2009						0.000**	0.000**	0.017*
Natron Lagoon December 2009							0.008**	0.855
Natron Marsh April 2009								0.217
Natron Marsh December 2009								

\* = significant at  $p < 0.05$

\*\* = significant at  $p < 0.01$

Appendix 1.e: Mann-Whitney U Test results for standing crop of the lake shore mud community.

	Bogoria April 2009	Bogoria August 2009	Elementeita August 2009	Nakuru Njoro River August 2009	Nakuru August 2009	Natron Lagoon April 2009	Natron Lagoon December 2009	Natron Marsh April 2009	Natron Marsh December 2009	Sonachi September 2009	Oloidien September 2009
Bogoria April 2009		0.221	0.046*	0.121	0.064	0.026*	0.037*	0.380	0.465	0.121	0.554
Bogoria August 2009			0.617	0.221	1.000	0.105	0.121	0.127	0.109	0.221	0.157
Elementeita August 2009				0.505	0.201	0.002**	0.002**	0.003**	0.001**	0.182	0.020*
Nakuru Njoro River August 2009					0.165	0.634	1.000	0.143	0.100	1.000	0.076
Nakuru August 2009						0.004**	0.007**	0.008**	0.004**	0.064	0.032*
Natron Lagoon April 2009							0.275	0.0378	0.016*	0.057	0.059
Natron Lagoon December 2009								0.015*	0.003**	0.037*	0.014*
Natron Marsh April 2009									0.933	0.040*	0.732
Natron Marsh December 2009										0.028*	0.563
Sonachi September 2009											0.076
Oloidien September 2009											

\* = significant at  $p < 0.05$

\*\* = significant at  $p < 0.01$

## Appendix 2: Primary productivity

Appendix 2.a: Mann-Whitney U Test results for primary productivity of planktonic community.

Planktonic Productivity		Bogoria Aug.-2008			Bogoria April-2009			Bogoria Aug.-2010			Sonachi Sept.-2009			Oloidien Sept.-2009		
		GPP	Resp	NPP	GPP	Resp	NPP	GPP	Resp	NPP	GPP	Resp	NPP	GPP	Resp	NPP
Bogoria Aug. 2008	GPP		0.000**	0.003**	0.004**	0.000**	0.005**	0.011*	0.000**	0.001**	0.337	0.714	0.056	0.140	0.153	0.971
	Resp			0.527	0.000**	0.010*	0.000**	0.800	0.087	0.231	0.007**	0.015*	0.539	0.007**	0.906	0.115
	NPP				0.000**	0.008**	0.000**	0.478	0.033*	0.101	0.020*	0.038*	0.434	0.012*	0.920	0.172
Bogoria April 2009	GPP					0.000**	0.570	0.000**	0.000**	0.000**	0.128	0.012*	0.000**	0.524	0.001**	0.932*
	Resp						0.000**	0.027*	0.100	0.132	0.001**	0.001**	0.254	0.005**	0.130	0.013*
	NPP							0.000**	0.000**	0.000**	0.114	0.009**	0.000**	0.768	0.001**	0.086
Bogoria Aug. 2010	GPP								0.136	0.313	0.012*	0.010*	0.737	0.011*	0.610	0.103
	Resp									0.503	0.001**	0.000**	0.494	0.005**	0.210	0.028*
	NPP										0.003**	0.004**	0.892	0.005**	0.386	0.041*
Sonachi Sept.-2009	GPP											0.462	0.010*	0.520	0.062	0.667
	Resp												0.004**	0.315	0.106	0.943
	NPP													0.018*	0.429	0.133
Oloidien Sept. 2009	GPP														0.047*	0.337
	Resp															0.276
	NPP															

(NPP= Net primary productivity, GPP= Gross primary productivity, Resp = Respiration)

\* = significant at  $p < 0.05$

\*\*= significant at  $p < 0.01$



**Appendix 2.b:** Mann-Whitney U Test results for primary productivity of the suspended epipelagic community.

Planktonic Productivity		Elementeita Aug.-2008			Natron Lagoon April-2009			Natron Lagoon Dec.-2009			Natron Marsh Dec.-2009		
		GPP	Resp	NPP	GPP	Resp	NPP	GPP	Resp	NPP	GPP	Resp	NPP
Elementeita Aug 2008	GPP		0.003**	0.182	0.003**	0.080	0.001**	0.077	0.182	0.000**	0.001**	0.006**	0.001**
	Resp			0.006	0.103	0.561	0.004**	0.977	0.213	0.000**	0.005**	0.292	0.006**
	NPP				0.004**	0.146	0.001**	0.173	0.339	0.000**	0.000**	0.011*	0.000**
Natron Lagoon April-2009	GPP					0.256	0.202	0.170	0.020*	0.113	0.264	0.859	0.263
	Resp						0.030*	0.588	0.114	0.004**	0.016	0.419	0.041*
	NPP							0.020*	0.001**	1.000	0.754	0.167	0.788
Natron Lagoon Dec.-2009	GPP								0.234	0.004**	0.020*	0.307	0.026*
	Resp									0.000**	0.001**	0.041*	0.001**
	NPP										1.000	0.158	0.750
Natron Marsh Dec.-2009	GPP											0.343	0.909
	Resp												0.211
	NPP												

(NPP= Net primary productivity, GPP= Gross primary productivity, Resp = Respiration)

\* = significant at p<0.05

\*\*= significant at p<0.01

**Appendix 2.c:** Mann-Whitney U Test results for primary productivity of the sedimented *Arthrospira fusiformis*.

		Bogoria August 2008			Bogoria August 2010		
		GPP	Resp	NPP	GPP	Resp	NPP
Bogoria August 2008	GPP		0.000**	0.008**	0.000**	0.000**	0.000**
	Resp			0.000**	0.003**	0.419	0.053
	NPP				0.007**	0.000**	0.001**
Bogoria August 2010	GPP					0.008**	0.418
	Resp						0.029*
	NPP						

(NPP= Net primary productivity, GPP= Gross primary productivity, Resp = Respiration)

\* = significant at p<0.05

\*\*= significant at p<0.01

Appendix 2.d: Mann-Whitney U Test results for primary productivity of the epipelagic community.

		Elementeita August 2008			Natron Lagoon April 2009			Natron Lagoon December 2009			Natron Marsh April 2009			Natron Marsh December 2009		
		GPP	Resp	NPP	GPP	Resp	NPP	GPP	Resp	NPP	GPP	Resp	NPP	GPP	Resp	NPP
Elementeita August 2008	GPP		0.000**	0.154	0.009**	0.000**	0.000**	0.100	0.001**	0.000**	0.018*	0.000**	0.001**	0.000**	0.000**	0.000**
	Resp			0.002**	0.779	0.012*	0.105	0.001**	0.250	0.956	0.055	0.089	0.381	0.719	0.354	0.001**
	NPP				0.009**	0.000**	0.000**	0.622	0.025*	0.001**	0.112	0.000**	0.010**	0.001**	0.000**	0.000**
Natron. Lagoon. April 2009	GPP					0.001**	0.030*	0.001**	0.498	0.640	0.046*	0.055	0.626	0.574	0.217	0.003**
	Resp						0.288	0.000**	0.002**	0.005**	0.000**	0.399	0.001**	0.038*	0.102	0.052
	NPP							0.000**	0.019*	0.210	0.001**	0.901	0.027*	0.276	0.611	0.035*
Natron. Lagoon. December 2009	GPP								0.020*	0.001**	0.308	0.000**	0.008**	0.001**	0.000**	0.000*
	Resp									0.250	0.278	0.028*	0.818	0.242	0.096	0.001**
	NPP										0.093	0.052	0.224	0.622	0.355	0.001**
Natron Marsh. April 2009	GPP											0.001**	0.199	0.036*	0.016*	0.000**
	Resp												0.011*	0.295	0.460	0.027*
	NPP													0.268	0.056	0.000**
Natron Marsh. December 2009	GPP														0.564	0.006**
	Resp															0.009**
	NPP															

(NPP= Net primary productivity, GPP= Gross primary productivity, Resp = Respiration)

\* = significant at p<0.05

\*\*= significant at p<0.01

### Appendix 3: Energy calculations

#### Appendix 3.a: Energy calculations for standing crop and net primary productivity of the planktonic food resource.

Planktonic					
	Month-Year	Amount of available food $\text{mg C m}^{-2} \times 10^3$	Energy of available food $\text{k cal m}^{-2}$	Rate of regeneration by net primary productivity $\text{mg C m}^{-2} \text{hr}^{-1}$	Rate of energy input by net primary productivity $\text{Kcal m}^{-2} \text{hr}^{-1}$
Bogoria	Apr-09	37.74	329.18	204.60	1.79
Bogoria	Aug-09	3.24	28.26	-	-
Bogoria	Aug-10	5.17	45.09	15.49	0.14
Nakuru	Apr-09	11.97	104.41	-	-
Nakuru	Aug-09	0.9	7.85	-	-
Sonachi	Sep-09	0.56	4.88	29.67	0.26
Oloidien	Sep-09	3.63	31.66	143.81	1.25

#### Appendix 3.b: Energy calculations for standing crop and net primary productivity of the suspended epipelagic food resource.

Suspended epipelagic					
	Month-Year	Amount of available food $\text{mg C m}^{-2} \times 10^3$	Energy of available food $\text{k cal m}^{-2}$	Rate of regeneration by net primary productivity $\text{mg C m}^{-2} \text{hr}^{-1}$	Rate of energy input by net primary productivity $\text{Kcal m}^{-2} \text{hr}^{-1}$
Elementeita	Aug-08	-	-	60.82	0.53
Elementeita	Aug-09	0.66	5.76	-	-
Natron lagoon	Dec-09	0.79	6.89	-23.88	-0.21
Natron marsh	Dec-09	4.29	37.42	-20.27	-0.18

#### Appendix 3.c: Energy calculations for standing crop and net primary productivity of the sedimented *Arthrospira fusiformis* food resource.

Sedimented <i>Arthrospira fusiformis</i>					
$\text{mg C m}^{-2}$	Month-Year	Amount of available food $\text{mg C m}^{-2} \times 10^3$	Energy of available food $\text{k cal m}^{-2}$	Rate of regeneration by net primary productivity $\text{mg C m}^{-2} \text{hr}^{-1}$	Rate of energy input by net primary productivity $\text{Kcal m}^{-2} \text{hr}^{-1}$
Bogoria	Apr-09	17.24	150.37	-	-
Bogoria	Aug-09	4.25	37.07	-	-
Bogoria	Aug-10	3.85	33.58	41.38	0.36

**Appendix 3.d:** Energy calculations for standing crop and net primary productivity of the epipellic food resource.

Epipellic					
mg C m <sup>-2</sup>	Month-Year	Amount of available food mg C m <sup>-2</sup> × 10 <sup>3</sup>	Energy of available food k cal m <sup>-2</sup>	Rate of regeneration by net primary productivity mg C m <sup>-2</sup> hr <sup>-1</sup>	Rate of energy input by net primary productivity Kcal m <sup>-2</sup> hr <sup>-1</sup>
Elementeita	Aug-09	4.92	42.91	-	-
Nakuru	Apr-09	46.67	407.07	-	-
Nakuru	Aug-09	5.81	50.68	-	-
Natron lagoon	Apr-09	12.74	111.12	14.37	0.13
Natron lagoon	Dec-09	20.69	180.46	21.62	0.19
Natron marsh	Apr-09	32.93	287.22	28.32	0.25
Natron marsh	Dec-09	26.68	232.71	3.71	0.03

**Appendix 3.e:** Energy calculations for standing crop and net primary productivity of the lake shore mud community food resource.

Lake shore mud			
	Month-Year	Amount of available food mg C m <sup>-2</sup> × 10 <sup>3</sup>	Energy of available food k cal m <sup>-2</sup>
Bogoria	Apr-09	55.46	483.73
Bogoria	Aug-10	0.53	4.62
Elementeita	Aug-09	2.83	24.68
Nakuru Lake	Aug-09	14.27	124.47
Nakuru Njoro river mouth	Aug-09	0.62	5.41
Natron lagoon	Apr-09	22.97	200.35
Natron lagoon	Dec-09	16.73	145.92
Natron marsh	Apr-09	44.08	384.48
Natron marsh	Dec-09	44.9	391.63
Sonachi	Sep-09	6.08	53.03
Oloidien	Sep-09	3.91	34.01



# UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS (SCI)

**Strategic Interventions to enhance adoption of Open Source Applications and Creative commons licensed Open Content in the Kenyan Government**

BY

**KING'OINA JANET MARANGA**

**P58/61647/2010**

**SUPERVISOR:**

**DR. WANJIKU NG'ANG'A**

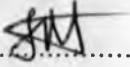
**OCTOBER 2012**

---

*A Research Project presented in partial fulfillment of the requirements governing the award of the Degree of the Master of Science in Computer Science. This report is the product of my own work except where indicated in the text. The report may be copied and distributed freely in part or whole as long as the source is acknowledged.*

### DECLARATION

I Janet Maranga hereby declare that this project has not been submitted for a degree or any other award in any other university or institution.

Signature:.....  ..... Date: 06<sup>th</sup> November 2012

Janet Maranga

Registration Number: P58/61647/2010

### DECLARATION BY SUPERVISOR

This research project has been submitted as part of fulfillment of requirements for the Master of Science in Computer Science with my approval as the School of Computing and Informatics (UoN) supervisor.

 ..... Date: 06<sup>th</sup> Nov 2012

Dr. Wanjiku Ng'ang'a

School of Computing & Informatics (UoN)

Supervisor

## **ABSTRACT**

Open source software (OSS) is a viable alternative for proprietary software (PS), particularly in the government sector globally for reasons such as lowering software costs, growing local software development industry, and bridging the digital divide. On the other hand CC licensed open content is also very useful and can be effectively used to supplement the content the government has in various thematic areas. However the Kenyan government has not harnessed the benefits that these afford. This research sought to realize the current status of OSS and CC licensed content usage in the Kenyan government by surveying top level ICT management in the e-government directorate, ICT staff as well as availability and skill of OSS software developers and willingness of content creators to share content appropriately licensed with an aim to proposing strategic interventions to enhance the adoption of these in the Kenyan government. The U.S Department of Labor E-Government framework was adopted to inform this research. The results indicate that OSS and CC licensed open content usage within the ministries is not yet extensive and measures need to be put in place to enhance the utilization of these. Various challenges and obstacles are hampering full OSS and CC licensed open content implementation and utilization within the ministries and agencies. These can however be combated and OSS and CC licensed open content adopted once these strategies have been adopted and implemented.

The entire study took a maximum duration of six (6) months.



## DEDICATION

This research is dedicated to the memory of my late mother Grace, who has inspired me for the best.

## ACKNOWLEDGEMENT

First I would like to express my most sincere thanks to God who has been a constant source of strength, direction and inspiration throughout and particularly during the Masters Study programme.

Sincere thanks and gratitude to my supervisor Dr. Wanjiku Ng'ang'a for all her support and encouragement during the course of this study. In spite of the deficiencies of my early proposal at the beginning of the year, she immediately believed in this research project and gave me the motivation I needed. Her professional, quick, and detailed responses have managed to keep my research on track. Quite frankly, the thesis would not have been completed without her support and guidance. The patience and kindness she has shown has amazed me so many times. Thank you Dr. Ng'ang'a, I have been in good hands and fortunate enough to be given the opportunity to be your student.

I like also to express my deepest gratitude to the panelists, Evans Miriti, Andrew Mwaura, Sam Ruhiu and Dr. Timothy Waema who provided me valuable comments and constructive criticism for improvement of my research work. I am also greatly thankful to many other researchers for the contributions they have provided to this research area.

There are numerous people in the Kenyan e-Government to whom I am thankful for the support and open-mindedness they have shown. Especially during the early stages of the study, the support shown by the top management of the e-Government was of utmost importance. For the support shown during the period January to July 2012, I am especially grateful to Dr. Katherine Getao whose constructive comments and criticism has managed to sharpen my thinking on the research topic and to improve my study. In addition to that the introduction to e-Government staff who served as a source of relevant and much needed information on the internal working of the e-Government and possible setbacks on the implementation of Open source software applications and suggestions on how these could be averted has been of great usefulness to me.

During the actual research of the Open Source Software applicability to the Kenyan e-Government, numerous staff members and colleagues in the e-Government and its administrative sector have surprised me with their positive attitude and willingness to participate in data collection activities to support the research. I gratefully acknowledge their support.

I am thankful to my Dad, Dr. Maranga for sponsoring my MSc degree studies without whose support this research would not have been possible. I am also grateful to my fiancé Derrick for the patience and support he has shown during the past months. My sisters Nancy, Emily and Naomi have been an inspiration to me through their optimism and practical touch with the complexities of life.

I would also wish to thank George Gitau who offered invaluable advice in many aspects of this research.

Finally, this research is especially dedicated to the memory of my mother Grace who died in May 2011 a few months before this study was commenced. Memories live, keeping the past not quite past.

**CHAPTER ONE: INTRODUCTION ..... 1**

1.1 BACKGROUND..... 1

1.2 OUTLINE OF THE REPORT ..... 2

1.3 PROBLEM STATEMENT..... 3

1.4 OBJECTIVES ..... 3

1.5 RESEARCH QUESTIONS..... 3

1.6 JUSTIFICATION OF THE STUDY ..... 4

1.7 SCOPE OF THE STUDY ..... 6

1.8 RESEARCH OUTCOMES AND THEIR SIGNIFICANCE TO KEY AUDIENCES ..... 6

1.9 ASSUMPTIONS AND LIMITATIONS OF THE RESEARCH ..... 6

1.10 DEFINITIONS OF IMPORTANT TERMS..... 7

**CHAPTER TWO: LITERATURE REVIEW ..... 9**

2.1 TERMINOLOGY ..... 9

2.2 FRAMEWORKS FOR E-GOVERNMENT ..... 12

2.3 OPEN SOURCE SOFTWARE IN E-GOVERNMENT ..... 15

2.4 OVERVIEW OF FOSS POLICY IN AFRICA ..... 16

2.5 TECHNOLOGY ADOPTION AND STRATEGIC PLANNING FRAMEWORKS..... 23

2.6 FRAMEWORKS AND OPEN SOURCE TOOLS ..... 26

2.7 FRAMEWORK TO BE ADOPTED FOR THE RESEARCH ..... 28

**CHAPTER THREE: METHODOLOGY ..... 29**

3.1 OVERVIEW ..... 29

3.2 RESEARCH PURPOSE AND DESIGN ..... 29

3.3 RESEARCH STRATEGY..... 29

3.4 SAMPLE SELECTION ..... 29

3.5 FRAMEWORK ..... 35

3.6 REQUIREMENT DETERMINATION..... 35

3.7 DATA COLLECTION ..... 36

3.8 DATA PROCESSING ..... 38

3.9 DATA CLEANING AND VALIDATION ..... 38

3.10 CONSTRAINTS ..... 38

3.11 VALIDITY AND RELIABILITY ..... 38

3.12 PROTOTYPE DEVELOPMENT TOOLS ..... 39

3.13 APPLICATION DEVELOPMENT METHODOLOGY ..... 39

3.14 TYPE OF SECURITY..... 41

3.15 IMPLEMENTATION OF THE CONTENT PORTAL USING OPEN SOURCE SOFTWARE FRAMEWORKS ..... 41

**CHAPTER FOUR: DATA ANALYSIS AND INTERPRETATIONS ..... 43**

4.1 INTRODUCTION ..... 43

4.2 DATA PROCESSING AND ANALYSIS ..... 44

4.3 DETAILED ANALYSIS OF DATA COLLECTED..... 45

**CHAPTER FIVE: SYSTEM ANALYSIS AND DESIGN..... 66**

5.0 INTRODUCTION ..... 66

5.1 SYSTEM ANALYSIS (REQUIREMENT DEFINITION)..... 66

5.2 SYSTEM DESIGN FLOW CHART..... 70

5.3 DATABASE DESIGN - E-R DIAGRAM..... 71

5.4 PROCESS NARRATIVE..... 72

5.5 SOFTWARE REQUIREMENTS AND CONFIGURATION ..... 73

5.6 SYSTEM TESTING ..... 73

**CHAPTER SIX: DISCUSSION** ..... 76

6.1 OBJECTIVES ..... 76

6.2. RESEARCH FRAMEWORK ..... 77

6.3. METHODOLOGY ..... 78

6.4 STRATEGIC INTERVENTIONS TO ENHANCE ADOPTION OF OSS AND CC LICENSED CONTENT IN THE KENYAN GOVERNMENT ..... 78

**CHAPTER SEVEN: CONCLUSION AND RECOMMENDATIONS** ..... 81

7.1 CONCLUSION ..... 81

7.2 FURTHER WORKS ..... 81

**REFERENCES** ..... 82

**APPENDICES** ..... 86

## List of Tables

TABLE 2.1: CONTRAST BETWEEN ADOPTION MODELS AND USE DIFFUSION MODELS .....	26
TABLE 3.1: SAMPLE SELECTION.....	34
TABLE 4.1: WILLINGNESS TO SHARE CC LICENSED OPEN CONTENT .....	47
TABLE 4.2: FACTORS TO ELICIT CC LICENSED OPEN CONTENT SHARING.....	48
TABLE 4.3: USAGE/ AWARENESS OF CC LICENSED OPEN CONTENT .....	49
TABLE 4.4: PROFICIENCY LEVELS OF OSS DEVELOPERS .....	50
TABLE 4.5: STABILITY .....	51
TABLE 4.6: PERFORMANCE .....	51
TABLE 4.7: SCALABILITY .....	52
TABLE 4.8: INTEROPERABILITY OF OSS.....	52
TABLE 4.9: EXTENDIBILITY.....	53
TABLE 4.10: STANDARDS .....	53
TABLE 4.11: DOCUMENTATION .....	54
TABLE 4.12: COMMUNITY SUPPORT.....	54
TABLE 4.13: FREQUENCY OF UPDATES.....	55
TABLE 4.14: FOSS ENVIRONMENTS.....	56
TABLE 4.15: PRESENCE OF OSS POLICIES.....	57
TABLE 4.16: WHY OSS IS NOT SUITABLE FOR GOVERNMENT.....	58
TABLE 4.17: E-GOVERNMENT STRATEGY AND OSS .....	58
TABLE 4.18: USAGE OF LINUX AS AN OS IN GOVERNMENT.....	59
TABLE 4.19: USAGE OF OPEN CONTENT LICENSING IN E-GOVERNMENT.....	59
TABLE 4.20: FLEXIBILITY TO ALLOW FOR OSS PILOT PROJECTS .....	60
TABLE 4.21: OPERATING SYSTEM.....	60
TABLE 4.22: DESKTOP OPERATING SYSTEM.....	61
TABLE 4.23: IN-HOUSE OSS CAPABILITY .....	62
TABLE 4.24: INCENTIVE PROVISION .....	62
TABLE 4.25: MONETARY VS NON-MONETARY INCENTIVES .....	62
TABLE 4.26: OPEN SOURCE SOFTWARE USAGE ENHANCEMENT .....	62
TABLE 4.27: CITIZEN INPUT IN CONTENT GENERATION .....	63
TABLE 4.28: PROFICIENCY TO SUPPORT OSS APPLICATIONS.....	63
TABLE 4.29: CONTENT AVAILABILITY .....	65
TABLE 5.1: ADMINISTRATOR USER CLASSES .....	67
TABLE 5.2: REVIEWER USER CLASS .....	67
TABLE 5.3: AUTHOR/ CONTENT CREATOR USER CLASS.....	68

## List of Figures

FIGURE 2.1: FOCAL DOMAINS FOR E-GOVERNMENT INITIATIVES.....	10
FIGURE 2.2: COMPONENTS OF THE U.S DEPARTMENT OF LABOR'S E-GOVERNMENT FRAMEWORK.....	13
FIGURE 2.3: THE E-GOVERNANCE FRAMEWORK FOR THE COMMONWEALTH .....	15
FIGURE 2.4: MALAYSIAN GOVERNMENT OSS POLICY.....	20
FIGURE 2.5: SWOT ANALYSIS .....	24
FIGURE 2.6: TECHNOLOGY ADOPTION CURVE FRAMEWORK .....	25
FIGURE 3.1: ICT HEADS SAMPLE.....	31
FIGURE 3.2: ICT STAFF SAMPLE .....	32
FIGURE 3.3: CITIZENS SAMPLE.....	32
FIGURE 3.4: OSS SOFTWARE DEVELOPERS SAMPLE .....	33
FIGURE 3.5: CONTENT CREATORS SAMPLE .....	34
FIGURE 3.6: WATERFALL METHODOLOGY.....	40
FIGURE 4.1: FREQUENCY OF INTERNET USAGE .....	46
FIGURE 4.2: AVAILABILITY OF CC LICENSED OPEN CONTENT AMONG KENYAN CITIZENS.....	46
FIGURE 4.3: OWNERSHIP OF CC LICENSED OPEN CONTENT .....	47
FIGURE 4.4: CONTENT CREATORS WILLINGNESS TO SHARE CONTENT.....	48
FIGURE 4.5: USAGE OF OSS IN APPLICATION/SYSTEM DEVELOPMENT.....	49
FIGURE 4.6: LEARNING CURVE .....	50
FIGURE 4.7: APPLICABILITY OF OSS IN GOVERNMENT.....	57
FIGURE 4.8: INPUT OF CITIZENS INTO PUBLIC SECTOR DECISIONS AND ACTIONS .....	61
FIGURE 5.1: SYSTEM USE CASE DIAGRAM.....	69
FIGURE 5.2: SYSTEM DESIGN FLOWCHART .....	70
FIGURE 5.3: E-R DIAGRAM.....	71
FIGURE 5.4: SAMPLE TEST DATA OUTPUT.....	73
FIGURE 5.5: MODERATION OF CONTENT .....	74
FIGURE 5.6: LOGIN TEST .....	75

## **List of Abbreviations**

**G.o.K** – Government of Kenya

**OSS** – Open Source Software

**CC licensed Open Content** – Creative Commons licensed Open content

**PHP** – Hypertext preprocessor

**GPL** - General Public license

**FOSS** – Free and Open Source Software

**IIS** - Microsoft Internet Information Server

**ICT's** –Information and Communication technologies.

**CRM** - Customer Relationship Management

**SSL** – Secure Socket Layer

**TLS** – Transport Layer Security

**DOL** – Department of Labour

**I.T** – Information Technology

**MIT** – Massachusetts Institute of Technology

**AGIMO** - Australian Government Information Management Office

## CHAPTER ONE: INTRODUCTION

### 1.1 Background

Governments around the world, over the years are recognizing the need for utilization of Information technology as an enabler in the various day to day tasks and are gradually embracing the various capabilities and efficiencies that these afford. These governments are also making or considering efforts to promote open source software (typically produced by cooperatives of individuals) at the expense of proprietary software (generally sold by for profit software developers).

e-Government is a key instrument for modernization and reform as governments face the continuing pressure of increasing their performance and adapting to the pressure of the new information society (Morven McLean and Tawfik Jelassi, 2003).

The recent trouble is that many government departments in Kenya still have little understanding of the many benefits that open source software can have and how to utilize it effectively in order to get optimal results from its use and subsequently lack a path toward making decisions in keeping with core government responsibilities. Cost savings, the naïve enticement, may not provide good enough motivation in the end. Although proprietary software (the complement to open source software) tends to come with high licensing fees, whereas open source can be downloaded without payment, monetary arguments for deploying open source software might be unsuccessful because of the high cost of conversion, retraining and developing an adequate base for support which can postpone the potential savings of open source software for many years.

E-government is enabling government organizations to provide better services to their constituents. Transactions such as filing taxes online, applying for jobs, renewing driver's licenses and ordering recreational and occupational licenses can now be conducted online, quickly and efficiently (West, 2002).

While some earlier e-government computer issues, such as office automation, may not be highly relevant to research today, many issues are, for example decision making service processes and values (Ake Gronlund, Tom Horan, 2005).

The Kenya government is embracing this change and is more open to adopting and making use of the numerous benefits it can reap from this automation. Over the last few years, efforts to automate most of the government processes have become a core undertaking of the various sectors in the government. The e-Government strategy has set out several best practices for benchmarking in Kenya some of which are coherent and compatible information processing and management policies and business processes, proper and adequate skills, knowledge and attitudes necessary for operationalization and sustainability of Communication within government, networked readiness by exploiting the literate population, innovative private sector and efficient government to stimulate economic growth, the use of Internet to ensure that e-Government services reach every citizen, business and institutions in Kenya, to mention but a few.



Open source software generally means software that is often developed in a public collaborative manner. Therefore the open source applications that will be built from it allow anyone to create modifications of the software, port it to new operating systems and processing architectures and share it with others. This comes with several benefits such as the software will continually be improved and tested by a wide community of software developers in Kenya and even beyond. Bugs will be fixed and knowledge will be drawn from a vast domain of knowledgeable persons.

Open source-software is usually copyrighted and its license may contain restrictions intended to preserve its open source status or to require notice of authorship or to control other developmental factors.

The open-source licensing model as evidenced by the GNU General Public license (GPL) contrasts sharply with licenses of proprietary mainstream software (i.e non-open or closed source equivalent to commercial product.) Examples are Sun Microsystems makes Star Office, alternative to Microsoft office, Apache Web server alternative to Microsoft Internet Information Server (IIS) or Netscape Web Server, GIMP (GNU Image Manipulation Program) alternative to Adobe Photoshop or Corel Photopaint, Postgre SQL relational SQL database alternative to Oracle (Oracle Corporation) or DBZ (IBM).

The Kenya government plays a crucial role with regard to ICT in general and open source and creative commons licensed open content in particular. With the recent launch of open data in Kenya, it is clear that the Kenya government is appreciating and warming up to the spirit of sharing. The Kenya government is in a position to drive strategic change throughout the whole country. According to the FOSS Policy toolkit (2005), the public sector is the biggest consumer of ICT and governments set the economic and regulatory boundaries that allow businesses to develop. Open source software has for a long time been in use among government agencies, and prospects for increased use have been greeted enthusiastically by both knowledgeable government employees and open source communities. A lot of open source software applications exist and continue to come into existence day by day which can be used in many different ways in order enhance service delivery in e-government in a life changing manner. But mobilizing the necessary forces in government to procure open source software has been difficult. This study attempts to explore the reasons so many efforts have stalled, the lessons learnt by the successful efforts as well as measures that can be put in place to facilitate the adoption of open source applications as well as increase the utilization of creative commons licensed open content.

## **1.2 Outline of the Report**

This report begins with an introduction to open source software and creative commons licenses. It also gives the problem statement, objectives and justification of the study in the first chapter. Chapter two will be the literature review, where various e-governments in the world that are using open source software are reviewed and also highlight the proposed solution to the problem posed in the previous chapter. Chapter three presents the methodology that was used in the research to build the prototype. Chapter four gives the results and findings of the analysis done, while a conclusion and recommendations are highlighted in chapter five.

## **Problem Statement**

The purpose of this study is to unearth the constraints that limit widespread adoption of open source software in e-government for some government services and utilize open source software to come up with a solution which will make it easy and enhance the use of Creative commons licensed open content by the government and also make it easy for content creators to channel this content to the relevant authorities in the government so as to increase awareness on its great potential so that it can be used in many life-changing ways.

Licensing can be complex and somewhat tricky. Creative commons was therefore founded in order to help give content creators an easy way to distribute their work while specifying some simple factors such as whether the work should be used commercially or modified.

The problem of closed source software applications is that they are not flexible in the sense that the vendors retain the source code and distribute the software in compiled form thereby preventing the user from understanding how it works or changing how it works and it also does not encourage collaboration on projects as the source code is not made publicly thereby stifling innovation.

It is this knowledge gap in open source software adoption in government in the Kenyan scenario and the subsequent software development process and utilization that we propose to address in this research.

## **Objectives**

The guiding research question in the study is to formulate strategic interventions to be used in the facilitation of the adoption of open source applications and creative commons open licensed content to be used in e-Government in Kenya. This study aims to:

- i) Investigate current e-Government frameworks in use globally in relation to OSS
- ii) Explore the flexibility of the current ICT policies and strategies in as far as adoption of OSS and CC are concerned.
- iii) Determine a model that enables content sharing and utilization of CC licensed open content from the literate population by the Government.
- iv) Determine strategic interventions to enhance adoption of OSS and CC licensed Open content and disseminate them to the Kenyan government.

## **Research Questions**

In order to achieve the objectives, the following research question will constitute the domain of investigation:

What are the strategies that need to be formulated to enhance the adoption of Open source applications and creative commons licensed open content in Kenya?

Other research questions that will be considered include:

- What is the potential of OSS as an appropriate relevant alternative to commercial software in Kenyan e-Government?
- To what extent and how adequately does the Kenyan e-Government strategy deal with Open source software and creative commons licensed open content?
- How and under what conditions can CC licensed open content be considered and utilized as a value add in the service delivery of the Kenyan e-Government?
- What new approaches towards content sharing can be implemented to enhance adoption of CC licensed content in the Kenyan e-Government scenario.
- What strategic interventions can be put in place to enhance and facilitate the use of OSS and CC licensed content to aid service delivery in the Kenyan e-Government?

### 1.6 Justification of the study

The need for governments to automate their processes and to provide value added services to their citizens and other stakeholders has always been a key undertaking, to keep pace with the current technological changes but also to identify possibilities to reap the benefits that come with these technologies.

Several researches done in the developed and developing countries have revealed that software applications for e-governance cannot and should not be limited to proprietary software applications only. Their counterparts, open-source software can also be utilized to play a vital role in service delivery to citizens. In this regard, effort must be put in devising a way of ensuring these two platforms work well together. (Working Group on e-Government in the Developing World, 2002)

Despite the advancements in technology, many government departments still have little understanding of fundamental goals of open source software let alone its massive capabilities. Although proprietary software (the complement to open source software) tends to come with high licensing fees, whereas open source can be downloaded without payment, monetary arguments for deploying open source software are usually unsuccessful because of the high cost of conversion, retraining and developing an adequate base for support which can postpone the potential savings of open source software for many years. Nevertheless, it is vital to produce locally based goods and services to substitute increasingly expensive imports and lower costs substantially. The weakening of the local currencies against the international major currencies like the dollar or the sterling pound will make the licenses on the imported software (not to mention other imports) prohibitively expensive. Whether or not the Kenyan shilling enjoys an upswing in future, it makes sense to minimize risks through avoidance where possible of dollar based software license fees and through vigorous encouragement of local software development.

The key trait distinguishing open source from proprietary software is not its availability free of cost, but its provision under a license that allows anyone to alter it and redistribute the altered form. Freedom to change, improve, and extend the software is the trait that draws a hard and fast line between software that can be defined as open source and software that remains locked to a particular developer. (Yayehyirad Kitaw, 2006).

Revealing source code to a particular customer or even to the general public is not enough to define a product as open source; it must also have a license that allows unlimited changes and redistribution by anyone.

In the light of these, the advantages of the traits of open source software are the ability to continue support and development if the original developer goes out of business, the capability to extend it in ways that the original developer does not find worth its while, the software developer community involvement in finding and fixing bugs quickly and also extending the functionality of the applications developed and continually enhancing the capabilities of the applications developed to keep in line with the changing needs of the citizens among other benefits. Nevertheless governments are additionally mandated with several responsibilities that make open source software particularly necessary such as:

- Vendor independence
- Access for all
- Archiving
- Special government needs
- Security

The approach that will be used is investigating and finding out how best OSS and CC can be adopted and be used widely and intuitively in the effective service delivery by the Kenyan e-Government. The project findings and recommendations will create awareness in the government on the importance and benefits of open source software as well as creative commons licensed open content and to remove the barriers to future open source development projects. The recommendations of this research will also provide a source for legalization of alternatives to lowering software cost as well as harnessing the myriad of other benefits accrued from adopting an open source software approach in addition to the utilization of the wealth of information and knowledge licensed using the Creative commons licensing model. It also intended to reveal that OSS and CC licensed open content has massive potential that can be successfully exploited and adopted to accelerate the pursuits and mandate of the Kenyan e-Government by utilizing the currently available information resident with the literate population and also offering a platform where updated content can be shared under various CC licenses which can be cleaned and used by the e-Government. The adoption of open source software is likely to make Kenya extremely well placed to compete in global software development market joining South Africa and Ethiopia among others in Africa, that are already participating in the global market for software development. The proposed CC licensed content sharing platform built using OSS will assist in expanding the information infrastructure, facilitate connectivity of all the Kenyan government agencies, initiate other systems geared towards increasing efficiency and non-replication of data as well as involvement of citizens and eliciting citizen participation and develop capacity of e-government I.T staff to monitor progress, maintain and utilize the content that is shared on this platform. It is also likely to provide a catalyst for the Kenyan government to develop national policies to promote the use of open source software in other sectors. The adoption of the open source software and creative commons licensed open content is also likely to provide a useful tool to enable a developing country like Kenya to leapfrog into the information technology age.

## **1.7 Scope of the Study**

The study will focus on the analysis and evaluation of the existing e-Government frameworks worldwide as well as technology adoption frameworks, select one which can map to the Kenyan scenario and utilize it to come up with recommendations and design a creative commons licensed content sharing model suitable to the Kenyan situation. The flexibility of the proposed strategies can, however, be reproducible in similar settings with a high degree of success for any other exercise that aims to increase adoption using the defined methodology.

## **1.8 Research Outcomes and their significance to key audiences**

The main outcome of this research process is to furnish concrete strategies that can enhance the adoption of FOSS and CC licensed open content in the Kenyan Government.

This will involve the creation of a content sharing platform where content can be shared using CC licensed open content which will foster interactions of the Government with its citizens, non-profits, businesses, other agencies, communities as well as the government within itself and the departments within it.

Utilization of Open source software will also feature prominently in the research outcome in that it will be used as a tool for building the content sharing platform which will enable and encourage sharing of content which will help in making critical therefore enhancing its value to more people than just the creator and enhance the e-governance process.

### **1.8.1 Summary of the Major Benefits of Open software and Open standards**

- Reduced costs and less dependency on imported technology and skills
- Affordable software for individuals, enterprise and government
- Universal access through mass software rollout without barrier of proprietary software and data formats.
- Access to government data without barrier of proprietary software and data formats.
- Participation in global network of software development

## **1.9 Assumptions and Limitations of the research**

The main underlying assumption in this study is that there is availability presence of creative commons licensed content improves e-government. How much it improves e-government depends on how well the availed content is collected and collated on a portal and thereafter implemented in facilitating various government projects.

Determining the availability of the CC licensed open content and having a platform for the publishing of this whilst utilizing open source software in this survey is meant to validate this assumption.

Other assumptions and limitations in this research are as outlined here under:

- i. The local software developer community is abreast with the current trends in open source software technologies.

- ii. The literate population in Kenya understand licensing especially using the Creative commons licensing model.
- iii. Time – All the aspects of E-governance and open source technologies as well as creative commons licensed open content might not be adequately covered due to the limited time allocated for the research. The flexibility of the proposed strategies can however be reproducible in similar settings with a high degree of success.
- iv. The developmental stage of this research area is in its infancy, this might be a limitation in terms of related work that may not be much detailed.

### **1.10 Definitions of Important Terms**

**E-Government** - E-Government is defined as the use of information and communication technology (ICT) to enable more efficient, cost-effective, and participatory government, facilitate more convenient government services, allow greater public access to information, and make government more accountable to citizens.

**Conceptual Framework** – A set of theories widely accepted enough to serve as the guiding principles of research within a particular discipline.

**Software framework** – A reusable set of libraries or classes for a software system (or software system)

**Application framework** – A software framework used to implement the standard structure of an application for a specific operating system.

**Strategy** – a long term plan or action designed to achieve a particular goal.

**Intervention** – This is the action or process of intervening. An influencing force or act that occurs in order to modify a given state of affairs.

**ICT's** – A general term that stresses the role of unified communications and the integration of telecommunications, computers, middleware as well as necessary software, storage and audio-visual systems, which enable users to create, access, store, transmit and manipulate information.

**FOSS** – Free/open-source software is software that is distributed together with its underlying source code, under a certain kind of copyright. FOSS copyright licenses allow everyone to read, modify, and redistribute the source code, so programmers can improve and adapt the software, and fix bugs.

**CC** – Creative commons licensed open content -Creative commons licensed open content is content that is released under licenses which allow creators to communicate which rights they reserve and which rights they waive for the benefits of recipients or other creators.

**Commercial software** – Software being developed for a business, which aims to make money from the use of software.

**Copylefted software** – Free-software whose distribution terms do not let redistributors add any additional restriction when they redistribute or modify the software.

**Freeware** – Refers to packages distributed free of charge (no license fee) which permit redistribution but not modification (and their source code is not available).

**GNU programs** – Software that is released under the auspices of GNU project

**Non-copylefted free software** – Non-copylefted free software comes from the author with permission to redistribute and modify and also add additional restriction to it.

**Proprietary software** – Software that is not free or semi-free. Its use redistribution or modification is prohibited or requires you to ask for permission, or restricted so much that you effectively cannot do it freely.

**SSL** - SSL is the secure communications protocol of choice for a large part of the Internet community. There are many applications of SSL in existence, since it is capable of securing any transmission over TCP. Secure HTTP, or HTTPS, is a familiar application of SSL in e-commerce or password transaction.

**TLS** - The protocol “allows client/server applications to communicate in a way that is designed to prevent eavesdropping, tampering or message forgery.

## CHAPTER TWO: LITERATURE REVIEW

This review focuses on various topical issues that will be covered in this research ranging from the outline of various e-Government frameworks available, countries that have adopted FOSS and their experience, driving forces of FOSS, the benefits and limitations of FOSS as well as creative commons licensed open content among others. Finally it will narrow down the research problem and come up with general principles to support the research question. This section will inform the conceptual frame work to be used in this study.

### 2.1 Terminology

#### Open Source Software

In the Toolkit for FOSS policy in Africa (2005), FOSS has been defined as follows: “Free/open-source software is distributed together with its underlying source code, under a certain kind of copyright. FOSS copyright licenses allow everyone to read, modify, and redistribute the source code, so programmers can improve and adapt the software, and fix bugs. And the software can be shared with others. The difference between “free” and “open” lies mainly in the fundamental beliefs and aims of the respective proponents. Open source software supporters tend to focus on pragmatic aspects of software development and use, whereas the free software community places the aspect of “freedom” at the centre of their activities. Free-software licenses require software developers to distribute their modifications and additions under a similar free-software license, whereas some open source-software licenses allow the inclusion of open source software in proprietary software.”

#### Creative Commons Licensed Open Content

Creative commons licensed open content is content that is released under licenses which allow creators to communicate which rights they reserve and which rights they waive for the benefits of recipients or other creators.

#### Overview of CC licenses

Each and every CC license has a short name and description which explains in a simple way what that license allows a person to do. There is also a full legal license in case where a content creator may wish to read it thoroughly.

There are some important things to note in regards to these licenses.

**Attribution** – When this is present the user of the content must attribute and link back to the original item. Attribution typically says something like “Photo by Janet Photographer” with a link to the page or portfolio where the item came from.

**Commercial** – This generally means that the licensed work can be used for commercial purposes. All non-commercial CC licenses explicitly say so.



**Public Domain** – Creative Commons also provides a public domain mark which can be used by content creators. Items put into the public domain can be used in any way (including without attribution). However, the public domain isn't technically a Creative Commons license, but the mark is a convenience which is offered by the organization for content creators.

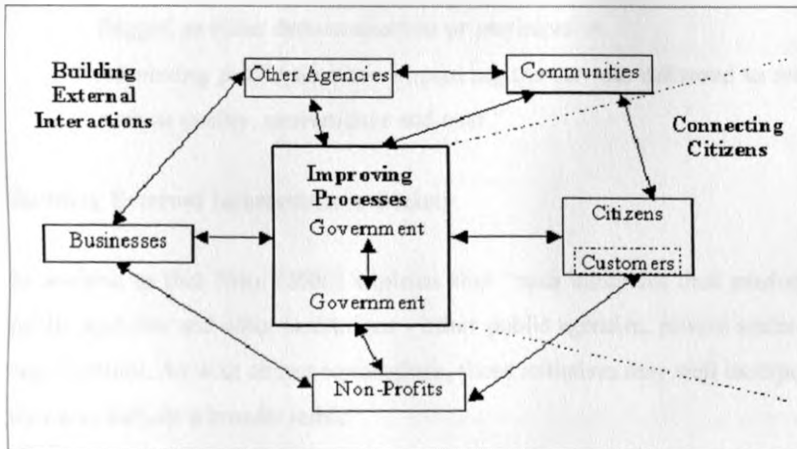
### An Overview of the Definition of E-Government

According to (Ntiro, 2000), e-Government is the use of information and communication technology (ICT) to enable more efficient, cost-effective, and participatory government, facilitate more convenient government services, allow greater public access to information, and make government more accountable to citizens.

E-Government is the use of Information and Communication Technologies (ICTs) to improve the activities of public sector organizations.

(Ntiro, 2000) goes further to expound that there are three main domains of e-Government, illustrated in *Figure 1.1*.

- Improving government processes: e-Administration
- Connecting citizens: e-Citizens and e-Services
- Building external interactions: e-Society



*Figure 2.1: Focal Domains for e-Government Initiatives*

Ntiro, (2000) further breaks down the different domains of e-Government as follows:

#### Improving Government Processes: e-Administration

Ntiro, (2000) states that “e-Government initiatives within this domain deal mainly with improving the internal workings of the public sector. They include:

- **Cutting process costs:** improving the input: output ratio by cutting financial costs and/or time costs.
- **Managing process performance:** planning, monitoring and controlling the performance of process resources (human, financial and other).
- **Making strategic connections in government:** connecting arms, agencies, levels and data stores of government to strengthen capacity to investigate, develop and implement the strategy and policy that guides government processes.
- **Creating empowerment:** transferring power, authority and resources for processes from their existing locus to new locations.”

### Connecting Citizens: e-Citizens and e-Services

The author goes further to elaborate that “such initiatives deal particularly with the relationship between government and citizens: either as voters/stakeholders from whom the public sector should derive its legitimacy, or as customers who consume public services. These initiatives may well incorporate the process improvements identified in e-Administration above. However, they also include a broader remit:

- **Talking to citizens:** providing citizens with details of public sector activities. This mainly relates to certain types of accountability: making public servants more accountable for their decisions and actions.
- **Listening to citizens:** increasing the input of citizens into public sector decisions and actions. This could be flagged as either democratization or participation.
- **Improving public services:** improving the services delivered to members of the public along dimensions such as quality, convenience and cost.”

### Building External Interactions: e-Society

In addition to that Ntiro (2000) explains that “such initiatives deal predominantly with the relationship between public agencies and other institutions - other public agencies, private sector companies, non-profit and community organizations. As with citizen connections, these initiatives may well incorporate process improvements. However, they also include a broader remit:

- **Working better with business:** improving the interaction between government and business. This includes digitizing regulation of, procurement from, and services to, business to improve quality, convenience and cost.
- **Developing communities:** building the social and economic capacities and capital of local communities.
- **Building partnerships:** creating organizational groupings to achieve economic and social objectives. The public sector is almost always one of the partners, though occasionally it acts only as a facilitator for others.”

e-Government in the view of Gordon (2002) is the use of ICT to improve the process of government. In a narrow sense it is sometimes defined as citizens 'services, re-engineering with the technology, or procurement over the internet.

Janet Caldwell (1999) on the other hand, defines e-Government as digital information and online transaction services to citizens.

In light of these definitions we move on to examine the various frameworks in place for e-governance.

## 2.2 Frameworks for E-Government

A growing body of research indicates that various frameworks for evaluating e-Government are in place. A sample review is provided hereunder:

### 2.2.1 U.S. Department of Labour e-Government Strategic Plan

The primary components of the Department's e-Government Framework (the Framework) are customer relationship management, organizational capability, enterprise architecture, and security and privacy. (Solis, 2011)

- **Customer Relationship Management (CRM)**

CRM comprises methodologies, technologies, and capabilities that help the Department identify customers, determine what customers want, and learn how to meet and continuously improve customer service. CRM requires developing a dialogue with customers. Advanced CRM is characterized by personalized services that are timely and consistently excellent. Customer relationship management helps Department of Labor prioritize e-Government projects.

- **Organizational Capability.** This component consists of the policies, plans, people, and management processes required to develop, implement, and sustain a high level of digital services in support of the Department's mission. This category includes strategic plans, investment review boards, IT capital planning processes, systems development methodologies, workforce plans, and training. Organizational capability helps Department of Labor select E-Government projects and ensures successful management of the projects and delivery of results.
- **Enterprise Architecture.** Department of Labour's DOL's enterprise architecture includes the explicit description and documentation of the current and the desired relationships among business and management processes and information technology. The enterprise architecture describes the current architecture and the target architecture. It also includes the rules and standards for optimizing and maintaining IT investments and portfolios. DOL's enterprise architecture helps the Department identify E-Government opportunities.
- **Security and Privacy.** This component of the Framework provides an integrated planning framework and a unified approach to developing and implementing security policies, procedures, and plans, including the analysis of threats and vulnerabilities, risk mitigation, and risk management. Security and privacy policies help create a secure and trusted environment for e-Government transactions.

Figure 2.2 depicts the components of the Department's e-Government Framework. As shown in this figure, the organizational capability, enterprise architecture, and security and privacy components, taken together, represent the Department's organizational readiness to meet customer service requirements. The CRM component is an indicator of the Department's customer awareness. The Department will address these components in an integrated manner. In addition, it will chart a forward course that matches organizational readiness to customer requirements.

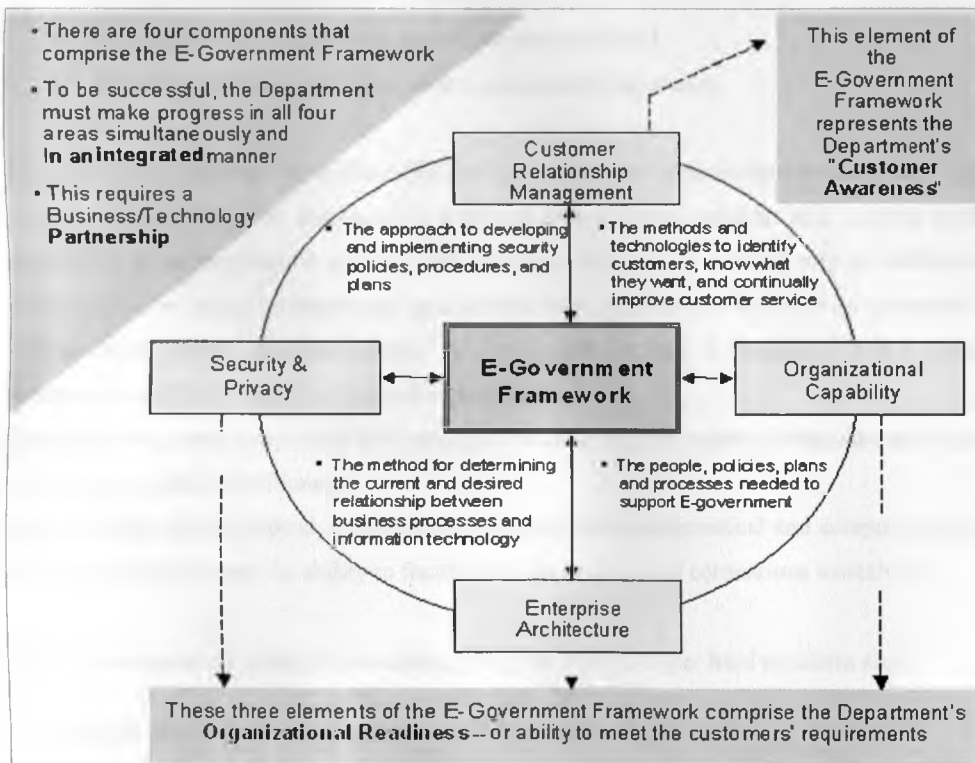


Figure 2.2: Components of U.S Department of Labour's E-Government Framework

### 2.2.2 E-Government Evaluation Framework

Gupta M.P. and Debashish Juma in their paper: E-Government evaluation, have postulated a broad categorization for determining information and servicing value attributable to the several aspects of e-Government benefits.

They go further to say that a range in the classification of methodologies in terms of degree of hardness or softness may be based on the clarity and nature of the influential variables of a problem situation.

Clearly defined problems are structured problems, while poorly articulated or unclear problem situations are categorized as ill structured problems.

They then expound that the methods that match the underlying characteristics of a problem situation are identified and represent an issue that needs to be considered especially in a complex situation.

In line with research done by Mingers and Gill (1997), the typical assumptions made by a hard OR/MS method are that:

- There is a single decision maker (or at least a consensual group) with a clear objective – in a case where there are multiple objectives these are usually reduced to a single metric
- The nature of the problem is agreed upon, even though a good solution may be difficult to find;
- The most important factors can be quantified and reliable data collected;
- A model, often mathematical or computer-based, can be used to generate solutions, and that this does not need to be transparent to the client(s);
- The role of the OR person is one of expert analyst; and
- Future uncertainties can be modeled using probability theory.

In contrast, soft methods can be characterized by generally not making these assumptions. Typically, there might be several decision makers or stakeholders involved, with different opinions and possibly conflicting objectives and definitions of the problematic nature of the situation. In as much as there may be difficulties in quantification of many important factors; transparency and accessibility of the model will be very important, thus often ruling out mathematical models; the OR person's role will often be one of facilitator with a group of participants and uncertainties will not simply be reduced to probabilities.

One important implication of this distinction is that these different types of methods require quite different skills and orientations in their practitioners.

Hard methods would demand a good analytical mind with mathematical and computing skills, while soft methods require people skills and the ability to facilitate often stressful and contentious workshops.

The key measurement criteria for measuring tangible benefits under hard measures are:

Cost Benefit Analysis and benchmarks in E-Government projects.

### **2.2.3 E-Governance framework in the Commonwealth**

Research where an assessment by the Commonwealth Secretariat was done through Governance and Institutional Development Division's Public sector informatics programme reviewed its observations and analyses of ICT case studies gathered in member country workshops and surveys during 2005/06 and came up with an initial e-governance framework which is portrayed in *Figure 2.3*.

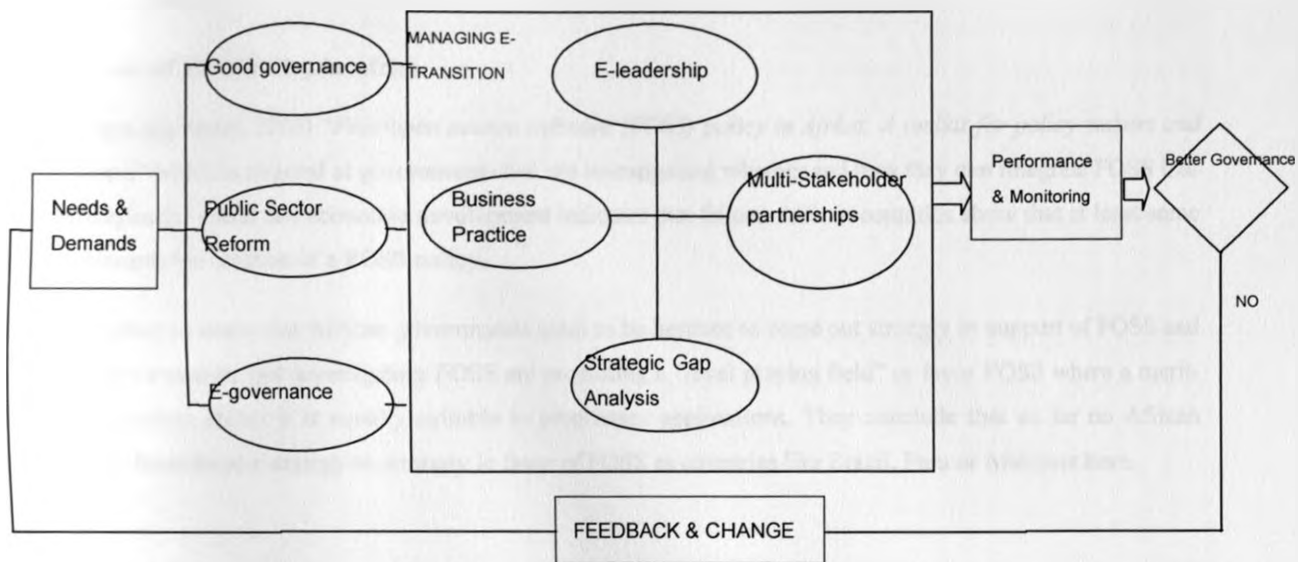


Figure 2.3: e-Governance framework for the Commonwealth

Gessi. *et al.*, (2006), contends that the novelty of this framework stems from its mapping of the confluence of four elements integral to managing e-transitions. First the framework sees e-leadership as the key to making strategic use of ICT's in e-governance initiatives and to assuring local commitment and ownership. Second, it emphasizes good business practice based upon private sector tools for innovation. Third, it focusses on and creates popular pressure for reform through strategic gap analysis. Lastly it features multi-stakeholder partnerships based on mutual trust and interdependence to build capacity.

The framework is about enabling change and redesigning roles and processes to achieve better governance. It responds to good governance principles and practices, public sector reform and ICT innovations. ICTs cross-cut citizens needs for better services and promote improved transparency, accountability and shared decision making. Deploying a set of interrelated planning tools and facilitates strategic responses to intractable problems. The framework also recognizes the need for strong monitoring and evaluation, with a feedback loop for corrective action. Citizens and stakeholders are expected to review governance improvements and to update their changing needs and demands. The combined use of these tools in the public sector increases the chances of successful strategic change management. This framework has been applied in practice as explicated using the four tools which are e-leadership, business practice, strategic gap analysis and multi stakeholder partnerships.

### 2.3 Open Source Software in e-Government

One important research study that demonstrates the need for Open Source Software in e-Government was conducted by the Danish Board of Technology (2002), whereby they cited that the public sector needs to change over to communicating digitally. They discussed that this development makes great demands both on the IT systems on which e-Government is based and on work processes in the public sector. They found out that from the economic point of view, the change-over poses great challenges, as huge investments will have to be made of what forms of Information technology it is anticipated will be used and who controls the ownership of this technology.

They conclude by posing the question: To what extent can open source software supplement or completely replace proprietary software?

## **2.4 Overview of FOSS Policy in Africa**

The (bridges.org report, 2005) "*Free/open source software (FOSS) policy in Africa: A toolkit for policy makers and practitioners*" which is targeted at governments that are investigating whether and how they can integrate FOSS into their strategies for social and economic development indicates that fifteen African countries show that at least some activity towards the creation of a FOSS policy.

They go further to assert that African governments seem to be hesitant to come out strongly in support of FOSS and that most governments that investigating FOSS are promoting a "level playing field" or favor FOSS where a merit-based comparison shows it is equally suitable to proprietary applications. They conclude that so far no African country has formulated a strategy as strongly in favor of FOSS as countries like Brazil, Peru or Malaysia have.

### **Example of FOSS policy: South Africa**

One important research study that demonstrates Open source policy was compiled in the report by Government Information Technology Officers' Council in South Africa in which they found out the following.

The South African Government has set out a policy for open source software (OSS):

"Government will implement OSS where analysis shows it to be the appropriate option. The primary criteria for selecting software solutions will remain the improvement of efficiency, effectiveness and economy of service delivery by Government to its citizens.

OSS offers significant indirect advantages. Where the direct advantages and disadvantages of OSS and PS (Proprietary Software) are equally strong and where circumstances in the specific situation do not render it appropriate, opting for OSS will be preferable." (Using Open Source Software In The South African Government, *A Proposed Strategy Compiled By The Government Information Technology Officers' Council*. 2003)

The following countries have formulated FOSS specific policies or references to FOSS or open standards. Adopted from the (bridges.org report, 2005) "*Free/open source software (FOSS) policy in Africa: A toolkit for policy makers and practitioners*".

#### **Angola**

The e-strategy recommends the use of 'open systems': computer software and hardware that are based on open standards. Adopted from the bridges.org report, "*Free/open source software (FOSS) policy in Africa: A toolkit for policy makers and practitioners*".

## **Benin**

The Government has identified the development of free software as a strategy and it sees the promotion of free software and the "free philosophy" as beneficial to software availability and development in Benin. It encourages civil society organizations to engage in the "battle" for wide scale adoption of FOSS, especially at the international level. A free software laboratory project (LABTIC) is being developed, with the support of the Agence de la Francophonie. (ibid)

## **Djibouti**

The government plans to research and test free software, in particular Linux, with a view to reducing the costs of software procurement. There is a plan to build a software laboratory to do this. There is some word of a plan to put Linux networks into schools, and a "Linux team" has been formed to carry this out. (ibid)

## **Kenya**

A FOSS strategy paper is in progress, but nothing has been published. The government has talked about making sure FOSS is always considered in the procurement of software by Government. There are active FOSS supporters in Kenya and a loose network of Kenyan "hackers" that develop FOSS software. (ibid)

## **Senegal**

Senegal has no FOSS policy but it is mentioned here for two reasons. First, Senegal has a notable amount of activity in ICT for development. It has a vocal Linux and free software society: Le Association Senegalaise pour Linux et les Logiciels Libres. The Senegalese chapter of the Internet society also has an interest in FOSS. Secondly, IT managers in Government are experimenting with FOSS solutions and are promoting them internally. (ibid)

## **South Africa**

In 2001, the South African Government began to openly debate the case for using open standards and open source software in government. The Government Information Technology Officers (GITOC) was subsequently tasked with forming an open source software working group to take this to policy level. The working group was provided with research carried out by the National Advisory Council on Innovation (NACI) in 2002, and in January 2003 presented a strategy paper recommending the use of open standards and open source software in government. To date, no official legislature has been passed by the South African Government endorsing these recommendations, but they have been approved by cabinet and are implemented by individual government departments. (ibid)

## **Tanzania**

Tanzanian policy makers have identified the challenges that face the ICT sector in the country, such as total reliance on imports for ICT equipment, no standards guiding the import of hardware and software, very little local software development and software license costs that are unaffordable to many. The National Information and Communications Technologies Policy, March 2003 lists "Development of local and open source software" as a challenge. (ibid)



## **Uganda**

Uganda has no official FOSS policy to date, but there are several organizations active in the field. Uganda Martyrs University began a complete migration of all software to FOSS around 2002. This initiative is being closely watched by the international development community as a case study of issues encountered during a large scale migration to FOSS. In April 2004, the "East African Centre for Open Source Software" (EACOSS) was opened. This is the first specialized FOSS training centre in the region. The Centre has introduced training, certification and access to FOSS. The Women of Uganda network provides a website describing all the FOSS initiatives active in Uganda, and the business sector is known to use FOSS extensively. The recent National ICT policy focusses on the employment of e-Government and there is a brief mention of Linux and Unix as operating systems to consider as alternatives to Microsoft Windows. However, in August 2004, the US Trade and Development Agency gave the Ugandan Government a grant to facilitate their e-Government strategy, and because this agency advances economic development and U.S. commercial interests in developing countries, there will likely be pressure on policy makers to use Microsoft products. (ibid)

## **Zambia**

Zambia makes some mention of Open Standards in its draft National Information and Communications Policy, 2004, in relation to the problems experienced when there is collaboration between institutions with different technology. (ibid)

The following additional countries show some activity that is relevant in the context of FOSS policy:

## **Burkina Faso**

Burkina Faso has no current FOSS policy, and an IICD-commissioned study showed there is only minimal uptake of FOSS. It found that due to the extremely risk-adverse nature of large businesses in Burkina Faso, there was a reluctance to try FOSS. The government may be considering FOSS due to an intention to develop a local software industry. The reason given is: "To reduce considerably the taxes and rights of customs on the importation of the computers, their elements and the basic software." (ibid)

## **Cameroon**

Cameroon has no published policy, but there are active Linux user groups and the Internet Society of Cameroon supports open source software. (ibid)

## **Ethiopia**

Ethiopia has no published policy, but there is an active Linux user group. There is growing evidence of FOSS use in Ghana, but not of local development of FOSS applications. (ibid)

## **Ghana**

Ghana has no published policy, but there is an active user group. There is growing evidence of FOSS use in Ghana but not of local development of FOSS applications. (ibid)

## **Namibia**

The last ICT policy document to be accepted by the Namibian Government was in 2002 and contained no reference to FOSS, despite there being a notable degree of FOSS activism in Namibia. SchoolNet Namibia (SNN) is a key organization that has led a successful, large scale campaign to put open source computer labs into schools. SNN was part of a working group that put a draft ICT policy for education before government in August 2004. (ibid)

## **Nigeria**

Nigeria has no FOSS policy to date, but a small but flourishing ICT industry and it is building ICT capacity. Many in Africa see Nigeria as a contender with South Africa for outsourcing contracts from overseas has a reasonable infrastructure (at least in urban areas) , and a relatively large number of people with relevant education qualifications. Currently, the Nigerian Government is investigating FOSS as an alternative to proprietary systems in Government departments. According to local sources, the situation is complicated by a generous (and as yet undisclosed) Microsoft offer for software.” (ibid)

### **2.4.1 Driving Forces behind the Adoption of Open Source Software**

There are several studies about how free and open source software is developed, who takes part in development, their motives for developing this kind of software and the reasons for their adoption. More and more governments around the world are requiring their agencies to use free or open source software and use proprietary software only as a last resort.

According to Festa ([www.egovos.org](http://www.egovos.org)), *open source and free software represent a budget priced alternative to Microsoft's Windows operating system and applications that can cost thousands of dollars a year to license. In addition, access to underlying source code means governments can fix problems or modify software to work effectively.*

In supporting Festa, Dan Kusnetzky says that one of the overriding drivers behind legislation appears to be a desire to break free of United States lock on global software market. He asserts that it is not just the United States government that they are worried about, but a single vendor exercising so much power over their government operations.

A government would not like to be under so much influence from any supplier. Governments, especially those of poorer nations with less money to spend on information technology are eager to reap the cost savings of using free software. It is a matter of choice for the governments, organizations etc. to opt for open source software over proprietary software.

### **2.4.2 Countries that have adopted the Concept of Open Source software**

It is generally suggested that knowing the current status and progress of OSS within different countries can be useful in improving OSS adoption and implementation in other countries, especially by learning from those countries that have successfully implemented OSS.

The market share of OSS has increased significantly over the past few years, particularly on the server side. (Brink et al. 2007). Several countries such as Brazil and Germany have migrated most of its local governments and state

agencies to OSS over the last decade. (Red Hat, 2006). According to Lewis, in 2007 there were about 286 OSS license initiatives within the government sector globally. Some governments are recommending OSS, some are mandating the adoption and others are merely doing research and development on OSS.

(Red Hat, 2006) reports that over 160 national, provincial and local governments around the world are utilizing OSS one way or the other. From various studies it is also evident that OSS is extensively implemented within the government sector. Formal academic reports on the current status of OSS usage within various governments are very scarce and in most instances only reports about the intended adoption and implementation of OSS are found.

One of the more interesting aspects of open source software movement is the role that governments are beginning to play. Detailed below are some of the countries that have extensively implemented OSS within their various state departments.

#### 2.4.2.1 Malaysia

The Malaysian government is one of the governments throughout the world that have articulated comprehensive implementation guidelines for OSS and open standards. (Thomas, 2007). An extensive feasibility study was done to provide proper guidelines for deploying Open Document format within the Malaysian government sector. (Red Hat, 2006). The investigation in to the use of OSS in Malaysian public sector began in 2004 where an encompassing Public Sector Policy on OSS (PSPOSS) implementation was adopted. The Policy is divided into eight areas (See Figure 2.4).

<b>ADOPTION</b>	OSS Adoption should be based on the least disruptive and fit for purpose implementation.
<b>PROCUREMENT</b>	OSS procurement should be based on merits, value for money, transparency, security and interoperability.
<b>OWNERSHIP</b>	OSS ownership should include software licensing that allows rights to use and modify the software.
<b>TECHNOLOGY</b>	OSS technology to be used in the public sector shall comply to worldwide open standards
<b>IMPLEMENTATION</b>	OSS implementation should be based on the guidelines specified in the Malaysian Public Sector OSS Technical Implementation Plan
<b>KNOWLEDGE SHARING</b>	Agencies implementing OSS must register their initiatives in the Knowledge Bank
<b>EDUCATION</b>	OSS education should be introduced through structured programmes in school IT labs for primary, secondary and tertiary education levels.
<b>TRAINING</b>	Agencies must be committed in educating and re-skilling its personnel to ensure their competency in OSS

Figure 2.4: Malaysian government OSS Policy

The eight areas are *adoption, procurement, ownership, technology, implementation, knowledge sharing, education and training*. The strategy for implementing OSS in Malaysia is divided into three (3) phases spanning over a period of five years.

Phase I includes the process of laying a foundation such as formulating guidelines and implementing small pilot projects (Thomas, 2007). Phase II focusses on accelerated adoption and Phase III is self-reliance where OSS usage is expected to be significant.

In 2006, the Malaysian government reported on the status of OSS implementation in the public sector (Malaysian Public Sector, 2008). Approximately 61% of IT personnel received training on OSS. In 2008, almost 3,000 government employees had been trained on various OSS products such as OpenOffice.org (The Malaysian Public Sector Open Source Software Master Plan, 2008). OSS is widely used on both the server and the client side in the Malaysian government sector with about 200 state agencies already using OSS (ibid). The OSS applications being used in Malaysia are mainly developed by the OSS community as well as by the Malaysia government open source competency centre (OSCC). Examples include MyWorkSpace (ibid) which was developed to replace MS-Exchange.

By March 2008, an estimated number of about 120 state agencies had migrated desktop users to OSS. The main motivation for adopting OSS in Malaysia is cost savings while lack of technical support is cited as one of the major challenges that affects OSS adoption. (Thomas, 2007).

#### **2.4.2.2 Germany**

According to (Rankin 2006), German government is one of the “visible adopters of OSS”. In 2002, the German federal office moved away from Windows NT to Debian Linux (Nagler, 2005). They further migrated fully from MS Exchange 5.5 to KOLAB, an in-house developed OSS groupware solution (ibid). In 2004, the Munich Municipality migrated 1400 of its Windows Desktop and laptop computers to Linux and OpenOffice.org (Kovacs et al. 2004).

In 2007, the German foreign office converted 10,000 of its desktop machines to OSS across 300 sites (Otter, 2007). What tends to be lacking in many governments that are in the process of adopting and implementing OSS is extensive, diverse and proper implementation guidelines and Germany is one exception. The Federal Ministry in Germany published a comprehensive guide to be used by all government offices when migrating software components on the server and desktop environments.

#### **2.4.2.3 Australia**

##### **2.4.2.3.1 Government Open Source Software Policy Principles**

The Australian government has outlined various principles that guide the procurement and use of open source software in their government and its various agencies. These are briefly highlighted below:

**Principle 1: Australian Government ICT procurement processes must actively and fairly consider all types of available software** (including but not limited to open source software and proprietary software) through their ICT procurement processes. It is recognized there may be areas where open source software is not yet available for consideration. Procurement decisions in such cases have to be made based on ‘value for money’. Procurement decisions should take into account whole-of-life costs, capability, security, scalability, transferability, support and manageability requirements.

**Principle 2: Suppliers must consider all types of available software when dealing with Australian Government agencies** -Australian Government agencies will require suppliers to consider all types of available software (including but not limited to open source software and proprietary software) when responding to the agencies' procurement requests.

**Principle 3: Australian Government agencies will actively participate in open source software communities and contribute back where appropriate** - The Australian Government, through AGIMO (Australian Government Information Management Office), will actively seek to keep up-to-date with international best practice in the open source software arena, through engaging with other countries and organizations. Australian Government agencies should also actively participate in open source software communities and contribute back where appropriate.

These principles go a long way in enhancing the use of Open Source software in Australia.

### 2.4.3 Initiative for Software Choice

To encourage continued software innovation, and promote broad choice, governments are encouraged to consider the following:

- Procure software on its merits not through categorical preferences.
- Promote interoperability through platform neutral standards and maintain a choice of strong intellectual property protection. Stanco (2000) notes that if governments want to create a culture of open source in the country to create an indigenous software industry (a noble goal) they are much better off working on the area of procurement policy.

### 2.5 Technology Adoption and Strategic Planning Frameworks

There are a good number of technology adoption frameworks and strategic analysis tools in place. Johnson and Scholes (1993) proposed a framework for strategic management which has three main elements.

1. Strategic Analysis (environment, culture and stakeholder analysis, and resources and strategic capability) – to understand the strategic situation.
2. Strategic choice (generation of strategic options, evaluation of options and selection of strategy)-to form the strategies.
3. Strategy implementation (planning and allocating resource, organizational structure and design, managing strategic change) – to implement the strategies (tactical)

For the purpose of this study we examine the following:

#### 2.5.1 SWOT Analysis - Strengths, Weaknesses, Opportunities, Threats

SWOT is a simple but powerful framework for assessing internal and external market dynamics. A SWOT analysis must first start with defining a desired end state or objective. A SWOT analysis may be incorporated into the strategic planning model. SWOT is defined as follows:

- **Strengths:** attributes of the person or company that are helpful to achieving the objective.
- **Weaknesses:** attributes of the person or company that are harmful to achieving the objective.
- **Opportunities:** external conditions that are helpful to achieving the objective.
- **Threats:** external conditions which could do damage to the business's performance.



Figure 2.5: SWOT Analysis

### 2.5.2 Cost-Benefit Analysis Framework

Cost Benefit Analysis is typically used by governments to evaluate the desirability of a given intervention. It is used to measure non-monetary as well as monetary costs and benefits to see if the benefits outweigh the costs. The aim is to gauge the efficiency of the intervention relative to the status quo. The costs and benefits of the impacts of an intervention are evaluated in terms of the public's willingness to pay for them (benefits) or willingness to pay to avoid them (costs). Inputs are typically measured in terms of opportunity costs - the value in their best alternative use. The guiding principle is to list all of the parties affected by an intervention, and place a monetary value of the effect it has on their welfare as it would be valued by them. (Gupta and Jana, 2003).

The practice of cost-benefit analysis differs between countries and between sectors (e.g. transport, health) within countries. Some of the main differences include the types of impacts that are included as costs and benefits within appraisals, the extent to which impacts are expressed in monetary terms and differences in discount rate between countries. Agencies across the world rely on a basic set of key cost-benefit indicators, including:

- PVB (present value of benefits);
- PVC (present value of costs);
- NPV (PVB less PVC);
- NPV/k (where k is the level of funds available) and
- BCR (benefit cost ratio, PVB divided by PVC).

The accuracy of the outcome of a cost-benefit analysis is dependent on how accurately costs and benefits have been estimated. Strategies adapted from (Anything Research, 2012).

### 2.5.3 Technology Adoption Curve Framework

The technology adoption curve framework is based on the notion that individuals will adopt an innovation if they perceive that it has the following attributes. First, the innovation must have some relative advantage over an existing innovation or the status quo. Second, the innovation must be compatible with the existing values, past experience, and practices of the potential adopter. Third, the innovation cannot be too complex nor perceived as difficult to

understand. Fourth, the innovation must have trialability; that is, it can be tested for a limited time without adoption. Fifth, the innovation must offer observable results (Rogers, 1995).

(Rogers, 1995) asserts that an adopter's experience with one innovation influences that individual's perception of the next innovation in a technology cluster to diffuse through the individual's system. Thus, if an adopter has a negative first experience with one computer application, he or she may regard all computer applications through this perspective. Diffusion theory provides a framework that helps to understand why IT is adopted by some individuals and not by others. This theory can explain, predict, and account for factors that increase or impede the diffusion of innovations.

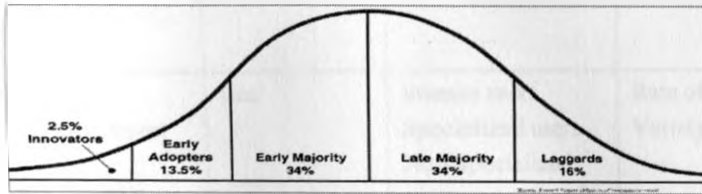


Figure 2.6: Technology Adoption Curve Framework

- **Innovators** tend to be more educated and prosperous, with a greater tolerance for risk
- **Early adopters** are younger, educated, and active in the community
- **Early majority** are more conservative, but open to new ideas and influential within the community
- **Late majority** may be older, less educated, conservative, and less socially active
- **Laggards** are highly conservative, oldest and least educated. They often are less prosperous and more risk averse.

The technology adoption curve is generalizable to any new product or market.

Getao (2004) in citing Sherry identified Rogers (1995) model where an innovation was considered an object with five relative attributes: relative advantage, compatibility, complexity, trialability and observability. The decision by the user to accept or reject the innovation is an event, a point in a linear process where time is the independent variable. The adoption process is made up of a series of choices and action over time based on internal factors within a social system. Getao (2004) contrasted technology adoption models by citing Shih (2004), who pointed out that adoption models concentrate on the diffusion of the technology to different categories of users with use models that concentrate on the different types of use to which the technology is put. *Table 2.1* exemplifies the contrast.



Model	Typology of population	Variable of Interest	Relevant criteria	Element unique to each model	Elements common to both models
Adoption	Adoption	Innovators Early adopters Early majority Late majority Laggards	Timing or rate of adoption	Observability Compatibility Triability	Innovativeness Social Communication Complexity Influence of media Relative Advantage
Use - Diffusion	Use	Intense users Specialized users Non-specialized users Limited users	Rate of use Variety of use	Product experience Competition for use Sophistication of technology Satisfaction	

Table 2.1: Contrast between adoption models and Use diffusion models

## 2.6 Frameworks and Open Source Tools

Frameworks are the base building blocks for most of the current generation applications. This is because of the fact that they help streamline application development, promote adoption of best practices, promote re-use and reduce total cost of ownership by reducing the amount of effort and time. Frameworks are pervasive in most domains of application development and this encompasses object assembly, runtime object management, transaction management, messaging infrastructure, data access, parallel processing, user interaction, service oriented orchestration, event processing, networking and legacy integration. J2EE and .Net which both relate to object oriented technologies are among the most well known and most comprehensive frameworks in the present time.

Nevertheless there are numerous other frameworks, especially those that provide the middleware infrastructure for applications built using Java, C# or the scripting languages like PHP, Python, Perl and Ruby. Java itself has several frameworks like the Apache Struts which implements the MVC pattern and effective user interaction or the Spring framework, that provides an alternative to J2EE and which leverages the dependency injection system. AJAX, Asynchronous JavaScript and XML is also a user interaction methodology or framework.

Many of these frameworks are open source initiatives. After the success of Linux and BSD, the best known open source operating systems, the next thrust to open source adoption has been in the area of middleware infrastructure. Frameworks, application servers, middleware components and shared libraries fall in this place. The debate on application level open source initiatives is still open, whereas the middleware, database and operating system open source options have become real and viable choices. (Rose India Technologies Pvt. Ltd, n.d.)

Some of the Open source frameworks available are:

### Open-source PHP5 web framework

Symfony aims to speed up the creation and maintenance of web applications, and to replace the repetitive coding tasks by power, control and pleasure. Symfony provides these features seamlessly integrated together, such as:

- simple templating and helpers
- cache management
- smart URLs
- scaffolding
- multilingualism and I18N support
- object model and MVC separation
- Ajax support
- enterprise ready

### **Apache Struts**

The Apache Struts web framework is a free open-source solution for creating Java web applications.

Web applications differ from conventional websites in that web applications can create a dynamic response. Many websites deliver only static pages. A web application can interact with databases and business logic engines to customize a response.

Web applications based on Java Server Pages sometimes commingle database code, page design code, and control flow code. In practice, we find that unless these concerns are separated, larger applications become difficult to maintain.

One way to separate concerns in a software application is to use Model-View-Controller (MVC) architecture. The model represents the business or database code, the view represents the page design code, and the controller represents the navigational code. The Struts framework is designed to help developers create web applications that utilize MVC architecture. (Rose India Technologies Pvt. Ltd, n.d.)

The framework provides three key components:

- A "request" handler provided by the application developer that is mapped to a standard URI.
- A "response" handler that transfers control to another resource which completes the response.
- A tag library that helps developers create interactive form-based applications with server pages.

The framework's architecture and tags are catchphrase compliant. Struts works well with conventional REST applications and with nouveau technologies like SOAP and AJAX.

## **Zend Framework Services**

This framework is employed for building performance-oriented, highly secure and modern Web 2.0 Application Development. It is based on OOP, UI design patterns, unit testing, loose coupling, corporate friendly licensing schemes, MVC implementation, and most importantly – a meticulously-tested agile code base. (ibid)

### **Advantages of using Zend Framework for PHP Web Development**

- Faster web development Makes Hybrid Website Development simpler with Rich inbuilt library
- Uncomplicated interfaces and abstract classes
- OOP-based, thus involves lesser coding and rapid development
- Low-cost maintenance
- Scalable over a period of time

### **Code Igniter**

Code Igniter is a powerful PHP framework with a very small footprint, built for PHP coders who need a simple and elegant toolkit to create full-featured web applications. (ibid)

#### **Advantages of Code Igniter**

- Has a small footprint
- Exceptional performance
- Broad compatibility with standard hosting accounts that run a variety of PHP versions and configurations
- Requires nearly zero configuration
- Does not require use of the command line.

## **2.7 Framework to be adopted for the Research**

It is summarized from the above literature that there are several frameworks that are in place to support e-Government as well as open source software applications initiatives. Technology adoption frameworks are also in place to facilitate and give direction to interventions that governments may wish to pursue. The Kenya government can seek the way of stakeholders input, top management support, local skill base in order to enhance the uptake of various open source technologies and encourage the sharing of content.

The Kenyan government will be interrelated with the U.S. Department of Labour e-Government framework which is the e-Government framework of choice for this study because its components map well with the ones intended to be researched on in the Kenyan government to form the research framework that will be adopted.

The User Input will consist of input from all the five categories identified. The users in this context are citizens, as well as officials in the Kenyan government. On creative commons licensed open content, factors such as who owns the data, how useful and relevant digital content can be used to supplement the data that the Kenyan government already has, the factors used in determining relevant content and so on will be researched on.

## **CHAPTER THREE: METHODOLOGY**

### **3.1 Overview**

The aim of this chapter is to present and justify the research and prototype development methods used in this research.

In carrying out the research, a systematic approach was followed.

Presentation and description of how the data was collected is outlined. The presentation of the chosen methodology includes discussions concerning the research approach, research strategy and research methods.

### **3.2 Research Purpose and Design**

All research approaches can be classified into one of three general categories of research:

Exploratory, descriptive and casual. These categories differ significantly in terms of research purpose, research questions, the precision of the hypothesis that are formed and the data collection methods that are used. (Aaker et al, 1998).

The methodology employed was exploratory Research as this research is used when one is seeking insights into the general nature of a problem, the possible decision alternatives and the relevant variables that need to be considered. The research methods in this category are highly flexible, unstructured and qualitative, as the researcher begins without firm preconception as to what will be found. The absence of structure permits a thorough pursuit of interesting ideas and clues about the problem situations. (ibid)

### **3.3 Research Strategy**

The first step to take when conducting research is to evaluate the research strategies.

Depending on the type of research, there are advantages and disadvantages to all the research strategies. The most important criterion for deciding what strategy to use is to look at the research questions/ objectives Davey (1991), Yin (1981).

In the literature review section, several frameworks were reviewed for e-Government, open source as well as technology adoption. In reference to these frameworks, for e-Government, the U.S Department of Labor e-government framework was adopted. This was mainly due to the fact that most of the elements could be married to the Kenyan context and form a good basis to evaluate the Kenyan scenario.

In addition to the above mentioned frameworks, the Code Igniter PHP development framework was the framework of choice as it has a small footprint, exceptional performance and has a broad compatibility with standard hosting accounts that run a variety of PHP versions and configurations.

### **3.4 Sample Selection**

Choosing a study sample is an important step in any research project since it is rarely practical, efficient or ethical to study whole populations. Martin N Marshall, (1996).

The major emphasis in this study is on the discovery of ideas and insights into the factors that influence the open source software technology adoption in e-Government in Kenya and aims to come up with ways to facilitate the adoption.

#### **3.4.1 Sampling Issues**

For the exercise to be effective and hence produce as meaningful results as possible, it was important that the various stakeholders be classified into different categories.

This put into perspective the various attributes that needed to be investigated. Each of the five categories was given an overview of the aim of the research and adequate understanding developed through discussions and feedback from the participating teams.

#### **3.4.2 Filtration of sampling**

For the purpose of this study, it was observed that the semi-illiterate and illiterate population would be inappropriate as most of them might not have been exposed to open source software nor CC licensed open content. Even though they were part of the customers of the Kenyan government and possible beneficiaries of the outcome of the findings, the research was more focused on Kenyans in the urban areas who were more literate and had some basic understanding of OSS and CC licensed open content. Members of the rural population as well as the semi-illiterate can be included as part of future research work.

#### **3.4.3 Sampling Design**

Due to the prevailing budgetary constraints the sampling procedure that was adopted was random sampling and the type of random sample that was drawn was a stratified sample as the parameters of interest in our research context were the literate populace who had some basic knowledge on open source software as well as content. As the population was homogenous sample bias was overcome by taking a stratified sample so that the stratified population structure was reflected in the sample structure and the criterion that was used was literacy levels.

The informants that were selected were those that had IT literacy and some knowledge about open source applications in addition to various licensing models. The focused population contained both students as well as professionals. The age group was from 21-40 because these age groups people are more interested about new services and have strong opinions about innovations and technology. They also constitute a large part of those involved in the operations of the e-Government and have relevant experience in open source technologies. Our problem is related to the government and its use of various classes of software, so it was necessary to ask them what they expected from software, what they currently had and what they would wish to get out of the software that they had invested in currently and would wish to in the future as well as their expectations as far as CC licensed open content was concerned. The open source software developers were also interviewed on more or less the same issues. This is why the data that was collected was from a sample population as they represented the requirements of their same target group. This helped to abstract the perception and awareness of the required open source applications and creative commons licensed open content from the relevant parties.

### 3.4.4 Sample Size and Sample Selection

The allocation of the sample was made on the basis of the size and population of the various target groups that were relevant to this research.

Different population sizes were used for the various target groups. Using a confidence level of 95% and confidence intervals of 23.77, 17.32, 4.62, 8.84, 9.8 respectively, the sample size for the various groups for ICT Heads, ICT Staff, Citizens, Open Source software developers and Content Creators was estimated at 17, 32, 450, 123 and 100 in that order. This was done by use of an online sample size calculator for statistics available at <http://www.surveysystem.com/sscalc.htm>, accessed on 1<sup>st</sup> April 2012.

#### 3.4.4.1 ICT Heads Sample

**Determine Sample Size**  
Confidence Level:  95%  99%  
Confidence Interval:   
Population:   
   
Sample size needed:

**Find Confidence Interval**  
Confidence Level:  95%  99%  
Sample Size:   
Population:   
Percentage:   
   
Confidence Interval:

Figure 3.1 ICT Heads Sample

### 3.4.4.2 ICT Staff Sample

**Determine Sample Size**

Confidence Level:  95%  99%

Confidence Interval:

Population:

Sample size needed:

**Find Confidence Interval**

Confidence Level:  95%  99%

Sample Size:

Population:

Percentage:

Confidence Interval:

Figure 3.2 ICT Staff Sample

### 3.4.4.3 Citizens Sample

**Determine Sample Size**

Confidence Level:  95%  99%

Confidence Interval:

Population:

Sample size needed:

**Find Confidence Interval**

Confidence Level:  95%  99%

Sample Size:

Population:

Percentage:

Confidence Interval:

Figure 3.3 Citizens Sample

### 3.4.4.4 OSS Software Developers Sample

**Determine Sample Size**

Confidence Level:  95%  99%

Confidence Interval:

Population:

Sample size needed:

**Find Confidence Interval**

Confidence Level:  95%  99%

Sample Size:

Population:

Percentage:

Confidence Interval:

Figure 3.4 OSS Software Developers Sample



### 3.4.4.5 Content Creators Sample

**Determine Sample Size**

Confidence Level:  95%  99%

Confidence Interval:

Population:

Sample size needed:

**Find Confidence Interval**

Confidence Level:  95%  99%

Sample Size:

Population:

Percentage:

Confidence Interval:

Figure 3.5 OSS Content Creators Sample

Population	Target	Respondents
ICT Heads	17	12
ICT Staff	32	28
Citizens	450	312
Open Source Software Developers	123	100
Content Creators	100	76

Table 3.1: Sample selection

### 3.5 Framework

The framework that was adopted to guide this study (the Framework) had these components: Customer relationship management, organizational capability, enterprise architecture, and security and privacy. (Solis, 2011)

- **Customer Relationship Management (CRM).**

This comprised the methodologies, technologies, and capabilities that would help the Government of Kenya identify its customers (citizens), determine what customers want, and learn how to meet and continuously improve customer service. CRM required developing a dialogue with customers. Advanced CRM was characterized by personalized services that are timely and consistently excellent. Customer relationship management would help the Government of Kenya prioritize projects. In our case, projects related to Open source software and CC licensed open content.

- **Organizational Capability.** This component consisted of the policies, plans, people, and management processes which were required to develop, implement, and sustain a high level of open source software digital services and generation of CC licensed content in support of the Government's mission. This category included strategic plans, investment review boards, I.T. capital planning processes, systems development methodologies, workforce plans, and training. Organizational capability will help the Government of Kenya select Open source projects and ensure successful management of the projects and delivery of results.
- **Enterprise Architecture.** This included the explicit description and documentation of the current and the desired relationships among business and management processes and information technology. The enterprise architecture described the current architecture and the target architecture. It also included the rules and standards for optimizing and maintaining IT investments and portfolios. The GoK's enterprise architecture helped the Department identify E- Government opportunities.
- **Security and Privacy.** This component of the Framework provided an integrated planning framework and a unified approach to developing and implementing security policies, procedures, and plans, including the analysis of threats and vulnerabilities, risk mitigation, and risk management. Security and privacy policies help create a secure and trusted environment for e-Government transactions.

### 3.6 Requirement Determination

The use of exploratory research was employed as from the literature review and research questions it was more relevant to use this. This entailed a fact finding exercise in which information/ facts about the various content needs of the Kenyan Government as well as software needs was discovered. The major emphasis in exploratory studies is on the discovery of ideas and insights. As such the research design appropriate for this study had to be flexible enough to provide opportunity for considering different aspects of the problem under study. Inbuilt flexibility in research design was needed because the research problem, broadly defined initially, was transformed into one with

more precise meaning in exploratory studies, which fact necessitated changes in the research procedure for gathering relevant data.

### **3.7 Data Collection**

#### **3.7.1 Field Data Collection**

Data collection was done immediately after the data collection tools had been prepared. The collection of data lasted for a period of 2 months. The supervisor oversaw the exercise and ensured the correct questions were asked to elicit the most relevant responses in the survey.

#### **Tools, Procedures and Methods for Data Collection**

The choice of data collection method is a critical point in the research process. The decision was not easy and many factors were considered and generally, the following three methods in the context of research design for this study were explored for establishing the information requirements of the prototype to be developed.

- Interviews
- Questionnaires
- Documentary Review

#### **3.7.2 Interviews**

Experience Survey means the survey of people who have had practical experience with the problem to be studied. The object of such a survey is to obtain insight into the relationships between variables and new ideas relating to the research problem. For such a survey people who are competent and can contribute new ideas may be carefully selected as respondents to ensure a representation of different types of experience. (Kothari, C.R., 2004).

This entailed a direct conversation with the sampled respondents with a specific purpose of obtaining information regarding the open source skills, proficiency levels, the availability of creative commons licensed open content etc.

An interview schedule was prepared for the systematic questioning of the informants. The interview was conducted in such a way as to ensure flexibility in the sense that the respondents were allowed to raise issues and questions which I might not have previously considered.

Different people in the various selected categories were surveyed about their experiences and expectations on open source technologies and creative commons licensed open content. The citizens who had basic IT knowledge and insight concerning the research question were surveyed rather than any sample group of citizens. This research type was more helpful in acquiring the results due to its flexible nature.

#### **3.7.3 Questionnaires**

In order to investigate the workability and applicability of Open source software and CC licensed open content, questionnaires were mainly used from the sampled population. Samples of the questionnaires are attached in the *Appendix*.

The questionnaires were initially piloted to test for ambiguity and ease of response using a few respondents. The validity of the questionnaire was tested by subjecting it to a few respondents. Content validity and relevance was analyzed through peer review and also supervision from my Academic supervisor and the e-Government liaison.

This involved the use of standardized, structured and unstructured questions that were designed to be used to supplement the interviews. Questionnaires were used because of their ability to reach geographically scattered correspondents conveniently and at a lower cost especially with the advent of emails.

The Primary data source was the feedback from the open source software developers on their familiarity of open source software and applications and their expectations in terms of working with the Government of Kenya to build applications and also content creators and the feasibility of sharing their content under creative commons licensed open content. Also sourced were the goings on at the Government with regards to the extent of usage of Open source software applications and systems and the extent to which the current IT strategy and various policies catered for the issue of usage of Open source software and CC licensed Open content in the Kenyan government.

In this case, open source software applications in e-Government were discussed from different aspects like what were the software needs for the government that could be serviced using open source technology. How these services could be made beneficial, effective or efficient by adopting an open source approach.

The questionnaires targeted the five main categories of respondents who helped disseminate useful information on various aspects of interest.

**E-Government decision/ policy makers** – This was to find out about the organizational capability of the government in terms of the policies, plans strategies and management processes required to mandate the development of Open source software applications as well as actualize the content sharing for content licensed under the creative commons licensed open content.

**E-Government I.T Staff** – This was to get feedback on the enterprise architecture of the Government in terms of the desired relationships among business management processes and information technology. They will also shed light on the rules and standards for optimizing and maintaining IT investments which in this case are open source software applications and portfolios and content sharing enablers.

**Citizens** – This was for customer relationship Management to enable us find out what the citizens who were the customers in our case needed and learned how to meet and continuously improve customer service in relation to CC licensed open content delivery.

**Open source software developers** – As this was the skill base that was going to be utilized going forward various aspects such as their skillset in Open source software, their perception of the various features and functionalities and robustness of various Open source software among others were explored.

**Creators of Content** – This was to find out from the literate population their willingness to share any useful content they may have created over the years or had access to under various CC licenses and which way their participation could be elicited in the most optimal way.

Utmost care was taken in order to present the collected expectations in their original way. This would provide strong basis for the Kenya government to find solutions and strategies for facilitating the uptake of open source applications as well as creative commons licensed open content in the light of collected data.

#### **3.7.4 Documentary Review**

The survey of relevant literature - Previous work will be thoroughly reviewed. Research questions stated by earlier workers may be reviewed and their usefulness be evaluated as a basis for further research. (Kothari, 2004).

This involved the inspection of existing literature on the OSS adoption concept. It assisted in providing facts on the governments in the world that have implemented OSS applications in their e-governance, their experiences and a critique of the OSS concept.

The documents that were reviewed included professional I.T journals, conference proceedings, newspapers, dissertations and other scholarly research literature. Documentary review formed the core fact finding technique as most of the information regarding the OSS concept, development and its implementation was found in secondary literature.

#### **3.8 Data processing**

The uniqueness of the survey required appropriate arrangements to be put in place so that it would be possible to make available the results within the shortest time possible once the data collection was complete.

Using Google Forms, an online survey tool proved requisite as it resulted in faster completion of the survey report. Also it would enable detection of any problems early during the data collection. The Data editing, processing and analysis took six days. Data was processed using the SPSS tool. Descriptive statistics was used mainly for analysis of the data. Frequency tables and charts were used for the presentation of the results.

#### **3.9 Data Cleaning and Validation**

The cleaning and validation processes were done during the data entry process. While data cleaning was a continuous exercise even during report writing, efforts were made to identify any invalid values within the data so that they would be sorted out early enough.

#### **3.10 Constraints**

When designing this survey, two major constraints were encountered. The first and most important was the financial resources available to undertake the survey. This constraint limited how many people could be surveyed and how much time the interviewer could spend with the respondents.

The second constraint was the willingness and ability of respondents to provide desired information. Majority of the participants were not very willing to provide the interviewer with the desired information and the few who were willing to give information couldn't disclose too much in as much as anonymity was guaranteed.

### **3.11 Validity and Reliability**

In this research, validity and reliability was achieved by focusing on key stakeholders in the government that were in charge of various matters related to various technologies among them open source and use of the creative commons licensed open content and various relevant categories of respondents. The validity was ensured throughout research by using relevant literature and the questionnaire was formulated to be as unambiguous as possible and to collect the expected information. Although in qualitative and partially quantitative research approaches, it is hard to maintain the reliability, but utmost care was taken in order to try to attain it by managing the contents, sequence and physical layout of questionnaires.

### **3.12 Prototype Development Tools**

PHP 5.3, MySQL 5 and Apache Web server were used.

### **3.13 Application Development Methodology**

The Waterfall model methodology was adopted as the prototype development methodology of choice as it was straightforward and let one know exactly what stage they were in the process. The steps that were followed in the development of the prototypes sequentially were outlined in these phases:

- Requirements Analysis Phase
- Design Phase
- Implementation Phase
- Integration and Test Phase
- Maintenance Phase

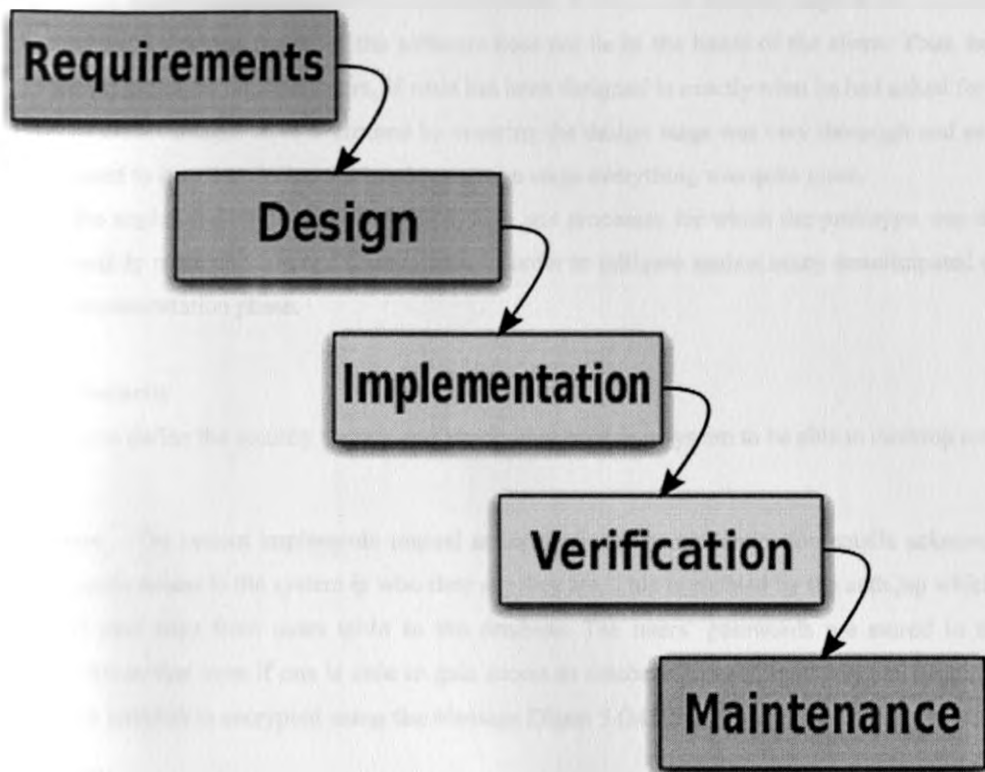


Figure 3.6: Waterfall Methodology

**Requirements Analysis Phase** – Here the various requirements were gathered, for example the actual process flow of the content sharing identified for a prototype to be developed.

**Design Phase** – E-R diagrams were drawn in order to define the various entities and their attributes as well as their relationships.

**Implementation Phase** – A prototype in the form of a platform for content sharing was developed for content sharing purposes using Code Igniter which is a PHP open source software framework.

**Testing Phase** – The tests shall be done by having various users upload content to the portal and check whether all the components function as required.

**Maintenance Phase** – This is where bugs that will have been identified will be fixed

### 3.13.1 Limitations of Methodology and how they are overcome

- You cannot go back a step; if the design phase has gone wrong, things can get very complicated in the implementation phase.
- Often, the client is not very clear of what he exactly wants from the software. Any changes that he mentions in between may cause a lot of confusion.
- Small changes or errors that arise in the completed software may cause a lot of problems.

- Another major disadvantage of the waterfall model is that, until the final stage of the development cycle is complete, a working model of the software does not lie in the hands of the client. Thus, he is hardly in a position to inform the developers, if what has been designed is exactly what he had asked for. These disadvantages were overcome by ensuring the design stage was very thorough and enough time was allocated to it so that during the implementation stage everything was quite clear. All the angles and requirements of the system and processes for which the prototype was developed were thoroughly researched on and documented in order to mitigate against many unanticipated changes during the implementation phase.

### 3.14 Type of Security

It is imperative to define the security threats and attacks that exist in a system to be able to develop mechanisms to avert them.

**Authentication** – The system implements mutual authentication. The authentication entails acknowledgement that the user who gains access to the system is who they say they are. This is enabled by the auth.jsp which compares the username and pass keys from users table in the database. The users' passwords are stored in the database in encrypted format so that even if one is able to gain access to database content, they will not be able to access user passwords. The pass key is encrypted using the Message Digest 5 (MD5) one way hash function.

#### Access Control

The content sharing system has user levels:

Registered user- Anyone can apply for membership as a registered user. A registered user can upload and edit his/her own content (videos/files/audio/applications among others).

All user transactions are logged by the system in such a way that all processes can be easily tracked.

The system performs automated daily backups in order to allow for quick recovery in the event of failure.

### 3.15 Implementation of the Content Portal using Open Source Software Frameworks

Code Igniter (which is an Open Source software) framework for PHP was used to develop the Creative commons licensed Open content portal.

After development, the next stage will be beta testing with a few users in order to evaluate various aspects of the system and also in order to evaluate if we have achieved our research objectives. Before then, a clear test plan will be developed, outline the test objectives, the test items/features, the pass and fail criteria for the tests, the expected and actual outcome etc. The outcome of the tests will be used to review the system and trigger enhancements consistent with the research objectives. Thus the development process will follow the waterfall software development methodology.



The flexibility of the proposed solution can however be reproducible in similar settings with a high degree of success for any other exercise that aims to enhance adoption rates of Creative commons licensed open content using the above defined methodology.

## CHAPTER FOUR: DATA ANALYSIS AND INTERPRETATIONS

### 4.1 Introduction

The purpose of collecting data for this project was to carry out a viability and applicability assesmnet of enhancing the adoption of Open source applications and CC licensed open content . This chapter presents the research findings and the researchers' interpretation from data collected from the respondents.

In addition, data collected through a closed online survey is descriptively and statistcially analysed. Main survey questions, including results received are also discussed and interpreted. It is important to note that respondents, ICT heads, top level management are used interchangeably to refer to the participants who completed the survey. Dynamic online questionnaires were used to provide an adaptive set of questions to the respondents. Questionnaires were administered to five sets of people so as to get various responses on vital aspects that were of concern in this research. The questionnaires consisted mainly of closed type questions, where respondents were compelled to choose between alternatives. Most questions used in the questionnaires had presefined alternatives (answers) with an option to type in other unavailable selections.

Detailed below is the breakdown of the Questionnaire distribution.

- i) **E-Government decision/ policy makers** – This was so as to find out about the organizational capability of the government in terms of the policies, plans strategies and management processes required to mandate the development of Open source software applications as well as actualize the content sharing for content licensed under the creative commons licensed open content.
- ii) **E-Government I.T Staff** – This was to get feedback on the enterprise architecture of the Government in terms of the desired relationships among business management processes and information technology. They also aimed at finding out the literacy levels in as far as Open source applications and software were concerned of the ICT Staff of the government so as to gauge whether they could successfully maintain and support Open Source software.
- iii) **Citizens/ Content Creators** – This was for Customer relationship Management purposes. It was aimed at the literate population and was to find out their willingness to share any useful content they may have come across or were in possession of under various CC licenses and which way they felt would be most beneficial to them to share this content.
- iv) **Open source software developers** – This was aimed at discovering the skill base of Open source software developers that would be utilized going forward as well as various aspects such as their skillset in Open source software, their perception of the various features and functionalities and robustness of various Open source software among others.

- v) **Creators of Content** – This is to find out from the literate population their willingness to share any useful content they may have created over the years under various CC licenses and which way their participation can be elicited in the most optimal way.

Out of 17 ICT heads in the various ministries in the Government of Kenya who were given the questionnaire for decision/ policy makers, only 12 responded.

Out of the 32 Government ICT Staff who the questionnaire was distributed to, 28 responded.

Out of the 450 citizens to whom the questionnaire for citizens was administered to, only 312 responded.

Out of the 312 who responded, 64 questionnaires had errors hence could not be used for statistical analysis.

This brought down the sample size to 248.

100 open source software developers responded to the questionnaires from total of 123 that were sent out.

For the content creators, 100 questionnaires were sent out and 73 responded.

## 4.2 Data Processing and Analysis

Data processing involved editing and tabulation of the collected raw data while analysis involved evaluation of some parameters from the data in order to get patterns or relationship among data items.

### 4.2.1 Coding the responses

In order to analyze the data using SPSS statistical software, as most of the questions were open-ended, similar ideas were identified, and grouped for ease of analyzing.

### 4.2.2 Reliability and Validity Testing

Before the data analysis was done, reliability and validity tests were carried out on the data collection instruments.

In this case, the techniques used were:

- Face Validity through peer review and experts judgment
- Content Validity using Factor Analysis

#### 4.2.2.1 Reliability Test

Reliability is the consistency of measurement, or the degree to which an instrument measures the same way each time it is used under the same condition with the same subjects. There are two ways that reliability is usually estimated namely *test/retest* and *internal consistency*. The idea behind *test/retest* is that you should get the same on several tests. On the other hand, *internal consistency* estimates reliability by grouping questions in a questionnaire that measure the *same concept*.

#### **4.2.2.2 Validity Test**

Validity refers to the best available approximation to the truth or falsity of a given inference, proposition or conclusion. Three commonly used validity testing techniques are construct, content and face validity.

*Construct Validity* refers to the totality of evidence about whether a particular operationalization of a construct adequately represents what is intended by theoretical account of the construct being measured. Such lines of evidence include statistical analyses of the internal structure of the test including the relationships between responses to different test items.

*Content validity* – is a non-statistical type of validity that involves the systematic examination of the test content to determine whether it covers a representative sample of the behavior domain. Such validity testing is done by a panel of experts who review the specifications of selected items. Through their recommendation, the content validity of a test can be improved.

*Face Validity* is also a non-statistical validation method used to get opinions on whether an instrument “looks like” it is going to measure what it is supposed to measure. While content validity requires more rigorous analysis by subject experts, face validity only requires an intuitive judgment.

#### **4.2.3 Reliability Analysis of the collected data**

##### **4.2.4 Face Validation**

In order to investigate the face validity of the research instruments, the questionnaire was given out to technical and non-technical people to check on whether the questions were clear and in line with the research questions. Changes were made before the questionnaires were administered as recommended by the reviewers.

#### **4.3 Detailed Analysis of Data Collected**

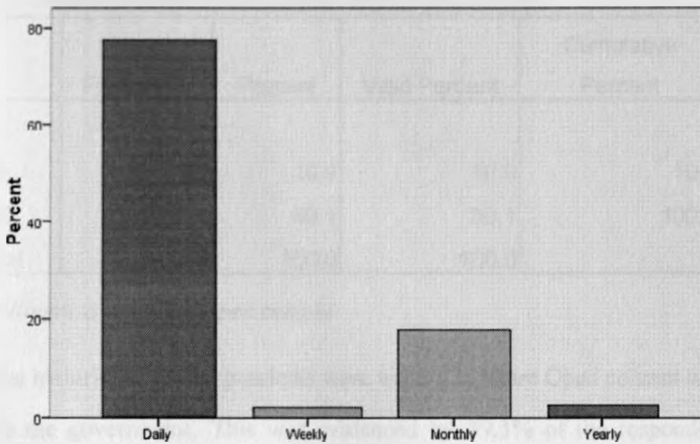
In this section, a detailed analysis and discussion of the valid data obtained from the preliminary investigations using 12, 28, 100, 248 and 73 for e-government top level management, e-government ICT staff, open source software developers, citizens and content creators respectively is presented.

##### **4.3.1 Availability of CC licensed open content among the citizens**

Siebel (2005), in his study indicates that most of the governments and government agencies offer e-government services through the Internet. Therefore the Internet was established to be a good medium for the government to use to reach its citizens and get CC licensed open content from them via a Content sharing portal made available on the Internet.

Figure 4.1 clearly depicts that on average 77.7% of our citizen sample access the Internet on a daily basis, 2% on a weekly basis and only 17.8% and 2.4% access the Internet on a Monthly or Yearly basis.

1. How often do you use the Internet?



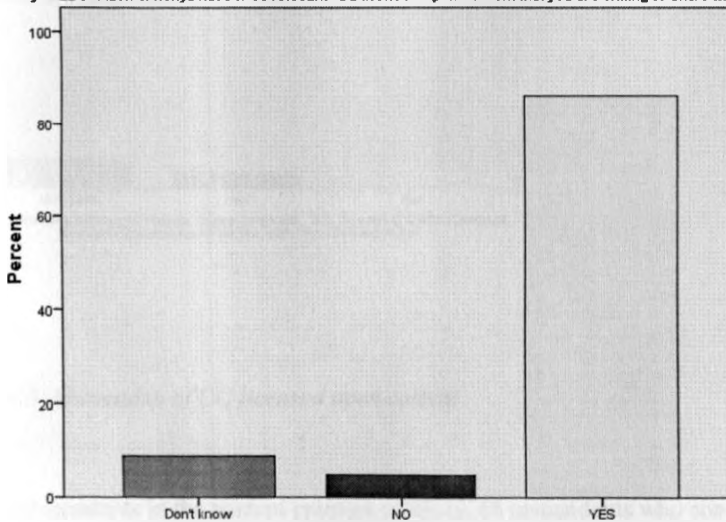
1. How often do you use the Internet?

Figure 4.1 Frequency of Internet Usage

Figure 4.1 above shows that accessibility to an online content sharing portal is feasible and accessible to a large number of citizens as regularly as daily.

Figure 4.2 below shows that a good percentage (83%) of CC licensed open content is available amongst the citizens of Kenya.

5. Do you as a citizen of Kenya have relevant CC licensed Open Content that you are willing to share ...



5. Do you as a citizen of Kenya have relevant CC licensed Open Content that you are willing to share with the Government

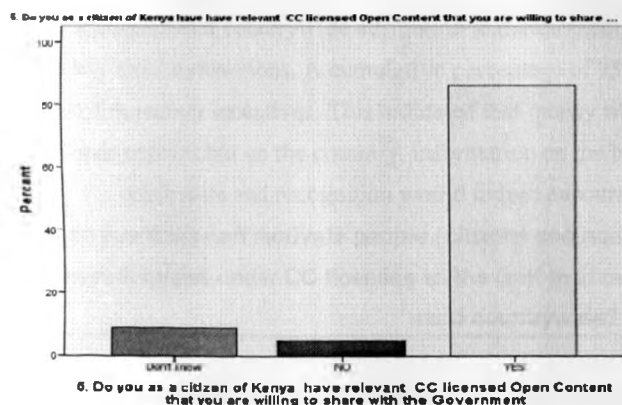
Figure 4.2: Availability of CC licensed Open content among the citizens of Kenya

**If you were requested to contribute any open digital content that is  
CC licensed you may be willing to share to the Government to enhance  
service delivery in the GoK would you do it?**

	Frequency	Percent	Valid Percent	Cumulative Percent
NO	27	10.9	10.9	10.9
YES	221	89.1	89.1	100.0
Total	248	100.0	100.0	

*Table 4.1: Willingness to share Open content*

The study established that majority of the respondents were willing to share Open content which they may be having or are able to access to the government. This was evidenced by 89.1% of the respondents who indicated their willingness to share content with the government. However, 10.9% indicated that they were not willing to share the content they may be having with the Government of Kenya. See *Table 4.1* above. This was an indication that the merits of sharing open content were yet to be fully understood by the citizens of Kenya.



*Figure 4.3: Ownership of CC licensed open content*

From the 73 respondents in the content creators category, 68 respondents who constitute 94.4% of our respondents in this category answered in the affirmative as being in a position to create content that they would share. This is depicted in *Figure 4.3* above.

4. Are you willing to share this content after having licensed it accordingly with the government?

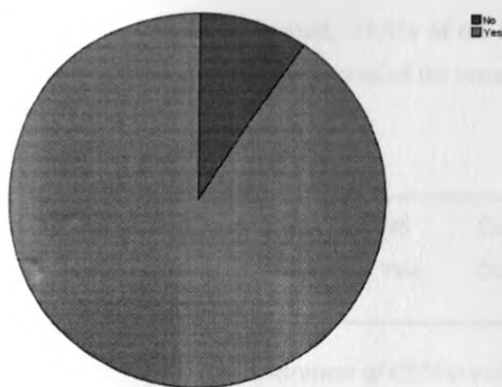


Figure 4.4: Content Creators Willingness to share content

The findings above clearly demonstrate that Kenyan citizens who could also double up as content creators have access to CC licensed open content in one way or other. The research further established that 90.4% were willing to share this content with the Kenyan Government after having licensed it accordingly to enhance service delivery.

**Factors to elicit Content Sharing**

The study also sought to establish the factors that would encourage content sharing from the literate population.

The study found that countrywide attribution and awareness of the benefits of content sharing were the widely cited motivations. A cumulative percentage of 75% of the respondents cited these. Only 25% cited monetary incentives. This indicated that money was not the only motivation for people to share open content but on the contrary, information on the benefits of the content shared as well as attribution and recognition would indeed encourage sharing of open content.

**What do you think can motivate people (citizens and non-citizens of Kenya) to supply content they have licensed under CC licenses to the GoK to allow the government to scale it up to be used countrywide?**

	Frequency	Percent
If they are given countrywide attribution and recognition	27	37.5
If they are informed on the benefits of content sharing	27	37.5
If they are paid in cash	18	25.0
Total	72	100.0

Table 4.2: Factors to elicit CC licensed open Content sharing

Concerning factors that would elicit more participation from the content creators in terms of incentives to share the content they had created or owned, 37.5% of the respondents said if they were given countrywide attribution and recognition, 12.5% cited being informed of the benefits of content sharing and 25% cited monetary rewards.

10. Have you ever used CC licensed content e.g MIT Open Courseware?	No	Count	45
	Yes	Count	27

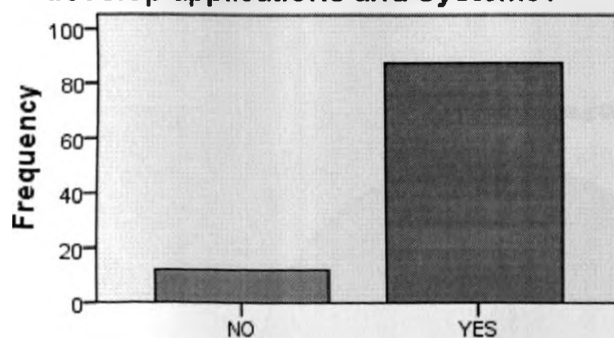
Table 4.3: Usage/ Awareness of CC licensed content

Out of the 72 valid responses from the content creators, only 27 had used CC licensed content before and 45 had not used any CC licensed content. This is a mere 37.5% of the total respondents. This showed that there was a need to raise awareness of the CC licensed content in Kenya so as to boost its usage to supplement the content that was already available to the government.

#### 4.3.2 Open Source Software Evaluation

For the purpose of evaluating open source application software, questionnaires were sent out to 123 software developers. Out of the 123, there were 100 respondents.

#### 1. Do you use Open Source Software to develop applications and systems?



1. Do you use Open Source Software to develop applications ...

Figure 4.5: Usage of OSS in application/ system development

Figure 4.5 above shows that 88% of our sample respondents build applications and software using open source software of various kinds. From this we can infer that the skill base for open source software is available.



The level of expertise of the developers however varied all the way from beginners to experts according to the frequency table below. See *Table 4.4*. However, it can be deduced from this finding that most (89%) of the developers ranged from intermediate to experts.

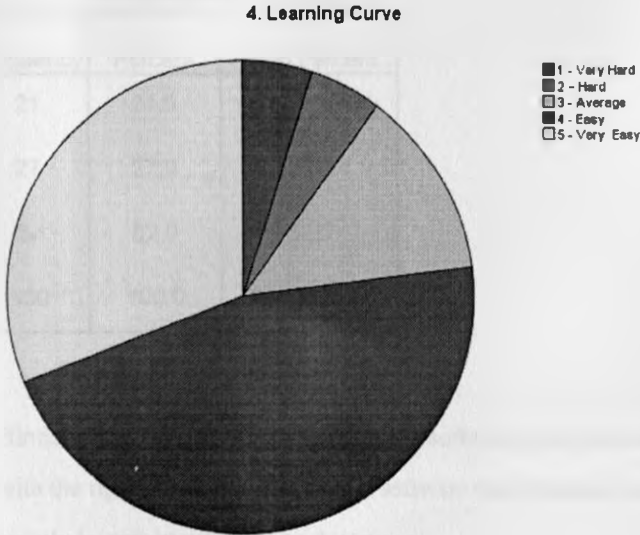
**Level of Expertise**

	Frequency	Percent	Valid Percent	Cumulative Percent
1 – Beginner	6	6.0	6.0	6.0
2 – Novice	5	5.0	5.0	11.0
3 – Intermediate	22	22.0	22.0	33.0
4 – Skilled	42	42.0	42.0	75.0
5 – Expert	25	25.0	25.0	100.0
Total	100	100.0	100.0	

*Table 4.4: Proficiency Levels of OSS developers*

The various Open source software, specifically PHP which is one of the most commonly used languages for open source development was evaluated using several criteria and the findings were as outlined below.

**4.3.2.1 Learning curve**



*Figure 4.6: Learning Curve*

The study also sought to establish the ease of learnability of open source software (PHP) as a software development language. 77% of the respondents cited the PHP language as being very easy or easy to learn so that it can be used to build applications and systems.

#### 4.3.2.2 Stability

In terms of stability during development, the developers' responses were as outlined below.

**Stability**

	Frequency	Percent	Valid Percent	Cumulative Percent
1 – Very Unstable	5	5.0	5.0	5.0
2 – Unstable	28	28.0	28.0	33.0
3 – Stable	26	26.0	26.0	59.0
4 – Very Stable	41	41.0	41.0	100.0
Total	100	100.0	100.0	

*Table 4.5: Stability*

It can clearly be established from these responses (67% of the respondents) that in terms of stability open source software (PHP) is seen to be fairly stable and thus can be used to build applications that can be used by the government.

#### 4.3.2.3 Performance

Open source software was also analyzed for performance and the outcome was as illustrated in the frequency table below.

**Performance**

	Frequency	Percent	Valid Percent
1 – Slow	21	21.0	21.0
2 – Average	27	27.0	27.0
3 – Fast	52	52.0	52.0
Total	100	100.0	100.0

*Table 4.6: Performance*

52% of the respondents affirmed the performance of Open source software, citing it as fast in development. From this we can conclude that with the right resources, open source software can effectively be used to build applications that can be utilized meritoriously in various government departments.

#### 4.3.2.4 Scalability

Open source software was also evaluated in terms of scalability and the findings from our sample respondents are displayed in the table below.

**Scalability**

	Frequency	Percent	Valid Percent	Cumulative Percent
1 – Poor	10	10.0	10.0	10.0
2 – Fair	10	10.0	10.0	20.0
4 – Good	24	24.0	24.0	44.0
5 – Excellent	56	56.0	56.0	100.0
Total	100	100.0	100.0	

*Table 4.7: Scalability*

56% of our sample developers were of the inclination that the scalability of the applications developed using open source software applications were scalable and 24% cited that the applications were fairly scalable. This shows that open source software applications can be scaled up accordingly based on the current needs and used at a larger scale than they were initially developed for and can also evolve as needed.

#### 4.3.2.5 Interoperability

In terms of interoperability with other applications built using different platforms, open source software was evaluated and the results are tabulated below.

**Interoperability**

	Frequency	Percent	Valid Percent	Cumulative Percent
1 – Very Hard	10	10.0	10.0	10.0
2 – Hard	20	20.0	20.0	30.0
3 – Easy	33	33.0	33.0	63.0
4 – Very Easy	37	37.0	37.0	100.0
Total	100	100.0	100.0	

*Table 4.8: Interoperability of Open source software*

37% of the OSS developers were of the disposition that interoperability of OSS applications with programs built using other kinds of software was very easy, which is very useful as it is vital that the applications that are built using OSS are interoperable with other programs for seamless integration.

#### 4.3.2.6 Extendibility

**9. Extendibility**

	Frequency	Percent	Valid Percent	Cumulative Percent
1 - VeryHard	5	5.0	5.0	5.0
2 - Hard	5	5.0	5.0	10.0
3 - Intermediate	15	15.0	15.0	25.0
4 - Easy	38	38.0	38.0	63.0
5 - Very Easy	37	37.0	37.0	100.0
Total	100	100.0	100.0	

*Table 4.9: Extendibility*

Extendibility in terms of the ease of adapting open source software products to changes of specification was evaluated. It was found that open source software was soft, and indeed is in principle as nothing can be easier than to change a program if you have access to its source code. This is the case with open source software and the responses above affirm this. 75% of the respondents attested to the fact that open source software applications are extendible to changes in specifications which is a very useful quality especially in government applications as needs change and they need to be continually addressed.

#### 5.3.2.7 Standards

**Standards**

	Frequency	Percent	Valid Percent	Cumulative Percent
1 - Poor	5	5.0	5.0	5.0
2 - Good	18	18.0	18.0	23.0
3 - Very Good	41	41.0	41.0	64.0
4 - Excellent	36	36.0	36.0	100.0
Total	100	100.0	100.0	

*Table 4.10: Standards*

Open source software standards were also assessed and 41% of the respondents responded that they were very good and 36% cited that they were excellent. This goes to show that open source software standards were relatively high and could be used to build applications that could be used in the government. The findings here indicated that open source software, and in this the example of PHP possessed most of the qualities of good software.

### 5.3.2.8 Documentation

**Documentation**

	Frequency	Percent	Valid Percent	Cumulative Percent
1 - Scanty	6	6.0	6.0	6.0
2 - Moderate	26	26.0	26.0	32.0
3 - Adequate	32	32.0	32.0	64.0
4 - Excellent	36	36.0	36.0	100.0
Total	100	100.0	100.0	

*Table 4.11: Documentation*

The survey also queried on the state of documentation of Open source software. The responses were generalized to all open source software. But the assumption was that other open source software would only defer minimally and that this would represent most of them adequately. 6% cited open source software as not being well documented whereas 36% which is slightly more than a third of our sample respondents recognized open source software as being more than adequately documented. A cumulative percent of 68 indicated that open source software was well documented.

### 4.3.2.9 Community Support

**Community Support**

	Frequency	Percent	Valid Percent	Cumulative Percent
1 - Poor	5	5.0	5.0	5.0
2 - Good	36	36.0	36.0	41.0
3 - Very Good	18	18.0	18.0	59.0
4 - Excellent	41	41.0	41.0	100.0
Total	100	100.0	100.0	

*Table 4.12: Community Support*

*Table 4.12* above summarizes the community support available for open source software. This was in terms of mailing lists, forums and so on that could provide support to open source software developers.

#### 4.3.2.10 Frequency of Updates

**Frequency of Updates**

	Frequency	Percent	Valid Percent	Cumulative Percent
1 - Very Infrequent	5	5.0	5.0	5.0
2 – Fairly frequent	32	32.0	32.0	37.0
3 - Frequent	22	22.0	22.0	59.0
4 – Very Frequent	41	41.0	41.0	100.0
Total	100	100.0	100.0	

*Table 4.13: Frequency of Updates*

The frequency of updates was also investigated. This was also a generalization as it was not easy to single out all the Open source software and thus they were evaluated in general. 41% alluded to very frequent updates of the software. Only 5% were of the opinion that the frequency of updates was not very high.

	Frequency	Percent	Valid Percent	Cumulative Percent
	90	90.0	90.0	90.0
Backtrack g++ Ubuntu metasploit	1	1.0	1.0	91.0
Codeigniter, Joomla, SMS Lib	1	1.0	1.0	92.0
PHP, Android	1	1.0	1.0	93.0
Linux Apache MySQL PHP	1	1.0	1.0	94.0
Linux OS, PHP, Python	1	1.0	1.0	95.0
Mysql Postgre sql Android				
Notepad++ Eclipse Netbeans	1	1.0	1.0	96.0
Php	1	1.0	1.0	97.0
PHP, MySql Linux	1	1.0	1.0	98.0
OS(Ubuntu) Subversion				
PHP/MYSSQL LINUX	1	1.0	1.0	99.0
CARTODB TILEMILL				
UNIX, Linux, Haiku (O S), Python, JAVA, PHP, Shell (Programming and scripting languages), MySQL and Postgres (database management systems), django and JAVA server Faces (Web development frameworks).	1	1.0	1.0	100.0
Total	100	100.0	100.0	

Table 4.14: FOSS Environments

The study also sought to determine which open source software environment the open source software developers were familiar with and a wide array of open source software platforms were floated which goes a long way to show that the local skill base is varied and capacity for building OSS applications of various kinds and with various functionalities for many useful government applications was indeed in place and what was needed was a link between the two disparate parties (the government and open source developers) for there to be meaningful utilization of the talent and also enhance operations in the various arms of the government.

### 4.3.2 Applicability of Open Source Software in Government

A questionnaire was also administered to top level management and policy makers in the Kenyan government as well as the I.T staff who were questioned about the applicability of open source software in the government among other things. Their responses were as summarized below.

#### 4.3.2.1 OSS Policies in Government

2. Is there an OSS policy in place in your Ministry, Agency or Department?

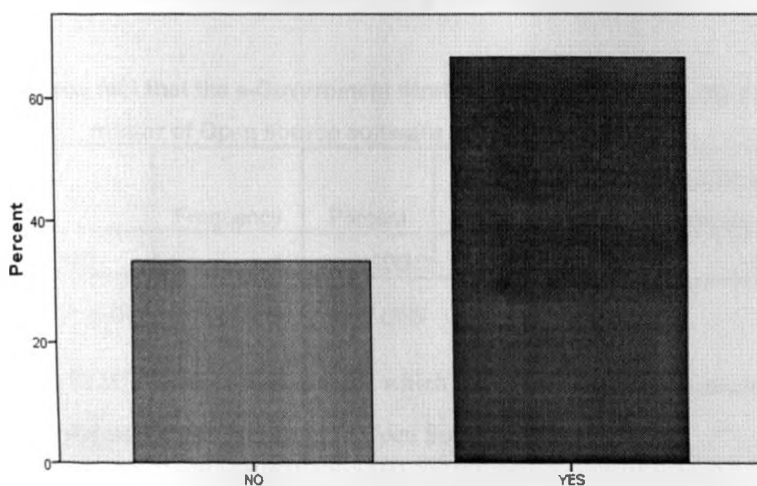
	Frequency	Percent	Valid Percent	Cumulative Percent
No	12	100.0	100.0	100.0

Table 4.15: Presence of OSS Policies

The ICT heads were questioned on the availability of Open source policies in their ministries. The responses clearly show that there is no Open Source software policy in place in the government as 100% of the respondents answered in the negative.

The question of whether Open source software systems were applicable for the government was also asked and the responses are as exemplified below.

4. Are Open Source systems or applications suitable for Government?



4. Are Open Source systems or applications suitable for Government?

Figure 4.7: Applicability of Open Source Software in Government.

66.7% of our sample respondents answered in the affirmative while 33.3% answered in the negative.

Those who answered in the negative were asked to elaborate on their response. This is detailed below.



**5. If your answer to #4 above is No, kindly state why not.**

	Frequency	Percent	Valid Percent
Support - No central point of support for Open Source Software, Learning Curve - Training and learning time required for familiarity with Open Source Software is high	4	33.3	33.3
Total	4	100.0	100.0

*Table 4.16: Why OSS is not suitable for Government*

As this was a question where only those who answered in the negative were to answer, the percentage of those who answered it was those who had answered in the negative and thus were 33.3%.

**4.3.3.2 e-Government Strategy and OSS current Status in the Kenyan Government**

The research also sought to investigate whether the current e-Government strategy dealt appropriately with the issue of Open source software. The responses are detailed below.

**1. Do you feel that the e-Government strategy deals adequately with the matter of Open source software use in Government?**

	Frequency	Percent	Valid Percent	Cumulative Percent
NO	12	100.0	100.0	100.0

*Table 4.17: e-Government Strategy and OSS*

All of the 12 ICT Heads who responded which is 100% of our sample indicated that the e-Government strategy did not deal adequately with the matter of Open Source Software.

#### 4.3.3.3 Usage of Open Source Operating Systems in the Government

Unix and/or Linux have been listed in the e-government strategy as some of the popular Operating systems for which standards will be developed. To what extent so far have these been deployed as part of the government's Operating systems?

	Frequency	Percent	Valid Percent	Cumulative Percent
>10%	12	100.0	100.0	100.0

Table 4.18: Usage of Linux as an OS in government

The percentage of usage of Open Source OS's in the government was also investigated and the responses showed that they had been deployed on a scale of less than 10% of the Total Operating systems. This showed that proprietary systems were still the Operating systems of choice in the government.

#### 5.3.3.4 Usage of Open Content Licensing in e-Government

3. Is Open Content licensing used in your Ministry, Agency or Department?

	Frequency	Percent	Valid Percent	Cumulative Percent
NO	6	50.0	50.0	50.0
YES	6	50.0	50.0	100.0
Total	12	100.0	100.0	

Table 4.19: Usage of Open Content licensing in e-Government

From the responses above, open content licensing was being utilized in the government and the level of usage was at 50%. This showed that there was room for improved usage of creative commons licensed open content for the government.

#### 4.3.3.5 Pilot Projects

**7. Does the e-government strategy and/ or related policies have the flexibility to allow for pilot projects to be undertaken in order to test, monitor and review selected OSS choices that might be considered for implementation in Government?**

	Frequency	Percent	Valid Percent
NO	3	25.0	25.0
YES	9	75.0	75.0
Total	12	100.0	100.0

*Table 4.20: Flexibility to allow for OSS Pilot projects*

The ICT Heads were also questioned on whether the current e-government strategy or related policies had the flexibility to allow for pilot projects to be undertaken to test, monitor and review selected OSS choices that might be considered for implementation by the e-government. 75% cited the flexibility was there but 25% mentioned inflexibility in as far as pilot projects were concerned. This reflects that there will be a need of an awareness campaign in order to win those in top level management who were opposed to having pilot projects to enable OSS usage.

#### 4.3.3.6 Operating System

**8. What is the most commonly used software in your Ministry, Government or Department on the desktop side?**

	Frequency	Percent	Valid Percent	Cumulative Percent
Microsoft Windows	12	100.0	100.0	100.0

*Table 4.21: Operating system*

**4. What is the most commonly used software in your Ministry, Agency or Department on the desktop side?**

	Frequency	Percent	Valid Percent	Cumulative Percent
	12	88.7	88.7	88.7
Mac OS	1	.9	.9	89.6
Microsoft Windows	11	10.4	10.4	100.0
Total	12	100.0	100.0	

*Table 4.22: Desktop Operating System*

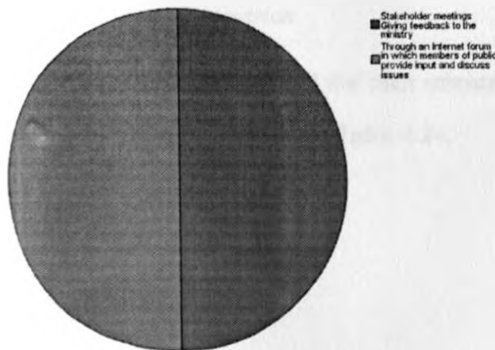
The study discovered that the operating system that was the most common on the desktop side was Windows. This showed that no Open source operating system was currently being used on the desktop. The first set of responses *Table 4.21* was from the ICT heads while the responses in *Table 4.22* were from ICT staff.

In general, most of the government departments indicated that they were using Open content licensing in one way or another but very few were actually using Open source applications. However, the study discovered that on the desktop side, proprietary software is still widely used. In particular Microsoft Windows operating systems. Other operating systems that are being used minimally on the desktop side are Mac OS.

**5.3.3.7 Citizen feedback to the Government**

One of the medium term goals of the e-Government strategy was to increase the input of the citizens into public sector decisions. The ICT heads were asked on how this was currently being addressed. 50% said through stakeholder meetings where they gave feedback to the ministry while 50% of the respondents said it was through an Internet forum in which members of the public provided input and discussed issues. This showed that the government was very willing to accept input that would better influence decisions and supplement service delivery from the citizens.

10. One of the medium term goals of the e-Government strategy is to increase the input of citizens into public sector decisions and actions. How is this being addressed currently?



*Figure 4.8: Input of Citizens into Public sector decisions and actions*

#### 4.3.3.8 In-house OSS Skill base in the Government

They were also questioned on whether there was capability in-house to initiate OSS projects. The responses were as below.

**12. Is there capacity in terms of skilled personnel who can be able to initiate in-house OSS development and customization projects?**

	Frequency	Percent	Valid Percent	Cumulative Percent
YES. But it can be supplemented when required.	12	100.0	100.0	100.0

Table 4.23: In house OSS capability

The responses in Table 4.23 above showed that as much as there was capability in-house, that is within the government to initiate OSS development and customization of projects, it could be supplemented as often as it was required especially when there was little in-house capacity.

#### 4.3.3.9 Incentives for Content provision

They were also probed on whether their Ministries would be willing to offer incentives to citizens who participated in useful content creation. Their responses were varied as illustrated below.

**17. Would your Ministry, Agency or Department be willing to offer some form of incentives to citizens who participate in useful content creation which can enhance service delivery?**

	Frequency	Percent	Valid Percent	Cumulative Percent
NO	6	50.0	50.0	50.0
YES	6	50.0	50.0	100.0
Total	12	100.0	100.0	

Table 4.24: Incentive provision

50% of the respondents answered that their ministries were willing to provide incentives to citizens who participated in sharing useful open content, see Table 4.24.

**18. If the answer to #17 above is YES, what form might these incentives most likely take?**

	Frequency	Percent	Valid Percent	Cumulative Percent
Monetary	6	50.0	50.0	50.0
Non-Monetary (NMR)	6	50.0	50.0	100.0
Total	12	100.0	100.0	

*Table 4.25: Monetary vs. Non-Monetary Incentives*

Half the respondents thought the incentives would take a Monetary form whereas 50% cited Non –Monetary incentives for useful content sharing. This confirms that either monetary or non-monetary incentives could be issued to those citizens and content creators who were willing to share their creative commons licensed open content with the government. See *Table 4.25*.

**4.3.3.10 Factors to enhance usage of OSS in Government**

**13. In your opinion, what do you think can be done to enable more use of Open source software and all its benefits in e-Government?**

	Frequency	Percent	Valid Percent	Cumulative Percent
Open source software training should be implemented and have the officers thoroughly trained on how to use it.	6	50.0	50.0	50.0
There should be a policy in place dictating ratios of proprietary and OSS usage in government	6	50.0	50.0	100.0
Total	12	100.0	100.0	

*Table 4.26: Open source s/w usage enhancement*

50% of the respondents cited that if a policy was put in place to dictate ratios of proprietary and OSS software usage in the government then OSS would have a better chance of being adopted and used in government. As this was an optional question, 50% of the respondents did not answer it.

#### 4.3.3.11 Citizen Participation in Content Sharing

In general it can be observed that half of the top level management and policy makers envisioned some participation from the citizens in terms of generation of open content for utilization by the government.

#### 15. Do you envision participation of citizens in generation of useful content properly licensed for use in Government?

	Frequency	Percent	Valid Percent	Cumulative Percent
NO	6	50.0	50.0	50.0
YES	6	50.0	50.0	100.0
Total	12	100.0	100.0	

Table 4.27: Citizen input in Content Generation

The ICT heads also had a section in the questionnaire where they were allowed to add any comments or suggestions they had about the topic. Most of them added that OSS required some sort of support agreement. They cited that this becomes complicated to governments annual budgeting cycles. As no ICT Officer would want to be chasing procurement to renew these agreements every financial year that is why most had a preference for proprietary software.

#### 4.4 Capacity to Support and Maintain Open Source Software Application in Government

From the 32 ICT staff in the government who the questionnaire was administered to, 28 responded. They were being asked about their proficiencies in open source software, deployment of open source applications among others things.

Their responses were as detailed below.

#### 11. What is your proficiency in terms of capability to troubleshoot and provide support and effective maintenance for Open Source Applications being used by your Ministry, Agency or Department?

	Frequency	Percent	Valid Percent	Cumulative Percent
Excellent	16	88.7	88.7	88.7
Good	3	2.8	2.8	91.5
Average	7	6.6	6.6	98.1
Poor	2	1.9	1.9	100.0
Total	28	100.0	100.0	

Table 4.28: Proficiency to support OSS Applications

Most of the respondents (88.7%) graded themselves as excellent in terms of ability to troubleshoot and provide support for OSS applications.

**4.4.4.1 Content Availability**

The question of Open content applicability was also asked. The responses showed that the majority of the ICT staff deemed it useful if they could be able to access data from a shared pool.

**15. Do you think it would be helpful if you could be able to access data from a shared pool of digital content to supplement the content your Ministry, Agency or Department has to improve services that you provide?**

	Frequency	Percent	Valid Percent	Cumulative Percent
NO	4	12.9	12.9	12.9
YES	24	87.1	87.1	100.0
Total	28	100.0	100.0	

*Table 4.29: Content Availability*

It is based on this literature and the responses obtained from this study that an open source Content sharing portal was designed and prototyped. This involved designing the process narrative, flowchart and then writing PHP scripts for actualizing the functionality. Tests of uploading and downloading data were carried out to ascertain the functionality and reliability of the system. This can be implemented at the Customer relationship management and enterprise architecture of the e-government framework to improve the interaction and contribution of citizens to CC licensed Open Content using open source software applications for efficient and reliable service delivery.



## **CHAPTER FIVE: SYSTEM ANALYSIS AND DESIGN**

### **5.0 Introduction**

The purpose of collecting data for this project was to carry out a current situation assessment of open source software usage and awareness as well as creative commons open licensed content in the government and also in the literate population with a view to enhancing the adoption and utilization of these in the Kenyan government. This chapter presents the research findings and the researcher's interpretation from data collected from respondents.

### **5.1 System Analysis (Requirement Definition)**

The first part of this section will deal with the specification of the requirements. Before deciding how the interface is going to work, one should always consider what kind of users will eventually be working with it, and what exactly is expected of the system. Therefore an analysis of User Classes is detailed below which enabled correct and intuitive system design in a manner befitting the users.

The second section of this chapter will focus on the analysis of the data that the desired system will have to manage. Which tasks will it be used for, how to structure all data in the scope of the system so that it's easy to manage, and what overall "look-and-feel" should be obtained. It also takes into consideration the expectations of the users who will be using the system.

#### **5.1.1 User Requirements**

- Consys should be usable countrywide
- The system should be web-based
- The system should allow registration of multiple users
- Consys should allow moderation of content
- Consys should be able to present data in a clear format
- Consys should have capability of filtering information

### 5.1.2 User Classes

#### Administrator

Type of user:	Support.
Experience with the system:	Expert.
Frequency of use:	When they need to check on the content that can be utilized and is resident on the portal. When they need to authorize I.T user staff.
Computer experience:	Advanced general computer skills.
Education/intellectual abilities:	A computer scientist or equal by experience.
Number of users:	1 (could be more)
Motivation for using the system	Keeping the system running, making everything possible. The administrator could be a selected personnel from e-Government
Tasks performed	Checking for new content

Table 5.1: Administrator User Class

#### Moderator

Type of user:	Direct.
Experience with the system:	Intermediate
Frequency of use:	Depending on the amount of content to be reviewed.
Computer experience:	Basic general computer skills
Education/intellectual abilities:	Expert in the subjects of the content assigned to them for review. Mostly people from the e-Government environment.
Number of users:	1
Motivation for using the system	Natural interest for the particular content they are assigned to review.
Tasks performed	Reviewing content for the various ministries.

Table 5.2: Moderator User Class

### Author/ Content Creator

Type of user:	Direct.
Experience with the system:	Novice – Expert
Frequency of use:	As often as they have new or revised content to upload.
Computer experience:	Basic general computer skills
Education/intellectual abilities:	Creative mind
Number of users:	No limit, every person is allowed to submit content.
Motivation for using the system	Presenting his research, artwork, content etc, exchanging knowledge.
Tasks performed	Submitting content in various formats

Table 5.3: Author User Class

### 5.1.3 System Use Case Diagram

The system prototype development began with specifying, visualizing, constructing and documenting the components of the systems. A System Use Case diagram was drawn to describe the set of scenarios of interaction between a user and the system. The actors and Use cases of the system were identified and represented as shown in Figure 5.1.

#### i) Actors

User

Administrator

Moderator/ Reviewer

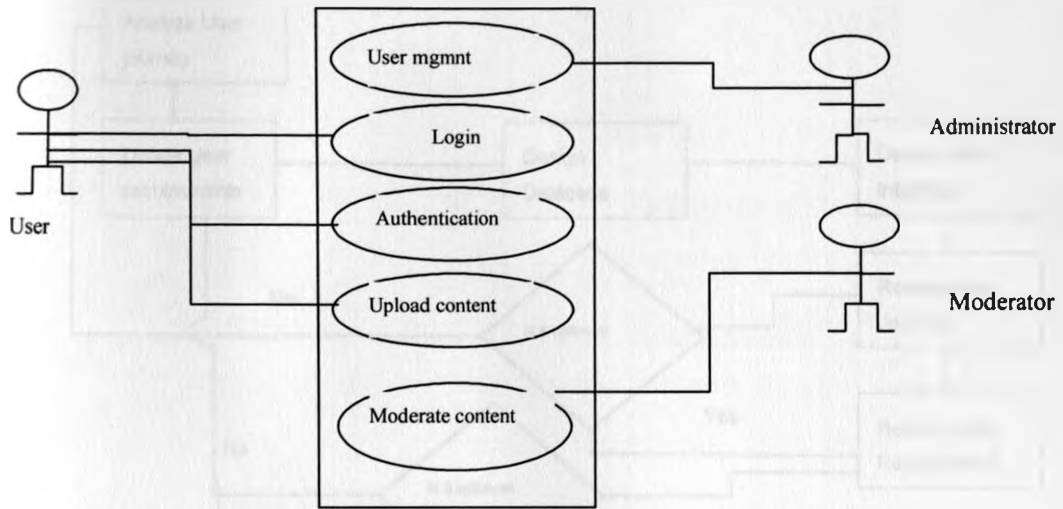
#### ii) Use Cases

Logging in

Uploading content

Managing users

Moderating conten



*Figure 5.1: System Use Case*

5.2 System Design Flow Chart

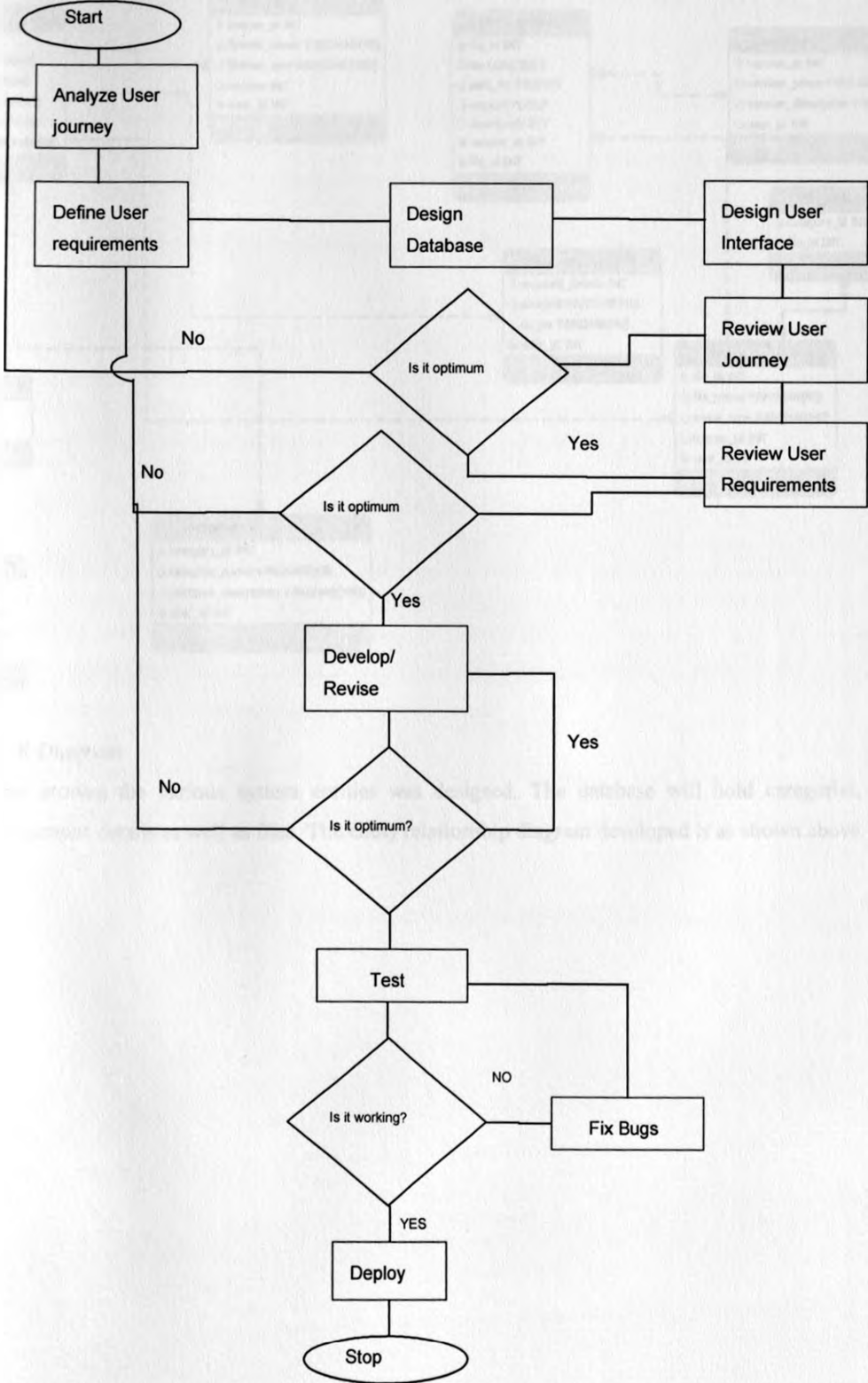


Figure 5.2: System Design Flowchart

### 5.3 Database Design - E-R Diagram

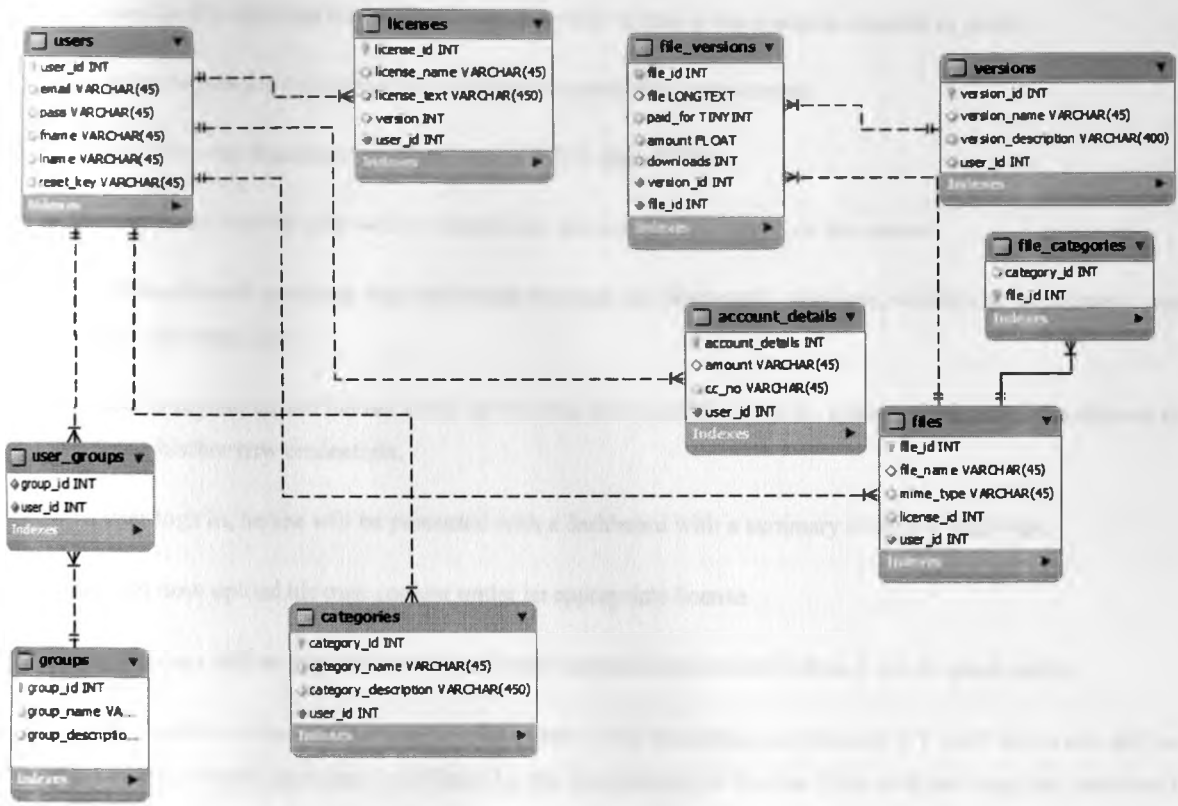


Figure 5.3: E-R Diagram

A database for storing the various system entities was designed. The database will hold categories, licenses, versions, user account details as well as files. The entity relationship diagram developed is as shown above.

#### **1.2. System Narrative**

- 1) When user land on the page for the first time, they will be able to view content marked as public.
- 2) User will be required to register online in order to upload their own content.
- 3) User interaction with the system will be through HTTP and HTTPS.
- 4) Information entered by the user will be routed over the internet and stored on the server.
- 5) User will be allowed to change their information except for their email addresses, which will be uniquely used to identify individual users.
- 6) Once a user is registered and his/her email address has been confirmed by the system, the user will be allowed to log using his/her new credentials.
- 7) Once a user logs in, he/she will be presented with a dashboard with a summary of his/her activities.
- 8) The user can now upload his own content under an appropriate license.
- 9) Uploaded content will be approved by the reviewer/ system administrator before it can be made public.
- 10) Content uploaded on the site can now be moderated by the designated government ICT staff whose role will be to find useful content and have it utilized by the Government of Kenya. They will be using the reviewer's comments and quality of work and relevance to subject matter.
- 11) The system will attempt to ease the work of the designated government ICT staff by allowing general users to rate content. Highly rated content will be assumed to be of high quality and the most relevant three will be displayed on the home page.
- 12) The rating will eventually contribute to winning of prizes and winners will be picked on the 1<sup>st</sup> of every month.

## 5.5 Software Requirements and Configuration

Open source software was used to build this system. The web server and PHP Code editor were installed on the same computer. The software used included:

### 5.5.1 XAMPP Server

Xampp is software that has a combination of apache (Web Server), PHP (Scripting Language) and MySQL (Database) on it. This is a webserver and was installed for local web administration using MySQL database and Apache. The version of XAMPP used for this project is XAMPP 2.5. A database named Consys was created.

### 5.5.2 Code Igniter

Code igniter which is a powerful PHP framework with a very small footprint was used for coding and testing PHP scripts.

## 5.6 System Testing

Testing the functionality of the system was done to ascertain whether the system could upload files under various licenses and have them downloadable to facilitate sharing of CC licensed Open content. Various users signed up and shared content that could be moderated and then shown publicly to the government and to the world.

### 5.6.1 Content Upload

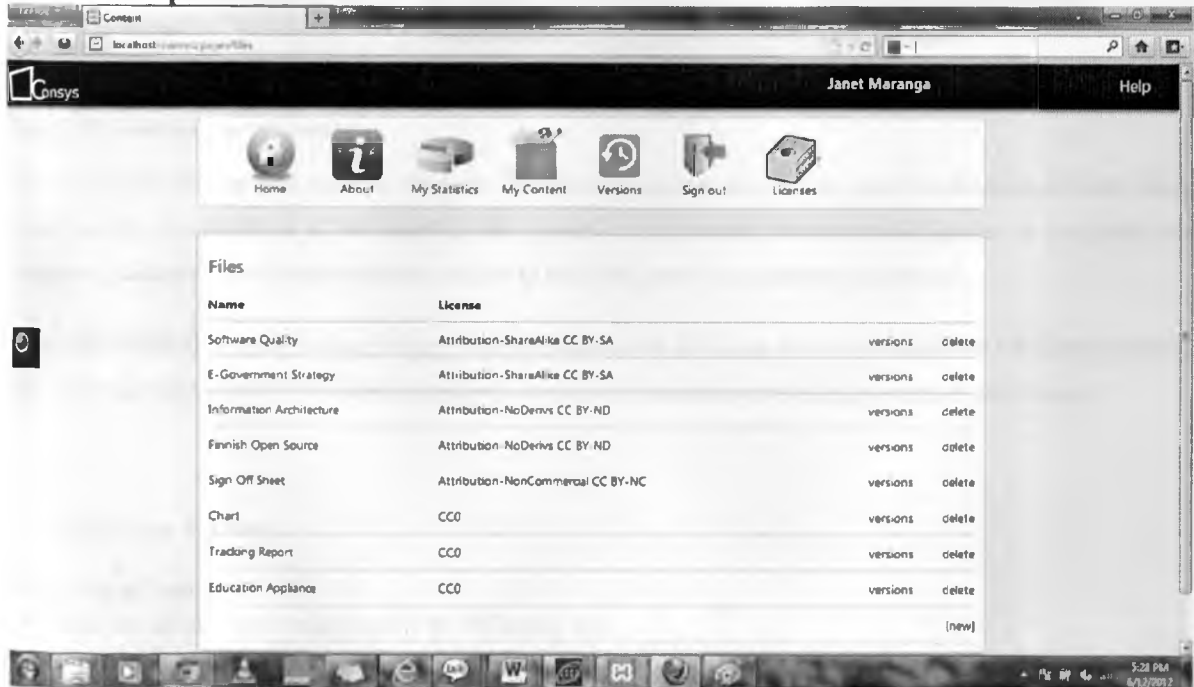


Figure 5.4: Sample Test Data Output



## 5.6.2 Moderation Testing



Figure 5.5 Moderation of Content

The moderation feature was tested to find out whether content that was put up could be moderated before being made publicly available to all the users of the system. All approved content would appear to the public but unapproved content would not be publicly visible to the public until the moderator approved it.

This study found that open content sharing with the appropriate licensing tools was feasible in the Kenyan context and there was a good match between the content creators/ citizens and content the government might need.

## 5.6.3 Validation Testing

### 5.6.3.1 Login Test

The validation of the log in was tested in the following way:

- entering invalid data into the fields
- pressing the login button
- checking that the login page is shown again
- entering valid data into the fields

- pressing the login button again
- checking that now the index page is shown

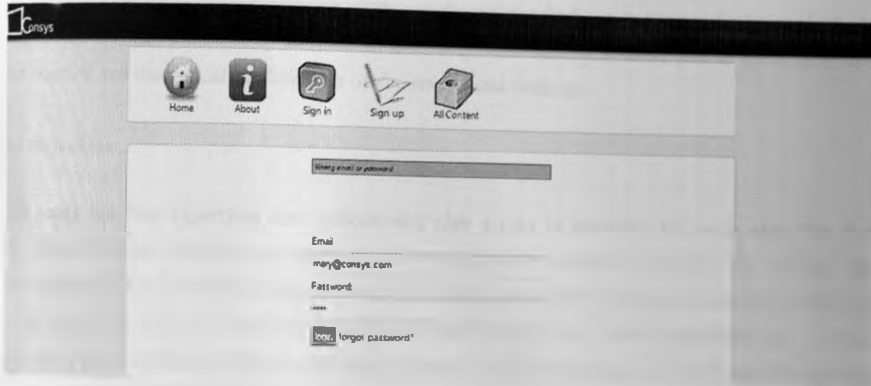


Figure 5.6: Login Test

## CHAPTER SIX: DISCUSSION

This section will discuss the implication of the results and the findings in relation to the research objectives, research framework and the methodology adopted in the study. This section is important in developing an understanding of the practical and theoretical implications of the results and findings.

### 6.1 Objectives

This study has four objectives that collectively play a role in attaining the main aim. The first objective is to investigate current e-Government frameworks in use globally in relation to OSS. The findings reveal a number of governments in the world both developed and developing which have various formalized guidelines that govern both procurement and usage of OSS in government and its agencies. These guidelines take various forms such as mandating the adoption of OSS, undertaking research and development on OSS, developing comprehensive OSS policies among others. All these require input from different stakeholders.

The second objective of the study was to explore the flexibility of the current ICT policies and strategies in as far as adoption of OSS and CC licensed open content were concerned. Pursuant to this objective, top level e-government officials and decision makers were questioned in regard to the current status of these. The same questionnaire was presented to the officials. The questionnaire was informed by the research framework which had the aspect of organizational capability which partly consisted of the plans and policies required to support and maintain a high level of digital services in support of the e-government's mission. The findings reveal that most of the e-Government ICT heads felt that the e-Government strategy did not adequately deal with the matter of OSS use in e-Government. Furthermore, the findings reveal that the ICT heads strongly felt that if clear policy guidance on OSS procurement and usage was developed for the Kenyan e-Government, and further train officers on open source products, it would go a long way in enhancing the uptake of OSS software and the subsequent applications that could be built from it. These findings suggest that the e-Government officials needed to reconvene and work on some detailed guidelines on OSS procurement and policy which would inform the usage of OSS in e-Government and its various agencies. It is however noteworthy that it was indicated that there was a FOSS strategy paper which had been indicated to be in progress, but nothing had as yet been published.

The third objective was to determine a model that enables content sharing and utilization of CC licensed open content. For a system to attain this objective, it should be able to accept information and offer mechanisms for licensing it under various CC licenses, store this information in a systematic and easily retrievable way so as to be usable and provide information that is relevant to various users. An examination of summary statistics reveals that the system developed allowed users to log in to the system, upload content under any of the various CC licenses, input the version of their document as well as view their statistics in terms of downloads. The uploaded content did not automatically go public so as to avoid offensive and misleading content being shared on the site. To mitigate against this a moderator role was created on the system which allowed for all the content to be approved/ moderated

before being viewable on the site. This will go a long way in ensuring the content on the portal reflects well on e-Government and remains continually useful and relevant. The model incorporated a provision where users would be able to rate content that had been made public so that the top three apps and open content were displayed on the home (landing) page. A competition was also launched which would serve to encourage contributors of apps based on open source software as well as contributors of open content. It worked using the rating of the apps and content where the contributors of the top three most highly rated apps and open content would be awarded with various prizes on the first of every month. The portal also has filtering capabilities which make information uploaded on the portal be able to be filtered by Ministry (i.e the ministry they had been uploaded under) thereby making retrieving information easy and intuitive.

The fourth objective was to determine strategic interventions to enhance the adoption of OSS and CC licensed open content and disseminate them to the Kenyan government. This had to be initiated with finding out the current OSS situation in the government so as not to duplicate efforts in areas which had already been previously covered. In our survey, we investigated the extent of the deployment of OSS technologies in government departments and state agencies. Our investigation indicated that there was some indication of usage of Open Source OS's in the government as there was a deployment of Linux as an operating system on a scale of less than 10% of the Total Operating systems. Regarding open source software in general, the respondents from the e-Government cited that open source software were good only if the officers were trained on how to use them. This led to the realization that training on OSS technologies was vital in enhancing the uptake of OSS in the government as the users will already be having the skills necessary to run these kinds of systems. It also came out clearly from the research study that there was need for top level management support in open source spearheading so that it could be used to catalyze strategic change throughout the country.

## **6.2. Research Framework**

The research framework that informed this study had four major aspects which together enabled the researcher follow a systematic and scientific way in this research.

The questionnaires that were used for data collection were also informed by our research framework of choice.

**E-Government decision/ policy makers** – This was to find out about the organizational capability of the government in terms of the policies, plans strategies and management processes required to mandate the development of Open source software applications as well as actualize the content sharing for content licensed under the creative commons licensed open content.

**E-Government I.T Staff** – This was to get feedback on the enterprise architecture of the Government in terms of the desired relationships among business management processes and information technology. They will also shed light on the rules and standards for optimizing and maintaining IT investments which in this case are open source software applications and portfolios and content sharing enablers.

**Citizens** – This was for customer relationship Management to enable us find out what the citizens who were the customers in our case needed and learned how to meet and continuously improve customer service in relation to CC licensed open content delivery.

**Open source software developers** – As this was the skill base that was going to be utilized going forward various aspects such as their skillset in Open source software, their perception of the various features and functionalities and robustness of various Open source software among others were explored.

**Creators of Content** – This was to find out from the literate population their willingness to share any useful content they may have created over the years or had access to under various CC licenses and which way their participation could be elicited in the most optimal way.

A critical examination of the design of the system shows that the components included in the system have targeted the four main sections namely customer relationship management, organizational capability, enterprise architecture and security and privacy.

Utmost care was taken in order to present the collected expectations in their original way. This would provide strong basis for the Kenya government to find solutions and strategies for facilitating the uptake of open source applications as well as creative commons licensed open content in the light of collected data.

### 6.3. Methodology

In relation to the methodology, there are a number of limitations highlighted in the findings. First, the conclusion that the Content sharing portal enhances OSS and CC licensed open content is reached subject to a number of assumptions. This is largely a result of the use of prototyping rather than coming up with a full-fledged system design. Further studies should look into the possibility of including treatment, control groups and blinding to accurately determining the implication of such a system on OSS and CC licensed open content adoption in the government. The inclusion of these measures would reduce the need for subjecting the findings to a myriad of assumptions. Another factor is that there is need for increase in the users of the system to even semi-illiterate persons and those who might not be I.T literate. In the current study, the participants were exposed (or used the system) for a one week period. It is not clear if this period is enough for the respondents that were drawn from the literate population. Making the system simple and easily understandable would enable all the users with different skills to be fully aware of the functionalities of the system so that they can help ensure they enjoy the full benefits of the system and its limitations. These methodological issues have to be addressed by future studies.

### 6.4 Strategic Interventions to enhance adoption of OSS and CC licensed content in the Kenyan Government

It is evident from the results of this research that OSS and CC licensed content usage within the Kenyan government departments is not yet extensive. This has been occasioned by various challenges such as lack of OSS policies in Kenya to govern OSS procurement and use, lack of awareness of OSS software products and benefits, no central point of support for OSS applications and so on. The proposed interventions are necessitated by the results of the

survey suggesting that despite OSS having several good attributes and qualities and being very relevant in usage in the government both from an applications point of view as well as operating system; the uptake has been very poor.

Governments such as those in Germany, Malaysia and Australia have developed comprehensive guidelines (models) that are used by their departments to proliferate OSS usage within ministries.

However there are some interventions that could be put in place to be utilized as a possible solution in overcoming the challenges and obstacles encountered currently by various government departments, thus improve and proliferate OSS usage within ministries and departments of the government.

- i) **Proper planning** – The first intervention in improving OSS usage within the Kenyan government should be proper planning. It is vital that all OSS implementation information be developed where there is a lack of these as is the situation currently. Come up with OSS policies, strategies and benefits of OSS and share them with all relevant stakeholders such as top level management, ICT personnel (support and development staff), external ICT service providers, software vendors, end users and business partners. This is also essential in minimizing resistance and negative influence. An element that would be of valuable importance is awareness campaigns. Within this part, we suggest that different awareness campaigns be initiated in order to ensure that every stakeholder, particularly end users who are the customers of the government are directly and indirectly involved in the implementation process. This will help in minimizing user resistance. Delegate the tasks of OSS implementation to certain ICT personnel rather than to all staff in order to not disrupt support for proprietary systems.
- ii) **Target environment** - Commence OSS implementation on the server side as there will be fewer users involved when servers are being migrated to new systems.
- iii) **Open Standards** – Select OSS alternatives which will enable interoperability with other solutions but ease future systems migrations.
- iv) **Re-skill** – ICT personnel and end users in the government should be trained thoroughly on open source software policies so as to enable them be better placed in maintaining the new software.
- v) **Pilot projects** – To ensure smooth implementation, pilot projects should be undertaken in order to test, monitor and review selected OSS choices. From the research findings the Kenyan government is flexible in as far as piloting is concerned so this can enable trialing of OSS applications and systems within a small environment. These will prove less disruptive but should be performed in a live user environment.
- vi) **Initiating in-house OSS development** and customization projects and supplementing with the OSS developers – This would enhance the skills of these developers.
- vii) Ensure some of the personnel are responsible for upgrading to newly released versions and customizing

where applicable as these are not done automatically.

- viii) Encourage developers to participate in various OSS activities such as in OSS forums, and global OSS projects which would in turn spill over to be utilized in government with some form of incentives.
- ix) Provide a forum for citizens' participation in Government activities e.g. scale up the Content sharing Prototype and incorporate more features that will aid citizen participation.
- x) Hold competitions on the content sharing portal in order to encourage software developers to compete to design and upload apps based on open source software to the portal for a chance to win fabulous prizes and CC licensed open content contributors to upload their content to the site and award prizes to the contributors of highly rated apps and content. Have a leader board which displays the names of the contributors of the best rated apps and content whereby earning a spot on the Leader Board will be a great achievement. This will motivate the contributors and engage citizens and ensure that they always strive to upload well thought out apps and content that will be useful in the Kenyan Government. This will encourage collaboration as well as visibility of the portal to more citizens and sharing as well as provide tangible and sustainable results which will reach a larger audience and can be used to facilitate the running of various sectors of e-government.

## **CHAPTER 7: CONCLUSION AND RECOMMENDATIONS**

### **7.1 CONCLUSION**

The motivation of this project was to come up with strategic interventions that would assist the Kenyan government adopt the use of Open source software applications and Creative commons licensed open content. Improving the content delivery method and awareness on the value of Open content for the Government and Government agencies as well as citizens and also linking open source software developers to the government would ensure full interaction between all these actors.

We believe the Kenyan government should be more proactive in drafting an open source policy and following the international developments in more detail. The government has an important function in signaling to software vendors what sort of standards and software are being needed in the public sector. Thus, we propose that the governments in the rich countries at large should in fact learn from the developing countries. More detailed and active open source policies for the different parts of the public sector can help in filling the current gaps in the software supply and demand. Both the local software companies and the public sector can potentially benefit.

This study found that it is possible for the citizens to provide open content to the government and also the developers would build various applications for the government under different agreements. This would greatly improve on information turn-around time within the government, its agencies and clients/ citizens.

### **7.2 FURTHER WORKS**

In the future, this work can be extended in the following directions:

#### **7.1.1. Knowledge management**

As Open Source Projects produce enormous amounts of data on mailing lists, web sites, repositories, in online communications, and other media and this data is very often dispersed across multiple sites and no one maintains it. Future efforts can be focused on extending the content sharing platform to have a knowledge bank where all this relevant information can be stored centrally. Also have staff in charge of maintaining this knowledge bank.

#### **7.1.2 Skill Matrices**

A central skill matrix can be included on the content sharing portal so that open source application developers and even contributors of CC licensed open content can classify themselves on a wide range of skills, the level they are at and their experience. This will make it easy to assess and plan resources as well as assign tasks on new open source projects and open content generation projects that the government may wish to undertake.



## REFERENCES

1. Aaker, D., et al, (1998), 'Marketing Research', Sixth edition
2. Anything Research, 2012. Anything Research, Better Data, Better Decisions. [online] Available at: <http://www.anythingresearch.com/> [Accessed 10 February 2012].
3. bridges.org report, "*Free/open source software (FOSS) policy in Africa: A toolkit for policy makers and practitioners*". [Online] Available at [www.ictregulationtoolkit.org/en/Document.3485.pdf](http://www.ictregulationtoolkit.org/en/Document.3485.pdf) [Accessed 13 December 2011].
4. Caldow, J., (1999), 'The Quest for Electronic Government: A Defining Vision'
5. Cost-Benefit Analysis Framework [Online] Available at <http://www.anythingresearch.com/Strategic-Planning/Cost-Benefit-Analysis.htm> [Accessed 23 January 2012].
6. E-Government evaluation: A framework and case study, M.P. Gupta\*, Debashish Juma
7. E-Government for development [Online] Available at <http://www.egov4dev.org/success/definitions.shtml#coverage> [Accessed 20 January 2012].
8. Flex Framework [Online] Available at <http://www.adobe.com/products/flex.html> [Accessed 23 January 2012].
9. Frameworks and Open Source Tools [Online] Available at <http://www.saventech.com/services-frameworks-opensourcetools.html> [Accessed 23 January 2012].
10. Gessi, Ramnarine and Wilkins (2006): *Introducing a New e-governance Framework in the Commonwealth: From Theory to Practice*. [Online] Available at <http://www.capam.org/documents/2008625112823.pdf> [Accessed 20 January 2012].
11. Getao (2004), Open Source Software Adoption in Kenyan Tertiary Education: Opportunities, Challenges and Methodology –*International ICT Workshop 2004 on Application of ICT in enhancing Higher Learning Education*, pp 34-51
12. Gordon, F. Thomas., (2002), 'E-Government Introduction', ERCIM, News No. 48

13. Gronlund, A., and Horan, T. (2005). Introducing e-Government: History, Definitions, and Issues. *Communications of the AIS*, Vol. 15, 713-729. pp 12-14.
14. Gupta and Jana *Government Information Quarterly* 20 (2003) pp.365-387
15. Johnson G., and Scholes K., (1993)'Exploring Corporate Strategy' 3<sup>rd</sup> Edition
16. Kothari, C.R. (2004) *Research Methodology, Methods & Techniques*
17. Lee, D.H.D.,(2005), 'E-Government: key success factors for value discovery and realization'
18. Literate population of Kenya - [Online] Available at  
<https://www.cia.gov/library/publications/the-world-factbook/fields/2103.html>  
[Accessed 1st June, 2012].
19. Marshall, M.N., (1996), 'Sampling for qualitative research'
20. Mclean, M. & J, Tawfik., (2003), 'The role of Information and Communication technology in the modernization of e-Government', PP 237-245
21. Mingers, J. and Gill, A. (1997) *Multimethodology: Theory and Practice of Combining Management Science Methodologies*.
22. Ntiro, S. (2000) *eGovernment in Eastern Africa*, KPMG, Dar-es-Salaam
23. Open Source Software in e-Government: *Analysis and recommendations drawn up by a working group under the Danish Board of Technology*, October 2002. [Online] Available at  
[http://www.tekno.dk/pdf/projekter/p03\\_opensource\\_paper\\_english.pdf](http://www.tekno.dk/pdf/projekter/p03_opensource_paper_english.pdf)  
[Accessed 21 January 2012].
24. Open source technologies. [Online] Available at  
<http://www.opensourcetechologies.com/technologies.html>  
[Accessed 23 January 2012].
25. Rama K.D. and Manas R.P., (2008) – *SOA for e-Governance in India: Potentials and Pitfalls*
26. Rogers E.M.; (1995); *Diffusion of Innovation*; 4<sup>th</sup> Edition; New York
27. Sherry L, Billig S, Tavalin F, Gibson D: *New Insights on learners technology adoption in communities of learners*
28. Shih C.F, Venkatesh A; *Beyond Adoption: Development and Application of a Use Diffusion model*; *Journal of Marketing* Vol. 68 (January 2004), 59-72
29. Strategic Planning with the SWOT Analysis. [Online] Available at  
<http://www.anythingresearch.com/Strategic-Planning/SWOT-Strengths-Weaknesses-Opportunities-Threats-Analysis.htm>

- [Accessed 23 January 2012].
30. Solis L. Hilda (2011), U.S. Department of Labor E-Government Strategic Plan. [Online] Available at [https://www.dol.gov/sec/e\\_government\\_plan/p10\\_e-gov\\_framework.htm](https://www.dol.gov/sec/e_government_plan/p10_e-gov_framework.htm) [Accessed 22 January 2012].
  31. Strategic Planning Technology Adoption Curve [image online] Available at: <http://www.anythingresearch.com/Strategic-Planning/Technology-Adoption-Curve.htm> [Accessed 12 January 2012]
  32. Strategic Planning SWOT Strength Weakness Opportunity Threat n.d [image online] Available at: <http://www.anythingresearch.com/Strategic-Planning/SWOT-Strengths-Weaknesses-Opportunities-Threats-Analysis.htm> <http://www.anythingresearch.com/Strategic-Planning/Technology-Adoption-Curve.htm> [Accessed 12 January 2012]
  33. Symphony Open source PHP 5 Web Framework. [Online] Available at <http://www.symfony-project.org/> [Accessed 23 January, 2012].
  34. Thomas, J. (2007). Malaysian public sector OSS program phase II: Accelerated Adoption. [Online] Available at [http://www.oscc.org.my/documentation/phase2\\_launching/OSS-Phase2Strategy-Plan-Launch.pdf](http://www.oscc.org.my/documentation/phase2_launching/OSS-Phase2Strategy-Plan-Launch.pdf) [Accessed 12 May, 2012]
  35. Using Open Source Software In the South African Government, *A Proposed Strategy Compiled By the Government Information Technology Officers' Council*. Version 3.3, 16 January 2003, <http://www.oss.gov.za> [Accessed 31 December 2011]
  36. Waterfall Methodology n.d. [image online] Available at: [http://www.learnaccessvba.com/images/application\\_development/Waterfall\\_model.png&imgrefurl=http://www.learnaccessvba.com/application\\_development/waterfall\\_method.htm](http://www.learnaccessvba.com/images/application_development/Waterfall_model.png&imgrefurl=http://www.learnaccessvba.com/application_development/waterfall_method.htm) [Accessed 20 January 2012]
  37. West, D. M. (2002). *Global e-Government, 2002*. [Online] Available at: <http://www.insidepolitics.org/egovt02int.html> [Accessed 30 December 2011].
  38. West, D.M., (2004), 'E-Government and the Transformation of Service Delivery and Citizen Attitudes', 'public Administration Review', Vol. 64, no.1 pp. 15-27
  39. Working Group on e-Government in the Developing World (2002): '*Roadmap for e-Government in the Developing World, 10 Questions E-Government Leaders Should Ask Themselves*'; Pacific Council on International Policy, The Western Partner of the Council on Foreign Relations; (Los Angeles: Pacific Council on International Policy, The Western Partner of the Council on Foreign Relations; (Los Angeles: Pacific Council on International Policy); 1-28. Available at

<http://unpan1.un.org/intradoc/groups/public/documents/apcity/unpan005030.pdf#search=%22Roadmap%20for%20E-Government%20in%20the%20Developing%20World%22>  
[Accessed 9 January 2012].

40. Rose India Technologies Pvt. Ltd. n.d. [online]  
Available at: <http://www.roseindia.net/opensource/open-source-frameworks.shtml>  
[Accessed 10 January 2012]
41. Yayehyirad Kitaw, 2006. E-Government in Africa-Prospects, challenges and practices  
Available at :< [http://people.itu.int/~kitaw/egov/paper/E-Government\\_in\\_Africa.pdf](http://people.itu.int/~kitaw/egov/paper/E-Government_in_Africa.pdf)>  
[Accessed 9 January 2012].
42. Yin, R.K. (1994), 'Case Study Research design and Methods. Thousand Oaks', 'CA: Sage publications'
43. Yin, R.K., (1981), 'The Case Study as a Service Strategy. Knowledge 3.', pp. 97-114

## APPENDICES

### APPENDIX A: Questionnaires

#### **Strategic Interventions to enhance Open Source Adoption and Creative Commons licensed Open content in the Kenyan Government Questionnaire**

##### **Introduction**

Dear respondent,

This is a survey I am conducting in order to find out Strategic Interventions that can be used to enhance adoption of Open Source Software applications and Creative Commons licensed Open Content in the Kenyan Government . However, the research requires you as a participant to have some basic understanding on software, both open source and proprietary software. Your responses are very important in enabling me to gain a better understanding about this topic.

I am a student at University of Nairobi. School of Computing and Informatics. This is in partial fulfillment of my course. You have been selected to take part in this study. I would be grateful if you would assist me by responding to all the attached questions in the questionnaire. The questionnaire should take you about 20 minutes to complete.

You will be asked a few questions about your opinion on factors affecting open source software applications adoption and creative commons licensed open content in the Kenyan government. Participation in this study is fully voluntary and you have the right to withdraw at any time with no penalty. I treat your participation as anonymously as practically possible. All data is treated as confidential and will be used for academic research purposes only. Your responses and others will be used as the main data set for my research project for my Master's degree in Computer Science at the University of Nairobi. The results of the study will be made available to the Directorate of e-Government after I have completed the data analysis. I hope that you will find completing this questionnaire a pleasurable experience. If you have any question or would like further information, please do not hesitate to email me at [janetmaranga@gmail.com](mailto:janetmaranga@gmail.com). Thank you very much for your time.

**Questionnaire for E-Government ICT Heads**

1. Do you feel that the e-Government strategy deals adequately with the role of Open Source Software and CC licensed open content use in e-Government?

\*Government in this case refers to your Ministry, Agency or Department.

- YES
- NO
- I don't know
- Other:

2. Is there an OSS policy in place in your Ministry, Agency or Department? \*

- Yes
- No

3. Is Open Content licensing used in your Ministry, Agency or Department? \*

- YES
- NO
- Other:

4. Are Open Source systems or applications suitable for Government?

\*Government in this case refers to your Ministry, Agency or Department.

- YES
- NO
- I don't know

5. If your answer to #4 above is No, kindly state why not.

Select all that apply

- Support - No central point of support for Open Source Software
- Learning Curve - Training and learning time required for familiarity with Open Source Software is high
- Unique requirements already catered for by Proprietary Software
- No guarantee of Regular Updates
- Other:

6. Unix and/or Linux have been listed in the e-government strategy as some of the popular Operating systems for which standards will be developed. To what extent so far have these been deployed as part of the government's Operating systems? \*Government in this case refers to your particular Ministry, Agency or Department

- >10%
- 11-25%

- 26-50%
- 51-75%
- <75%

7. Does the e-government strategy and/ or related policies have the flexibility to allow for pilot projects to be undertaken in order to test, monitor and review selected OSS choices that might be considered for implementation in Government?

\*The Government in this case refers to your Ministry, Agency or Department

- YES
- NO

8. What is the most commonly used software in your Ministry, Government or Department on the desktop side? \*Select all that apply.

- Linux/ Unix
- Microsoft Windows
- Sun Solaris
- Mac OS
- I don't know

9. Which of these applications are currently being used in the Government? \*Government in this case refers to your Ministry, Agency or Department. Select all that apply

- OpenOffice.org
- KDE office
- Microsoft Office
- Squirrel Mail
- Mozilla Thunderbird
- Mozilla Firefox
- Internet Explorer
- Other:

10. One of the medium term goals of the e-Government strategy is to increase the input of citizens into public sector decisions and actions. How is this being addressed currently?

\*You can explain how it is being addressed in your Ministry, Agency or Department.

11. Does the I.T training strategy currently in place in the e-Government for training Government personnel cover use of open source technologies?

\*You can explain how this is covered in your Ministry, Agency or Department

- YES
- NO
- I don't know

12. Is there capacity in terms of skilled personnel who can be able to initiate in-house OSS development and customization projects? \*This is in your Ministry, Agency or Department

- YES
- NO
- YES, But it can be supplemented when required.

13. In your opinion, what do you think can be done to enable more use of Open source software and all its benefits in e-Government?

14. Providing a forum for citizens' participation in Government activities is one of the specific objectives of the e-government. How is this being addressed currently in your Ministry, Agency or Department?

15. Do you envision participation of citizens in generation of useful content properly licensed for use in Government? \*

- YES
- NO

16. If yes to #15 above, please mention some categories of content that can be useful to your Ministry, Agency or Department. Kindly indicate your Ministry, Agency or Department in your answer.

17. Would your Ministry, Agency or Department be willing to offer some form of incentives to citizens who participate in useful content creation which can enhance service delivery? \*

- YES
- NO

18. If the answer to #17 above is YES, what form might these incentives most likely take? \*

- Monetary
- Non-Monetary (NMR)
- Other:



19. Please give your comments or suggestions about other ways citizens can take a more participatory role in enhancing the quantity and quality of services and material offered by Government. \*Government in this case refers to your Ministry, Agency or Department.

20. I would appreciate any thoughts you might like to add to your responses or to the topic in general.

21. Email (Optional)

Note:

Your personal detail will be kept confidential and will not be used for any other purpose apart from Academic Research purposes. Thank you for your kind cooperation.

Quote

**Questionnaire for E-Government I.T. Staff**

1. In your opinion are Open Source Software Systems suitable for Government? \*Government in this case means your Ministry, Agency or Department

- YES
- NO
- Don't know

2. If your answer to 1 above is No, kindly state why not. \*Select all that apply

- Support - No central point of support for Open Source Software
- Learning Curve - Training and learning time for familiarity with Open source software is high
- Unique Requirements already catered for by proprietary software
- No guarantee of regular updates
- Other: \_\_\_\_\_

3. To what extent have Open Source Operating systems e.g Linux been deployed as part of your Ministry, Agency or Department's operating systems? \*

- >10%
- 11-25%
- 26-50%
- 51-75%
- >75%
- I don't know

4. What is the most commonly used software in your Ministry, Agency or Department on the desktop side? \*Select all that apply.

- Linux/ Unix
- Microsoft Windows
- Sun Solaris
- Mac OS
- Don't know

5. What is the most commonly used Operating System installed on the Servers in your Ministry, Agency or Department? \*e.g. Linux, Windows 2003 Server, Windows 2008 Server

6. What is the Web Server of choice in use in most of the servers in your Ministry, Agency or Department? e.g. Apache, Tomcat etc.

7. What are some of the considerations of Software choice to be deployed in your Ministry, Agency or Department?

8. Have you faced any challenges in maintenance of I.T infrastructure in your Ministry, Agency or Department's services using any kind of Open Source technologies ?\*

YES

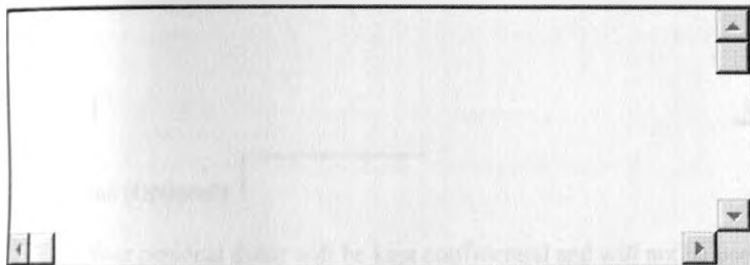
NO

Other:

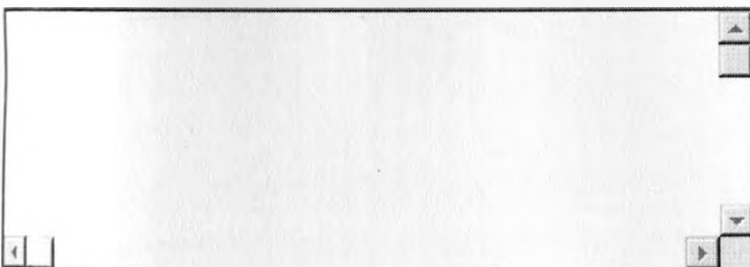
9. If the answer to Question 5 above is YES, kindly mention what kind of challenges these were.

NO

16. If yes to #15 above, please mention some categories of content that can be useful to your Ministry, Agency or Department? Kindly also indicate the name of your Ministry, Agency or Department in your answer.

An empty rectangular text box with a thin black border. It contains no text. There are small navigation icons (back, forward, up, down) in the corners.

17. Give some brief measures that you think the Government of Kenya can take to create a culture of sharing useful digital content by content creators to be used in enhancing the Government services?

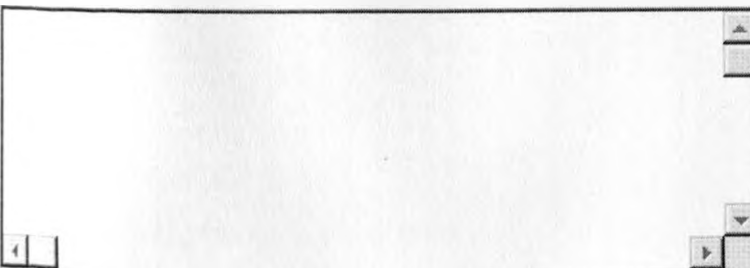
An empty rectangular text box with a thin black border. It contains no text. There are small navigation icons (back, forward, up, down) in the corners.

18. Has any of the training you have gone through at the Government covered the use of Open source technologies? \*

YES

NO

19. Please give your comments or suggestions about other ways in which citizens can take a more participatory role in enhancing the quantity and quality of material offered by e-Government in your Ministry, Agency or Department.

An empty rectangular text box with a thin black border. It contains no text. There are small navigation icons (back, forward, up, down) in the corners.

20. I would appreciate any thoughts you might like to add related to your responses or to the topic in general. (Kindly add below, if the space above is too limited)



**Questionnaire for Citizens**

**Brief explanation of Creative Commons licenses**

CC licensed open content stands for Creative Commons licensed open content and it helps you share your knowledge and creativity with the world. CC licenses that work can be published under include:

- Attribution alone (by)
- Attribution + NoDerivatives (by-nd)
- Attribution + ShareAlike (by-sa) Attribution + Noncommercial (by-nc)
- Attribution + Noncommercial + NoDerivatives (by-nc-nd)
- Attribution + Noncommercial + ShareAlike (by-nc-sa)

CC0 - a legal tool for waiving as many rights as legally possible, worldwide when releasing material into the public domain.

1. How often do you use the Internet? \*

- Daily
- Weekly
- Monthly
- Yearly
- Other: \_\_\_\_\_

2. Do you create/ own content that you can share? \*Articles, Photos, Research material, Books etc

- Yes
- No
- Not sure

3. If answer to 2 above is YES, what kind of content is it? \*Select all that apply

- Articles
- Photos
- Books
- Research Material
- Educational content e.g Exam questions
- Music
- Other: \_\_\_\_\_

4. Are you willing to share this content after having licensed it accordingly with the government? \*

- Yes
- No

5. If yes in 4 above, how would you like to be compensated?

- Payment in Cash
- Payment by use of electronic means e.g MPESA
- Payment in kind e.g getting recognition countrywide as the creator of some content
- If my content sharing can assist in one way or other in advancing knowledge in Kenya
- If the research findings/ outcomes can be shared with me
- If the content sharing objectives are explained to me
- I am willing to share content free of charge
- Other:

6. If No in 4 above, kindly explain why

7. What do you think can motivate people (citizens and non-citizens of Kenya) to supply content they have licensed under CC licences to the GoK to allow the government to scale it up to be used countrywide? \*Select all that apply

- If they are paid in cash
- If they are given countrywide attribution and recognition
- If they are informed on the benefits of content sharing
- I don't know
- Other:

8. Are you concerned about security of content you may have licensed using CC licenses and shared with the government? \*

- Yes
- No

9. If appropriate security measures were put in place to secure your content (the one you've shared) you were assured that you would be attributed as per your license, would you share the content? \*

- Yes
- No

10. Have you ever used CC licensed content e.g MIT Open Courseware? \*

Yes

No

11. If yes, to 10 above, did you find it useful?

12. Do you envision more interaction and positive change in the government services once you share your CC licensed Open content with the GoK? \*

Yes

No

13. Do you use Social Media? \*

Yes

No

14. If yes to 13 above, please tick the ones you use

Facebook

Twitter

Google+

LinkedIn

MySpace

Flickr

Tagged

Other:

15. Have you ever shared any content (photos, notes etc) on any of these social media platforms?

Yes

No

16. If yes to 15 above, then to some extent you were sharing content. Did you realize this at the time?

Yes



No

17. Email

I would appreciate any thoughts you might like to add related to your responses or to the topic in general.

Note:

Your personal detail will be kept confidential and will not be used for any other purpose apart from academic research purposes. Thank you for your cooperation.

Submit

## Questionnaire for Open Source Software Developers

For the purpose of this survey, the particular focus is on PHP software.

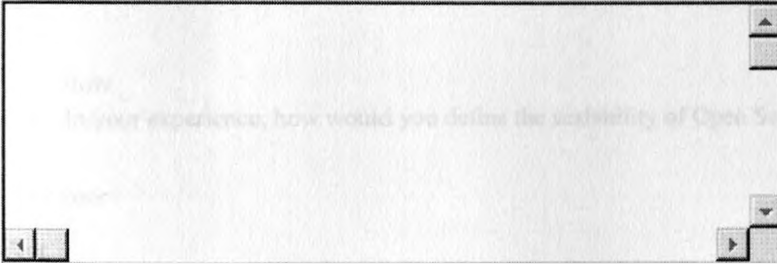
Top of Form

1. Do you use Open Source Software to develop applications and systems? \*

YES

NO

2. Which FOSS environments do you have experience in? \*Please list all of them below.



3. Level of Expertise \*What is your personal experience level with this software?

No answer

1 - Beginner

2 - Novice

3 - Intermediate

4 - Skilled

5 - Expert

4. Learning Curve \*Do you think it is easy or hard to learn how to use Open Source Software as a developer?

1 - Very Hard

2 - Hard

3 - Average

4 - Easy

5 - Very Easy

5. Stability \*Do you think Open Source Software is stable and reliable for production use? (e.g few critical bugs)

1 - Very Unstable

2 - Unstable

3 - Fairly stable

4 - Stable

5 - Very Stable

6. Performance \*In your experience, how would you define the performance of Open Source Software?

1 - Very Fast

2 - Fast

3 - Fairly fast

4 - Slow

5 - Very Slow

7. Scalability \* In your experience, how would you define the scalability of Open Source Software?

1 - Very Poor

2 - Poor

3 - Average

4 - Good

5 - Very Good

8. Interoperability \*How would you define the integration of Open Source Software with other technologies?

1 - Very Hard

2 - Hard

3 - Average

4 - Easy

5 - Very Easy

9. Extensibility \*Is it easy to extend Open Source Software functionalities with external plugins/ add-ons?

1 - Very Hard

2 - Hard

3 - Average

4 - Easy

5 - Very Easy

10. Standards \*How would you define the Open Source Software support for widely adopted standards?

1 - Very Poor

2 - Poor

- 3 - Average
- 4 - Good
- 5 - Excellent

11. Documentation \*What do you think of the documentation of the Open Source Software? (e.g readability, completeness, quality, useful examples, etc)?

- 1 - Very Poor
- 2 - Poor
- 3 - Fair
- 4 - Good
- 5 - Very Good

12. Community Support \*How would you define the technical support for Open Source Software offered on the technical forums/ mailing lists?

- 1 - Very Poor
- 2 - Poor
- 3 - Fair
- 4 - Good
- 5 - Very Good

13. Frequency of Updates \*How would you define the new releases containing new features, improvements and bug fixes?

- 1 - Very Rare
- 2 - Rare
- 3 - Average
- 4 - Frequent
- 5 - Very Frequent

14. Why do you think the Government of Kenya is not using a lot of Open Source Software? \*

- Lack of Awareness of Open Source Software
- Absence of skilled manpower to build Open Source Software Applications
- Lack of capacity of Open Source Software applications that are relevant
- Other:

15. Do you think the Kenyan government should embrace and use more OSS applications? \*OSS - Open Source Software

- YES

NO

16. What is your highest level of completed education? \*

17. Was FOSS part of this education?

Yes

No

18. Given a chance, would you build OSS applications to be used by the Kenyan e-Government? \*

YES

NO

19. If your answer to Question 15 above is YES, how would you like to be compensated?

20. If your answer to Question 15 above is NO, why wouldn't

Comments

I would appreciate any thoughts you might like to add to your responses or to the topic in general

Email: (Optional)

Note:

Your personal detail will be kept confidential and will not be used for any other purpose apart from Academic Research purposes

Submit



Consys USER MANUAL

**APPENDIX B: User Manual**



**Consys USER MANUAL**

## **Introduction**

This section will detail:

- How Users, Moderators and Administrators log in to the content sharing platform.
- An overview of the default buttons Consys uses.
- How to perform various tasks on Consys.

## **How to Access the Home Page**

In order to access the Consys Content sharing portal, type in the URL.

## **How to Access the Portal as an Administrator**

Input your Username and Password in their respective fields and click the Login button. The Control Panel (Home page) is displayed.





Home



About



Sign in



Sign up



All Content



## Welcome

Consys is a creative commons licensed contents sharing site, where you can publish and share most media types with the Kenyan Government. You have practically unlimited flexibility in choosing what you want to share, as long as it is your original work or work licensed under a Creative commons license which allows for free remixing and distribution of that work but with the proper attribution to the original content creator.

- [Click here](#) to sign up and start enjoying the benefits!
- A ready a member? [Click here](#) to sign in

As an administrator, log in with your credentials. Input your username and password.

This will be the landing page.

Administrator Consys



Home



About

Users

Groups



Sign out



All Content

### Profile

Name	Administrator Consys	<a href="#">edit</a>
password		<a href="#">change</a>
Account		<a href="#">Account</a>

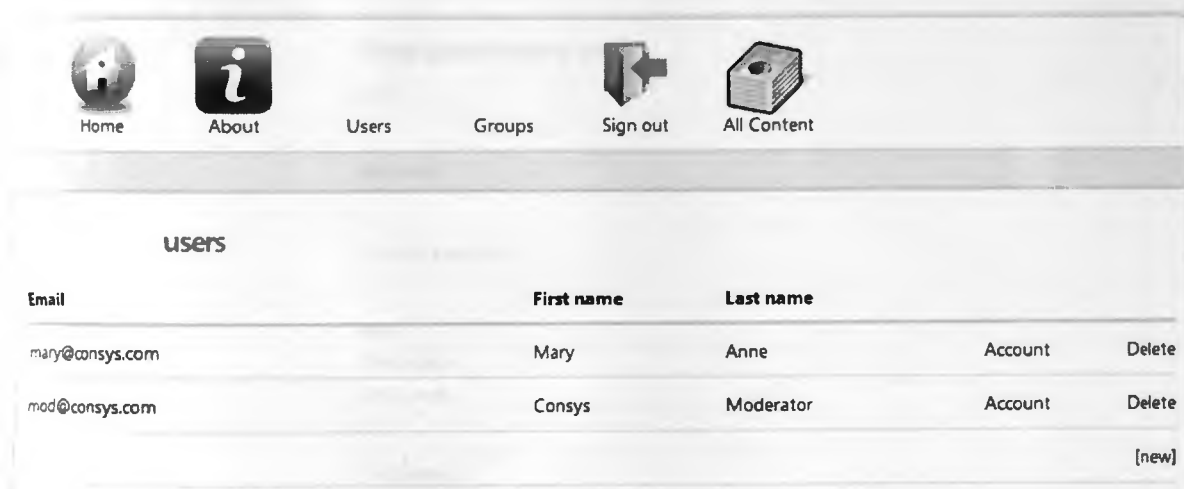
You can edit your profile by clicking on the 'edit' link.

To create and manage users and groups and perform other functional activities you click on one of the links as shown in the following screenshot and this will give you access to various sections of the administration interface.

## Users

The Users link does just what you might guess by its title. It manages users. You can create new users; delete existing ones, change passwords and so on. Lastly, but perhaps most importantly, you can change the user's user group, giving the user different access levels and different abilities in the portal.

Click on the 'Users' link and it will take you to the page illustrated below.



The screenshot shows a navigation bar with icons for Home, About, Users, Groups, Sign out, and All Content. Below the navigation bar is a section titled 'users' containing a table with columns for Email, First name, Last name, Account, and Delete. The table lists two users: mary@consys.com (Mary Anne) and mod@consys.com (Consys Moderator). A '[new]' link is visible at the bottom right of the table.

Email	First name	Last name	Account	Delete
mary@consys.com	Mary	Anne	Account	Delete
mod@consys.com	Consys	Moderator	Account	Delete

[new]

To add a new user, click on the 'new' link and the page shown below will be displayed which will allow you to enter the credentials of the new user and assign them to the appropriate group.

### User details and parameters

You will see different fields where you can fill in or edit information of the user. These are:

**E-Mail:** The e-mail address from the user is displayed here. When a user wants to log in, he has to fill in this email.

**Password:** Fill in a (new) password.

**Confirm password:** Fill in the password from the field above again, to verify it. This field is required when you filled in the Password field.

**Role:** The user's Group. The following Groups are available:

- **User:** Normal visitors who register at the site. Can view Menu Items and can submit articles.
- **Moderator:** Can review and approve articles to be published on the content sharing platform.



Home



About

Users

Groups



Sign out



All Content

### New government user

Email

Password:

Confirm password

Role:

Moderator

First name

Last name

save

### Adding New Groups

As an Administrator, you can be able to add groups with different rights and assign new users to new groups which you create. You can also be able to edit existing groups as is necessary.



Home



About

Users

Groups



Sign out



All Content

## Groups

Group name	Group description		
User	Application user	edit	delete
Administrator	Application Administrators	edit	delete
Mod	Moderator	edit	delete
IT. Staff	Application Maintainers	edit	delete
			[new]

To add new groups, click on the new link and the page shown below will appear.



Home



About

Users

Groups



Sign out



All Content

## New group

Group name

Category description

Enter the details of the new group which you wish to add and then click on Save.

To edit an existing group, click on the 'edit' link next to the group which you wish to change.



Home



About



Users



Groups



Sign out



All Content

## Edit group

### Editing group

Group name

Administrator

Category description

Application Administrators

save

## Viewing Content

As an administrator, you can also be able to view all the content that has been uploaded on the portal by clicking on the 'All Content' link.



Home



About

Users

Groups



Sign out



All Content

Filter:

Pick a Ministry



Filter



### Creative Commons Licenses -

These are CC licenses.



### Malaysia Government -

This a diagram on how Malaysian government is.



### Sign Off Sheet -

This is the design stage sign off sheet



### E-Government Strategy -

This is the Kenyan e- Government Strategy



### MSc Guidelines -

These are the MSc guidelines.



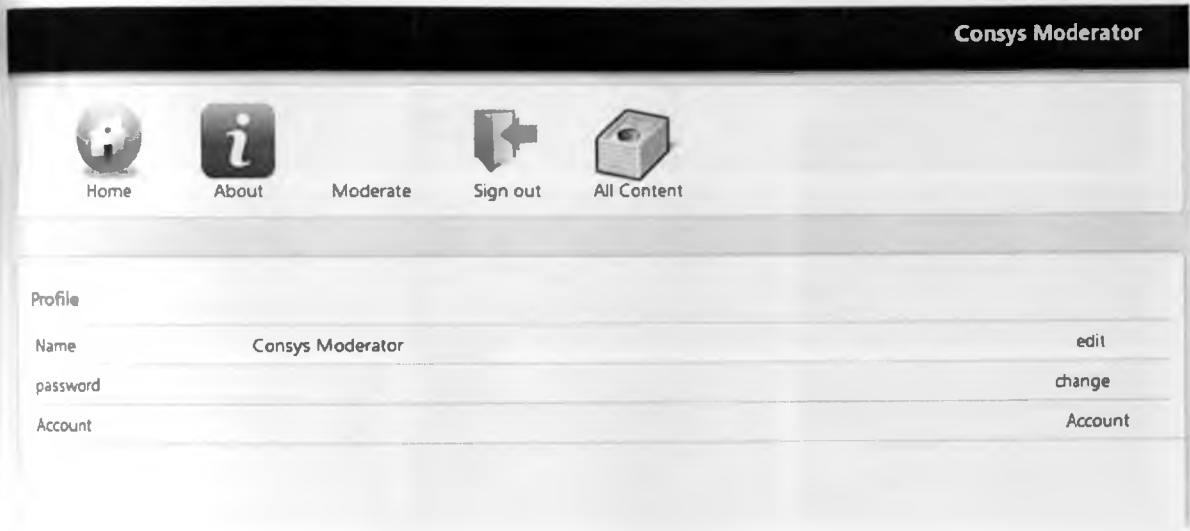
### Transmission Media -

Different transmission media

You can also be able to filter the content which has been uploaded by a specific Ministry or search for specific content.

## How to Access the Portal as a Moderator

Input your Username and Password in their respective fields and click the Login button. The Home page is displayed.



You can edit your profile or change your password from here.

### Moderation

The role of the moderator is to review all the content that has been uploaded on the content sharing portal and approve it for general viewing on the portal.

In order to moderate the content on the portal, click on the 'Moderate' link.

Once you click on the link, the following screen will appear.



To moderate content, click on the 'Approve' link on the specific content that you would wish to approve for general viewing on the site. You can be able to reverse the approval at any time by clicking on the 'Reverse Approval' link for any content.

Approved content will not be highlighted but all Unapproved content will be highlighted as depicted above.


You can also be able to filter the content per ministry as shown below for moderation





### How to Access the Portal as a User

In order to upload your content as a user, Sign up on the content sharing platform as shown below by filling in your details.



**Sign up below...**

Email

Password:

Confirm password

First name

Last name

### Getting around the Portal

Navigating the portal interface from anywhere within the content sharing site is accomplished by a global menu bar to guide you through the various areas of the portal. Located at the top of every screen (which is why we refer to it as global) the links are named according to the areas they lead to. The global menu bar is as shown in the following screenshot:



## Home

This takes you back to the Home page which contains links to all of the key site areas, which include the following:

- Top Rated Content
- Versioning information
- Licensing details among others.

## About

This links takes you to the section where you can get information on what Consys is all about and what to expect from it.

## My Statistics

This takes you to a page where you can see a list of all the content that you have uploaded on the Consys content sharing portal, the license with which you licensed your work as well as the number of times your content has been downloaded.

Files			
Name	License	Downloads	
CISA	Attribution-NonCommercial-ShareAlike CC BY-NC-SA	0	versions
Malaysia Government	Attribution-NonCommercial CC BY-NC	1	versions

## My Content

All the information that you have uploaded on the site and has been approved by the moderator can be viewed through here.



### Files

Name	License		
Peeling back the Mask	Attribution-ShareAlike CC BY-SA	versions	delete
Salty Reg Keys	Attribution-ShareAlike CC BY-SA	versions	delete
Salty Killer	Attribution-NoDerivs CC BY-ND	versions	delete
Malaysia Government	Attribution-NonCommercial CC BY-NC	versions	delete
			[new]

## All Content

This is the link via which you can see all the content that has been published on the portal, not just by you but by other users as well. It will appear as shown below.

Filter: Search

Pick a Ministry

Filter

**i** Creative Commons Licenses -

These are CC licenses.

**i** CISA -

This is the CISA Exam Guide

**i** Malaysia Government -

This a diagram on how Malaysian government is.

**i** Sign Off Sheet -

This is the design stage sign off sheet

**i** E-Government Strategy -

This is the Kenyan e-Government Strategy

**i** MSc Guidelines -

These are the MSc guidelines.

**i** Transmission Media -

Different transmission media

**i** E-Govt Strategy -

Kenyan E-Gov Strategy

**i** Peeling back the Mask -

Peeling back the Mask by Miguna Miguna

**i** Salty Killer -

Salty Killer

**i** Salty Reg Keys -

Salty Reg Keys

## Help

A comprehensive list of terms and key phrases are contained within a database of information. This is the site-wide Help database and contains broad information for all areas of the site. It is available on most of the pages on the content sharing portal.



**Content** Here you are offered the facility of submitting your high quality original articles and content licensed using any one of the various Creative Commons licenses for use by the Kenyan government which will increase exposure of your work as well as have your work have a greater impact in the society at large.



**Statistics** Here you can monitor download statistics generated by your files and content that you have uploaded on the portal.



**Versions** This is where the provision for adding versions to your work is permitted. Therefore you can use an incrementing value to indicate various versions of the same file you might want to upload on the portal. An old copy of your file will still be maintained in the portal. You also have the flexibility of deleting any of your files in case you deem it obsolete or have a revised version which supersedes the older one.



**Categories** These are the various categories under which you can publish your work on the portal. In case a category that you wish to publish your work under is not listed here, you can suggest it to the Portal Administrator who will review it and then add it.



**Licences** The Creative Commons copyright licenses and tools forge a balance inside the traditional "all rights reserved" setting that copyright law creates. These tools give everyone from individual creators to large companies and institutions a simple, standardized way to grant copyright permissions to their creative work. The combination of these tools and our users is a vast and growing digital commons, a pool of content that can be copied, distributed, edited, remixed, and built upon, all within the boundaries of copyright law. For more information, [Click Here](#)

Close

Click on close once you finish and you will be taken back to the page you were on before you clicked on 'help'.

## Uploading New Content

To upload new content on the portal, click on the 'My content' link. Then click on the 'New link' as shown below.

### Files

Name	License		
Peeling back the Mask	Attribution-ShareAlike CC BY-SA	versions	delete
Salty Reg Keys	Attribution-ShareAlike CC BY-SA	versions	delete
Salty Killer	Attribution-NoDerivs CC BY-ND	versions	delete
Malaysia Government	Attribution-NonCommercial CC BY-NC	versions	delete

[new]

This will open the page shown below. Select the file or zipped folder containing your content or application from your computer, fill in the title in the title field, type in a short description of your article or application. Select the Version, category (ministry) and last but not least the CC license with which you are sharing your work. Then click on save.



**New file**

Title

Description

Versions

--select version--

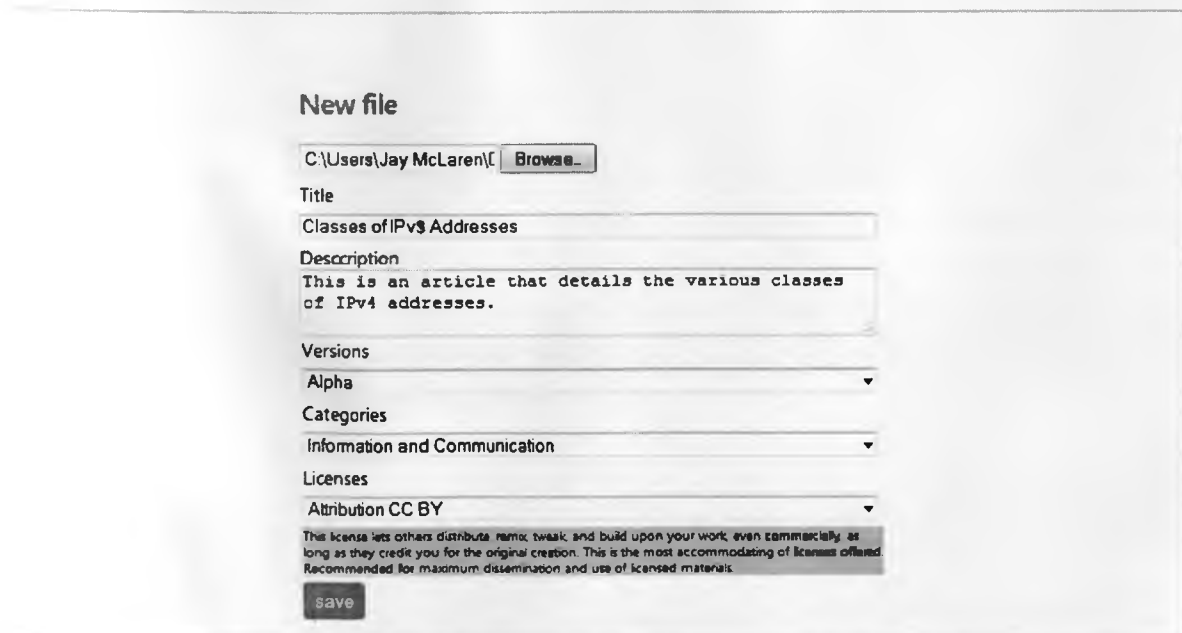
Categories

--select category--

Licenses

--select license--

A filled in upload section would look as follows.



**New file**

C:\Users\Jay McLaren\

Title

Classes of IPv4 Addresses

Description

This is an article that details the various classes of IPv4 addresses.

Versions

Alpha

Categories

Information and Communication

Licenses

Attribution CC BY

This license lets others distribute, remix, tweak and build upon your work, even commercially, as long as they credit you for the original creation. This is the most accommodating of licenses offered. Recommended for maximum dissemination and use of licensed materials.

Once you click on Save, your content will be saved in the database. But it can only be visible on the website once the moderator has reviewed and approved it.

But you can keep track of your uploaded content by using the Statistics section where you can be able to view all the content that you have uploaded on the site and how many times your work has been downloaded.

Files			
Name	License	Downloads	
CISA	Attribution-NonCommercial-ShareAlike CC BY-NC-SA	0	versions
Malaysia Government	Attribution-NonCommercial CC BY-NC	1	versions
Open Knowledge Festival	Attribution CC BY	0	versions
Salinity Killer	Attribution-NoDerivs CC BY-ND	1	versions
Salinity Reg Keys	Attribution-ShareAlike CC BY-SA	0	versions

## Competition

### How to enter the competition

The content sharing platform also offers you a chance to win fabulous prizes by participating in uploading open content as well as apps based on OSS. All you need to do is follow the instructions on the home page as highlighted below.

### Welcome

Consys is a creative commons licensed content sharing site, where you can publish and share most media types with the Kenyan Government. You have practically unlimited flexibility in choosing what you want to share, as long as it is your original work or work licensed under a Creative commons license which allows for free remixing and distribution of that work but with the proper attribution to the original content creator.

### Join the competition

#### How It Works

#### Upload Item

#### Vote

By uploading content to this site you stand a chance to win fabulous prizes.

#### Steps

1. Register an account on the site
2. Upload your open source content using your account.
3. Ask your friends to rate your uploaded content.
4. Content with highest total vote wins

*Winners will be picked on the 1st day of every month*

- [Click here](#) to sign up and start enjoying the benefits!
- A ready a member? [Click here](#) to sign in

## Leaderboard

The Leader Board displays the top six rankings of the leading content contributors to the portal. Ranks are based on the points that users achieve in each of their uploaded items that have been rated. That is, the highest position on the Leader Board at any particular time is given to the user whose app or content has achieved the highest points. The Leader Board only keeps track of the six highest scores. Therefore, earning a spot on the Leader Board is a great achievement, and the top three users are awarded for doing so on the 1st of every month.



## Welcome

Consys is a creative commons licensed content sharing site, where you can publish and share most media types with the Kenyan Government. You have practically unlimited flexibility in choosing what you want to share, as long as it is your original work or work licensed under a Creative commons license which allows for free remixing and distribution of that work but with the proper attribution to the original content creator.

## Join the competition

How It Works		Leader Board	
Name	File Title	File Description	Total Points
1 Ann Consys	MSc Guidelines	These are the MSc guidelines.	40
2 Janet Maranga	Creative Commons Licenses	These are CC licenses.	38
3 Well Ness	CISA	This is the CISA Exam Guide	36
4 Janet King	Malaysia Government	This a diagram on how Malaysian government is.	30
5 Anna Lora	Transmission Media	Different transmission media	25
6 Janet Maranga	Chart	This is a chart	23

## Prizes


The prizes that users will compete for are:

**Top prize:** A Samsung S3 phone, **Runners up:** A Samsung Galaxy Tab 2, **Third Prize:** A Samsung Galaxy Young phone. These are great prizes, so the more your friends rate your app or content, the higher your chances of winning.

The prizes are as shown below.

Prizes


**1st Prize**



**Key Features**

1. HD Super AMOLED, 4.8" , 1280 x 720, 306 PPI with Gorilla Glass 2.0
2. Android Ice Cream Sandwich 4.0
3. 8 megapixels with LED, autofocus, records video
4. 3G, HSPA+, 4G LTE (and WiFi)


**2nd Prize**



**Key Features**

1. 1GB (RAM), 16/32GB (ROM) + microSD (up to 32GB)
2. 1.2GHz dual-core processor
3. Built-in 4,000mAh battery
4. Back camera : HD(720p) Video recording, 3MP auto-focus camera with LED flash)

**3rd Prize**



**Key Features**

1. Androidâ., v2.3.5 (Gingerbread)
2. 832MHz Processor
3. TOUCHWIZ v3.0 User Interface (up to 7 pages widget desktop)
4. User Memory: 160MB