UNIVERSITY OF OSLO
Department of Informatics

HEALTH INFORMATION SYSTEMS INTEGRATION

A Data Warehouse Architecture Model for the Ministry of Health in Mozambique

Teotónio G. S. Fumo

Master Thesis

2nd May 2003

HEALTH INFORMATION SYSTEMS INTEGRATION:
A Data Warehouse Architecture Model for the Ministry of Health in Mozambique

By

Teotónio Glória dos Santos Fumo

Supervised by
Dr. Jens Kaasbøll

A dissertation submitted in partial fulfillment for the degree of Master of Science in Information Systems (IS).

Department of Informatics
University of Oslo, Norway

May 2003

Declaration

I, Teotónio Glória dos Santos Fumo, declare that this thesis embodies only my original work, except where acknowledgement indicates otherwise, and that no part of it has been, or is being submitted for a degree at this or any other University.

Signature…………………………………

Place and Date……………………………

Dedicated to,

My parents Teófilo Fumo and Serafina Matavel, who brought me up and gave me the value of education, my sisters Natércia Fumo and Maria Fumo, my brother Eládio Fumo and my girlfriend Guilhermina Banze, I will always love you.

Does not exists a difficulty
that resists to the
attack of the persistence.

Unknown author

Abstract

This thesis presents a study about a data warehouse architecture model for the Ministry of Health (MoH) in Mozambique. The study combines two areas: health care and information systems areas. It was conducted using two research methodologies: system development and case study of the Ministry of Health. The model aims at integrating data from different sources in the Ministry of Health. The fieldwork was carried out in the southern area of Mozambique, in the Ministry of Health in Maputo province and in two districts of Gaza province: Manjacaze and Chibuto. The choice of these sites for the fieldwork was based on the implementation of the District Health Information System project.

This research is a contribution to the current process of data and information integration in the Ministry of Health. Integration of information and data warehouse (DW) technology are tools that offer resources to obtain managerial information needed to establish control over management process. Data warehousing aims at providing, managing and exploiting a set of integrated data for decision support within an organization.

The research methods applied to this study include interviews, observations, questionnaires, document analysis and analysis of existing systems. During the development of the model I applied part of the data warehouse life cycle. Taking into account the system development phases, I covered the first three phases of the data warehouse life cycle, in order to document the existing legacy systems, to create a model of data warehouse and to clean the data. As a result, I present a data warehouse architecture model for the Ministry of Health. The proposed model integrates different heterogeneous systems and provides integrated information for health workers (administrative personnel and managers).

The empirical findings proved that for the data warehouse project the Ministry of Health will need to put more effort on the data quality control, because the quality of data influence the decision-making process.

Keywords: Health Information Systems, Database, Data Warehouse, Data Mart, Data Integration, and Legacy Systems.

Acknowledgements

First, to my Lord.

Secondly to many people who contributed to and helped me shape this thesis, I owe them all a debt of gratitude. Those listed below deserve special mention and thanks.

My sincere thanks to Dr. Jens Kaasbøll, my supervisor, who supervised my work from the very beginning – from collection of vague ideas to the completion of the project as it stands now, for his guidance, advice, comments, suggestions and criticisms during this work.

My thanks to my colleagues from the Masters in Public Health course for their support. Special thanks go to my colleagues at the Eduardo Mondlane University and University of Oslo, Gertrudes Macueve, Juma Lungo, Faraja Mukama, Emílio Mosse, Virgílio Culpa Leopoldo Nhampossa, Baltazar Chilundo and Carlos Mavimbe, for their support.

I would also like to express my gratitude to NORAD (Norwegian Agency for Development Cooperation) for the financial support during my study and to the coordinator of the Masters Course in Information Systems in Mozambique, Dr. Esselina Macome, for her trust and availability.

To all my teachers during the Masters Course: Dr. Judith Gregory, Dr. Sundeep Sahay, Dr. Jorn Braa, Dr. Knut Roland, Dr. Arthur Heywood, Dr. Yussuf Adam, Dr. Humberto Mukingue, for their support, trust and guidance during the different courses.

To my dearest family and friends for their understanding and unconditional support. My special thanks to my girlfriend, for her patience and good will. Last but not least, I would like to thank all individuals who directly or indirectly helped me in this work.

Table of Contents

List of Figures

List of Tables

List of Acronyms

AIDS – *Acquired Immune Deficiency Syndrome*

BCG – *Bacillus Calmette-Guerin vaccine*

BEM – *Boletim Epidemiológico Mensal / Monthly Epidemiological Bulletin*

BES – *Boletim Epidemiológico Semanal / Weekly Epidemiological Bulletin*

BESC – *Boletim Epidemiológico Semanal Computarizado*

BI – *Business Intelligence*

CBES – *Computerized Weekly Epidemiological Bulletin*

CSIS – *Computerized Sistema de Informação de Saúde*

CO – *Central Offices*

COBOL – *Common Business Oriented Language*

CRM – *Customer Relationship Management*

CS – *Centro de Saúde / Health Centre*

DB – *Database*

DDH – *District Directorate Of Health*

DDS – *Direcção Distrital de Saúde* (*District Department of Health*)

DHIS – *District Health Information System*

DM – *Data Mart*

DM – *Data Mart*

DMi – *Data Mining*

DPC – *Direcção de Planeamento e Cooperação / Planning and Cooperation Directorate*

DPS – *Direcção Provincial de Saúde* (*Provincial Department of Health*)

DPT – *Diphtheria, Tetanus and Pertussis vaccine*

DSS – *Decision Support Systems*

DW – *Data Warehouse*

DWH – *Data Warehousing*

DWS – *Data Warehouse System*

EIS – *Executive Information System*

EPI – *Expanded Program of Vaccination*

F1 – *Keyboard key that open the HELP application*

FRELIMO – *Frente de Libertação de Moçambique (Mozambique Liberation Front)*

GACOPI – *Gabinete de Cooperação de Projectos de Investimento / Investment Project Coordination Office*

GDP – *Gross Domestic Product*

HF – *Health Facilities*

HID – *Health Information Department*

HIS – *Health Information System*

HISP – *Health Information System Project*

IBM – *International Business Machine*[â]

IHS – *Integrated Health Systems*

INDER – *Instituto de Desenvolvimento Rural / Rural Development Institute*

INE – *Instituto Nacional de Estatísticas / National Institute of Statistics*

IS – *Information Systems*

LAN – *Local Area Network*

LS – *Legacy Systems*

MCH – *Mother and Child Health*

MoH – *Ministry of Health*

MOLAP – *Online Analytical Process*

MTS – *Meticals*

NE – *Nucleo de Estatística / Statistics Team*

NEP – *Núcleo de Estatística Provincial / Provincial Statistics Team*

NHI – *Nucleus of Health Information*

NHIS – *National Health Information System*

NHS – *National Health Services*

NORAD – *Norwegian Agency For Development Cooperation*

OBDC – *Open Database Connectivity*

ODS – *Operational Database Systems*

OLAP – *Online Analytical Process*

OLE – *Object Link Embedded*

OLTP – *Online Transaction Process*

PDH – *Provincial Directorate of Health*

PESS – *Plano Estratégico do Sector de Saúde/Strategic Plan for the Health Sector*

PF – *Planeamento Familiar / Family Planning*

PHC – *Primary Healthcare*

PS – *Posto de Saúde / Health Post*

RENAMO – *Resistência Nacional Moçambicana (Mozambique National Resistance)*

ROLAP – *Relational Online Analytical Process*

SCM – *Supply Chain Management*

SDLC – *Software Development Life Cycle*

SIMP – *Sistema Integrado de Planificação / Planning Integrated System*

SIS – *Sistema de Informação de Saúde / Health Information System*

SMI – *Saúde Materno Infantil / Mother and Child Health*

SQL – *Structured Query Language*

STD – *Sexually Transmitted Diseases*

TB – *Tuberculosis*

UNHCR – *United Nations High Commissariat for Refugees*

UNICEF – *United Nations Children's Fund*

USAID – *United States Agency for International Development*

VAT – *Vacina Anti-Tetanica (Anti-Tetanus Vaccine)*

WB – *World Bank*

WHO – *World Health Organization*

Chapter 1

Introduction

This chapter presents an introduction to the topic of data warehouse development, the problem statement, the main objective of the research, my research questions, the target group, the personal motivation and research method in brief.

1.1 Introduction

At independence (1975), the new government launched an ambitious programme in the health sector (Ministério da Saúde, 1979). All existing health assets were nationalised within a unified health system, whose dominant approach was Primary Health Care (PHC). The National Health Service (NHS) was expected to provide health care to all citizens, under a central planning system. Health expenditures expanded, to fuel the expansion of the peripheral network and of Primary Health Care services. The health workforce increased significantly, through the creation of new Primary Health Care -oriented professionals and accelerated training activities. An innovative and highly successful policy for medicines (Barker, 1983) was introduced. The coverage of comprehensive, integrated basic services increased (Walt and Melamed, 1983).

In the Ministry of Health, the National Directorate of Health is responsible for managing health programmes, community health, medical care and epidemiology. The Planning Directorate has the responsibility for long-term planning, statistical collection and analysis of health data. This directorate is subdivided into various departments, of which the Health Information Department (HID) is one of them. This directorate manages the health information system through its health information directorate (Macome, 1992).

Several systems from different departments compose the health information system. These systems are operational database systems. Operational database systems (ODS) are data repositories manipulated by information systems inside the organizations. These databases play an important role in organizations, and are composed of millions of lines of source code that are frequently written in third generation languages like Dbase or COBOL (Common Business Oriented Language), and most of them are not documented. Operational database systems are found in most public organizations in Mozambique, such as the Ministry of Health.

Operational database systems sometimes cause serious problems for the Ministry of Health due to the fact that they are essential to their daily activities and were conceived with an approach to transactions processing rather than to analytical processing, and they are not subject oriented. This approach of database construction does not allow managers to obtain important information about historical data in order to make decisions.

Due to organizational or operational constraints, these data sources do not generally lend themselves to being fully replicated or consolidated under a single database. Information integration technology enables integrated, real-time access to traditional and emerging data sources, transforms information to meet the needs of managers and manages data placement for performance, currency and availability (Hayes, 2002).

The topic of data warehousing comprises architectures, algorithms, models, tools, organizational and management issues for integrating data from several operational systems in order to provide information for decision support, e.g., using data mining (DMi[1]) techniques or on-line analytical processing (OLAP[2]) tools. Thus, in contrast to operational database systems which contain detailed, atomic and current data accessed by on-line transactional processing (OLTP[3]) applications, data warehousing technology aims at providing integrated, consolidated and historical data (Vavouras, 2000). In the Ministry

[1] The process of and utilizing the results of data exploration to adjust or enhance management strategies.
[2] Processing that supports the analysis of organizational trends and projections.
[3] Processing that supports the daily organizational operations.

of Health "*islands of systems*" are found, so the information is not integrated and the analysis process is difficult.

Integrating data and information is a requirement for modern management. The challenge becomes extracting and integrating data and information, ensuring its accuracy, and making it easily available to those who need it. One approach to solving this problem – building and maintaining a data repository – focuses on collecting and aggregating information over time. The data repositories contain data that will migrate to the data warehouse.

In the healthcare area, health systems integration is an idea that is currently popular in the world as a means to provide better effectiveness and efficiency in the healthcare system. Health systems integration exists under a variety of names in different parts of the world, such as clinical systems integration, medical and health systems or health data integration.

Creating integration architecture is recognized as a competitive differentiator that can enable businesses to respond more quickly to a changing environment and new opportunities. Complete integration architecture will necessarily include multiple technology approaches to match the integration issue (Hayes, 2002).

> "Providing a unified view of the company's information assets to underpin business intelligence is what data and information integration is all about".
>
> (Steensboe, 2002)

In today's health sector, the demand to integrate data and information across and beyond the organization is a competitive mandate. In the last ten years, the Government of Mozambique approved a law that allows privates to operate in the health sector, so the health sector may have to become more competitive.

Hayes (2002) argues that, initiatives, such as customer relationship management (CRM), supply chain management (SCM) and business intelligence (BI) are based on successfully integrating data and information from both structured and unstructured data sources.

In Mozambique, the Ministry of Health currently has several systems in different departments that are not integrated. From the research results, different Managers recognized that relevant decisions could only be taken based on data coming from different existing systems. Managers interviewed during the research argued that decision makers need to integrate different information sources in order to make correct decisions.

This research is based on a field research in Gaza Province. The choice of this province is due to the fact that the Health Information System Project (HISP) is conducting the implementation of the District Health Information System (DHIS) in the province, and as member of the project, I was able to easily collect data during our trips to Gaza.

1.2 Problem Statement and Research Questions

An integrated information system is essential for monitoring and assessing performance in the entire healthcare sector in Mozambique. The system should be integrated to combine data and information from different subsections inside the Ministry of Health in Mozambique. In the context of addressing the research problem, the topic is:

- What is the current status concerning integration of data in the health sector in Mozambique?

My research questions are:

1. How to collect or extract health data from heterogeneous sources?
2. How to clean and transform heterogeneous health data?
3. How to integrate heterogeneous health data?

4. What type of methodological approach should be used to design a data warehouse model for the health sector?
5. How can data quality influence the decision-making process

1.3 Target Groups of Interest

The intended audiences for the study findings are:

- Technical staff working at information technology (IT) departments in any healthcare organization;
- Anyone interested in integration of heterogeneous healthcare information systems.
- Experts involved in systems analysis, design and implementation;

The content of the thesis is of general interest and written in a commonly understandable way, which also makes the material accessible for anyone working in the computer science area.

1.4 Personal Motivation

According to Maslow (1954), we need to be motivated to carry out a research. I am motivated to carry out this research because I am interested in learning more about health information systems. I am particularly interested in database systems, data warehousing and data integration in the healthcare sector, focusing on techniques to make integration easier and more reliable. My aim is to present a research study that can be used as reference for new health information Masters students in Mozambique.

Why an architecture model?

Although many data warehouses have already been built, there is no common methodology that supports database system administrators in designing. The problem with architecture models for data warehouses is that practice has preceded research in this area and continues to do so. Consequently, the task of providing an abstract model of architecture becomes more difficult. Formally, an architecture model corresponds to the schema structure of the meta-database that controls the commonly distributed and heterogeneous set of data warehouse components and, therefore, is the essential starting point for design.

The proposed architecture model should not restrain from any single aspect of the management. As a matter of fact, a comprehensive model will address managerial, organizational or social aspects. This means looking at managerial aspects to ensure that the management style and philosophies are compatible with the model. The organizational aspects should be analyzed to ensure that different processes, policies, and procedures support the model. The social aspects should also be addressed.

1.5 The Objective and the Research Method in brief

The main objective of this research was to propose a data warehouse architecture model to integrate data from different systems at the Ministry of Health in Mozambique.

Based on the main objective, the following activities were carried out during the research:

- Study of the data warehouse technology as a way to integrate heterogeneous health systems at the Ministry of Health in Mozambique;
- Study of the information flow of the health information system in Mozambique based on the information cycle proposed by Heywood and Rohde (2002); and
- Design of a representative data warehouse architecture model for the integration of different health heterogeneous systems.

The research approach to this thesis was based on qualitative research methods such as interviews, observations, questionnaires, documents analysis, literature review and analysis of existing systems.

I started my research with a broad study of the relevant literature about data warehouse development, health information systems and integration of systems.

While in Norway, I participated in a data warehouse discussion forum group on the Internet, from September to November 2003. I also discussed the topic with a group of five Masters students from Universidade Católica de Brasília in Brazil, working on data warehouse. These discussions helped me understand the concepts about data warehousing.

My primary access for my fieldwork research in Mozambique was within the Provincial Directorate of Health in Gaza. Three sites were selected: Chibuto and Manjacaze districts in Gaza Province and the Ministry of Health, in Maputo Province.

Health workers such as doctors, nurses, administrative personnel and managers were interviewed in the three sites. I conducted some observations in the districts of Chibuto and Manjacaze. Observations included the daily work activities of health workers. I was able to analyze documents and existing manuals and automated systems at all – national, provincial and district – levels.

1.6 Structure of the Report

This thesis is organized in seven chapters. This chapter (chapter one) deals with the introduction to the study and focuses on the problem statement, research problem and objectives of the study.

Chapter two describes the research method approach adopted and the different methods used in the research. Chapter three includes literature review and the theoretical framework for the thesis. Chapter three reviews research and studies carried out in the past and presents the theoretical base for analysis and discussion.

Chapter four presents a description of the Mozambican Health Information System. A brief description of the Ministry of Health is presented. Chapter five presents the empirical findings, as a result of field research at different levels of the National Health Information System with focus on information flow. Chapter six presents the proposed data warehouse architecture model and a description of different components of the model.

Chapter seven presents the discussion, conclusions and recommendations of the study. It discusses the results obtained from the data analysis. In this chapter the research questions are discussed. Chapter seven concludes with the summary of findings and recommendations.

Chapter 2

Research Methods

This chapter presents the methodology used in the research. Qualitative methods and research techniques are briefly described. There were three major goals for this research. First, to study the information flow of the health information system in Mozambique based on the information cycle proposed by Heywood and Rohde (2002). Secondly, to study the data warehouse technology as a way to integrate heterogeneous systems at Ministry of Health in Mozambique. Finally, to design a representative data warehouse architecture model to integrate the different heterogeneous health systems.

2.1 Research Problem

> Everyone has their own data repository so no one has a clear idea of the corporation information and knowledge – the data is effectively not integrated. (Barker, 2000)

The research problem can be formulated as follows. Each department at the Ministry of Health has their own legacy system running, so it is difficult to know the amount of information that flows inside the organization. There is no integration of all data that flows inside the Ministry of Health. The problem is that, when the decision process happens, managers have incomplete and non-integrated information to support decisions taken, because systems inside the Ministry of Health are not integrated. Due to the operational nature, such systems are inadequate to the production of temporary reports and of difficult use for the managers of the organization. A review of the literature showed that data gathering process from different systems takes time and that managers prefer to make decisions without any information support.

2.2 Research Settings

The location of this project was the Ministry of Health in Maputo city and in two districts in Gaza province, Manjacaze and Chibuto. The selection of Gaza province was based on the fact that I am a member of the Health Information System Project (HISP) and the project is currently implementing the District Health Information System (DHIS) in Gaza province.

"Health Information Systems Project (HISP) was initiated in 1994 by researchers from Norway and the Universities of Western Cape and Cape Town. With its inspiration drawn from the "Scandinavian action research" tradition coupled with the fervor of the South African freedom struggle, the aim of HISP was primarily to empower the new and emerging local health management structures and health workers through improved and locally based information systems. Nearly a decade later, the research in South Africa is still going strong as HISP has officially taken on the status of the "national standard in Health Information Systems" and through a series of various planned and unplanned events, HISP processes have been initiated in various other countries including Mozambique (since 1998), other countries in Africa (like Malawi and Tanzania, since 2000 and 2001), India (since 2000) and Cuba and Mongolia (since 2002). In this way, HISP became an important actor in the national as well as local processes of standardization of health data" (Braa *et al*. 2002).

Figure 2.1 shows the location of Manjacaze and Chibuto districts at Gaza province.

Map of Gaza Province

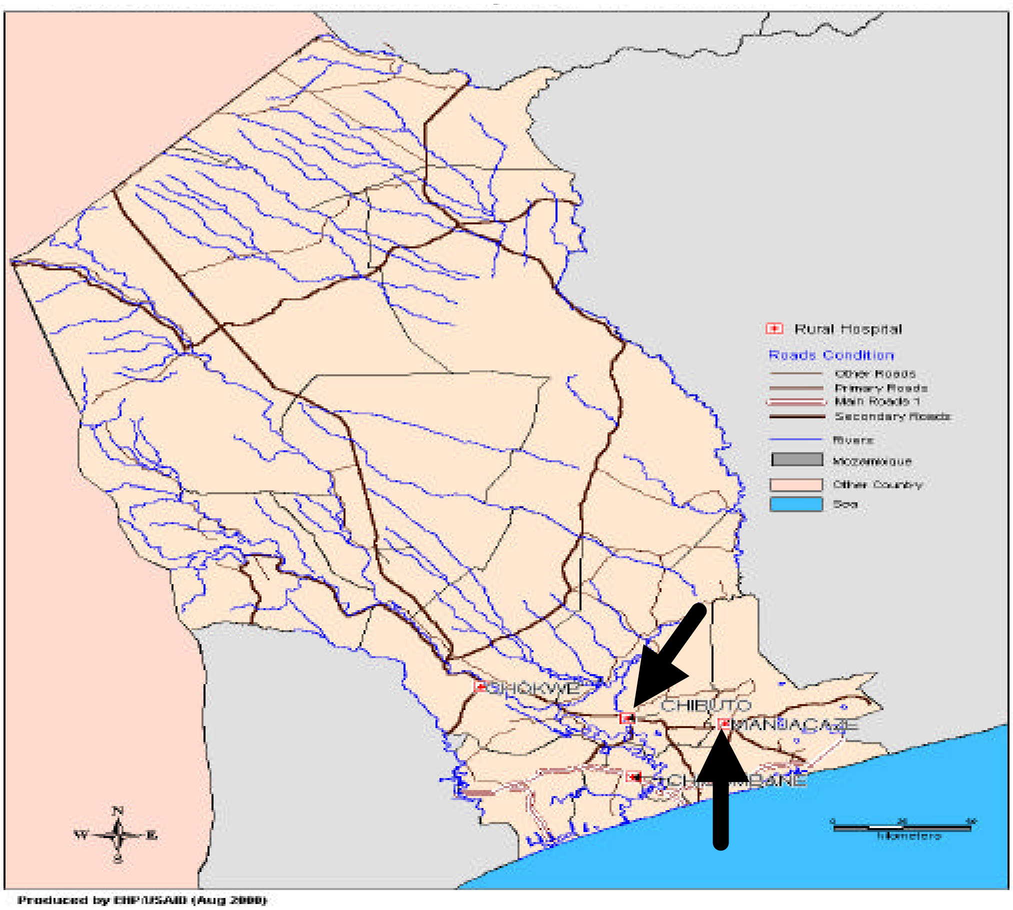


Figure 2.1: Map of Gaza Province (Source: Census 1997/INE).

Key:

→ - Visited districts in Gaza (Chibuto and Manjacaze).

The research reported in this thesis started in 2001 and it was conducted at the Ministry of Health and in Gaza province (Manjacaze and Chibuto districts) in Mozambique. I have developed a data warehouse architecture model and described all its components.

The research followed a qualitative case study approach. By the term qualitative research, Strauss and Corbin (1990) mean, "any kind of research that produces findings not arrived at by means of statistical procedures or other means of quantification". To understand the context and the process it was necessary to record the answers and explanations that the different stakeholders gave to different questions. Hence my decision to adopt a qualitative research method.

Qualitative methods, also described as a naturalistic approach (Silverman, 1993; Savenye and Robinson, 1996), have been dominant in investigating the system integration. The research methods adopted for conducting this research included, interviews, observations, questionnaires, documents analysis, literature review and analysis of existing systems. The key elements of the research were the case study approach and collection and analysis of data from primary and secondary sources. Primary sources of data included interviews, observations and questionnaires and secondary sources of data included analysis of reports and existing systems.

2.2.1 The Ministry of Health

The Ministry of Health was used as the main point of data collection for my research. I conducted five interviews with personnel from different backgrounds. One nurse, one physician and three consultants were interviewed in the Ministry of Health. The interviews with the nurse and the physician were informal interviews while the interviews with the consultants were more formal. The consultants are epidemiologists who have been working in the information system for a long time. I had several meetings where I discussed important points about health information system with consultants. The Ministry of Health also organized all trips to Gaza province, Manjacaze and Chibuto districts. The research in the Ministry of Health helped me gather essential data for my thesis and to understand the healthcare sector in Mozambique related to information systems.

2.2.2 Gaza Province

In Gaza, at the Provincial Directorate of Health, I conducted interviews with four individuals. The interviewees included staff working at the Nucleus of Health

Information, the Manager of Provincial Directorate of Health and some staff responsible for different health programmes at province level. Contact was first established with these personnel because they are the responsible for all programmes running at district level.

Figure 2.2: HISP Team visiting the Provincial Directorate of Health in Gaza.

Figure 2.2 shows the head of the Weekly Epidemiological Bulletin (BES) explaining to the Health Information Systems Project team members how the system works.

2.2.3 Manjacaze District

In Manjacaze, I conducted interviews with six individuals (See Annex C). Four out of six of the interviewees work at the Nucleus of Health Information and the rest include administrative personnel. These interviews were conducted to understand the data collection, processing, analysis and use processes at the district level and to gather explanations about some reports, such as supervision reports, health facilities reports and district reports. Observations on how health workers deal with daily tasks, such as data collection and analysis, and meetings were conducted. I spent approximately three weeks in the observations. I also participated in one meeting of the Nucleus of Health Information.

Figure 2.3: Students of Masters in Public Health visiting Manjacaze.

Figure 2.3 shows students receiving nurse information about the patient admission process at the Manjacaze Rural Hospital.

During my visit to Manjacaze district one of my tasks was to install the District Health Information System in the computers available in the District Directorate of Health. I managed to install the software and trained health workers on how to operate and use the software.

2.2.4 Chibuto District

In Chibuto, I conducted interviews with six individuals (See Annex C), the same number of health workers interviewed in Manjacaze. Three out of six of the interviewees work in the Nucleus of Health Information and the rest were managers and administrative personnel. In Chibuto, I carried out the same work performed in Manjacaze, software installation and training for health workers.

Figure 2.4: Meeting with District Directorate of Health Manager in Chibuto.

Observations of health workers daily work were conducted. Approximately three weeks were spent in observations in the different visits to this site. Observations on how health workers deal with daily tasks, such as data collection and analysis, and meetings were conducted.

2.3 Additional Field Research

In addition to the research in the Ministry of Health, I conducted three additional field studies as part of my Masters studies.

2.3.1 Field Research on Internet Forums

My first research was conducted while I was in Oslo. The research was carried out through the support of a Brazilian group of Masters students in information systems and another on Internet in a group discussion about data warehouse. Five students with background in computer science formed the group of students from the Universidade Católica de Brasília, Brazil.

The first research involved an exchange of information with Brazilian students based on email discussions. We exchanged information about data warehouse using email. A set of questions was elaborated (See Annex A) and everyone presented their points of view, based on experience.

The second research was conducted in the data warehouse discussion forum website on the Internet (http://www.datawarehouse.com/forums/). The forum is for people interested in data warehouse topics and is a good source of information. The website is divided in several forums, from implementing a data warehouse to Extraction, Transformation and Load (ETL) processes topics.

In these two researches, I experienced some different difficulties such as, some answers led to other questions, there was no feedback for some of the questions made, each individual in the group of Brazilian students and in the forum had different points of views about data warehouse topics. The limitations were related to the fact that initially we were discussing the topic on data warehouse, more in terms of business perspective than related to the healthcare sector, but these problems did not affect the research, and the whole process were useful. In terms of what I learnt, the main finding is on the fact that data warehouse is a project team, where each individual has a specific role. Findings related to the data warehouse advantages and disadvantages were important.

2.3.2 Training Health Workers

I was engaged in the health workers training due to the fact that I am a member of the Health Information System Project (HISP). The Health Information System Project implemented the District Health Information System (DHIS) at the time of my fieldwork in Gaza Province.

The training was carried out during different visits to the districts. During the visits, I performed different activities, such as software installation and users training. In Chibuto and Manjacaze we conducted courses on how to use the District Health Information System. The courses were delivered to three to five health workers of different fields, like nurses, administrative personnel and physicians.

From 23rd September to 3rd October 2002, the Ministry of Health conducted a course on health information management and information technology in Chicumbane district in Gaza. Health workers with different backgrounds such as, administrative, physicians, nurses and managers were involved in the course. The Provincial Directorate of Health selected two to four workers from twelve districts of the province to participate in the training. During the training I was one of five computer technicians that delivered material related to District Health Information System and other computer tools (e.g. Microsoft Office) to the trainees.

2.4 Data Collection

The most commonly used method of data collection is individual interviews. This procedure requires the interviewers to ask prepared questions and to record the respondents' answers. This can be done through the use of questionnaires.

2.4.1 Primary Data Collection

Primary data collection involved going to the field and gathering information regarding different processes that occur in the health area. Multiple methods were used to collect primary data: structured and semi-structured interviews, observations and questionnaires. The main sources for the primary data collection were the Ministry of Health and the fieldwork conducted in Gaza province.

2.4.1.1 Interviews

Interviews were the most typical data gathering method. There were two types of interviews: a structured interview and a semi-structured interview (Järvinen, 2001). According with Järvinen (2001),

> "an interview can be described as a conversation between interviewer and respondent with the purpose of eliciting certain information from the respondent. The type of interview differs depending on the research approach used".
>
> (Järvinen, 2001, p.130)

In the theory-testing approach, Bell (1993, p.93) recommends the completely formalized interview where the interviewer behaves as much as a machine as possible. In the case of evaluation research, if the criteria to be used in evaluation are known and predetermined, the formalized structured interview can be used, but if criteria are still sought for, the semi-structured interview is recommended. Järvinen (2001) argues, that in the formalized or structured interview the interviewer tries to stay as neutral as possible and interact with all the respondents in a similar way, whereas in the non-formalized or unstructured interview the interviewer tries to gather descriptions of the world of the interviewee with respect to interpretation of the meaning of the described phenomena.

> "How do you make a record of what the informant says? If you tape record the interview you can later transcribe it, but that is a long and arduous task, so you will usually find that you will want to transcribe only certain key interviews or a few parts of a particular interview. If you have not tape-recorded the interview, you will have to write down later what you are able to remember. Depending on your relations with particular informants, you may find it useful to take your write-up of the interviews to them. Interviewing should be seen as part of the whole fieldwork process, rather than as an isolated exercise".
>
> (White, 1997).

The researcher registered 24 interviews from the fieldwork conducted in different sites (See Annex C). Most of the interviews were semi-structured and not tape-recorded. Instead, notes were taken (see Figure 2.5).

Figure 2.5: Example of notes taken from the fieldwork interviews.

Among the interviewees were nurses, doctors, district managers, program managers, people who process and analyze data at different levels (national, provincial, district, health facility), consultants at the Ministry of Health. Interviewees express their feelings about different topics, such as the National Health Information System, the information flow and other healthcare services.

The interviews in the districts occurred when a team from Health Information System Project, of which I am a member, visited Gaza province to teach health workers to use district health information system software. The selection of interviewees did not follow any particular strategy; I worked only with people available at the different levels. At the Ministry of Health the interviews were not regular and were dependent on the interviewee's availability. From the beginning of my fieldwork in the Ministry of Health, I established four visits per month. After three months of visits, the scheduled visits were not possible to follow as such because the interviewees were not available for the meetings. From the interviews it was possible to understand the information flows and study more about the existing legacy systems in the Ministry of Health.

2.4.1.2 Observations

Everyone observes the actions of others. We look at other people and listen to them talking. Researchers should have good observations skills. Using observations researcher can confront what interviewees say and what interviewees do. About observation, Järvinen (2001) states that,

> "scientists seek reliable and objective observations from which they can draw valid inferences. They treat the observation as a part of a measurement procedure. Observations must be naturalistic; observers must be immersed in ongoing realistic and natural situations and must observe behavior as it occurs in the raw. Direct observation may be more reliable than what people say in many instances. It can be particularly useful to discover whether people do what they say they do, or behave in the way they claim to behave".
>
> (Järvinen, 2001, p.137)

There are two types of observation – participant observation and non-participant observation. Bell (1993) defined participant observation as "the transfer of the whole person into an imaginative and emotional experience in which the fieldworker learned to live in and understand the new world".

As a researcher, I was allowed to participate in a total of about ten meetings in different sites such as Ministry of Health, Provincial Directorate of Health and Districts Directorate the Health. The majority of the meetings were about management, planning, data analysis, and information production. During these meetings, I played two different roles and responsibilities. First as a non-participant observer, I took notes as in the interviews; about the way health workers presented, interpreted and analyzed data and information, and how they perform their daily tasks. Secondly as a participant observer, I was able to solve technical problems related to the functioning of computers and the District Health Information Software.

The observation technique was adopted to enhance understanding regarding the functioning of the National Health Information System in different ways. Field notes and data collected were fundamental for the task of reconstructing the context and understanding the process in the interpretating stage of the research. The notes and data collected provided the researcher with the opportunity to understand the organization under study, the Ministry of Health.

2.4.1.3 Use of Questionnaires and Interviews Schedules

In the research, three types of questionnaires were designed (See annexes D.1, D.2 and D.3). Using these questionnaires it was possible to collect the data for the thesis. Järvinen (2001) argues that,

> "a questionnaire in a paper or electronic format contains many structured or unstructured questions intended to be answered by the selected people. Questionnaire is mostly used data gathering data technique, one of the theory-testing research approaches".
>
> (Järvinen, 2001, p.134)

Compared to my Internet research in Norway, where I did not have access to all places, access in Mozambique was granted to selected sites. In these sites it was possible to interview and observe health workers in their daily work. Interviews conducted in the different sites helped to validate the observations. Regarding communication with the interviewees, it is important to note that the interactions were very formal in some cases. Interviewees did not volunteer any information beyond very specific answers to the questions, and the answers were often short. Though questions had been structured around different issues of data flow in health sector.

2.4.2 Secondary Data Collection

Two methods were used to collect secondary data: analysis of documents and existing systems. The main sources for the secondary data collection were the Ministry of Health, especially the Health Information Department, the Provincial and District Directorates of Health and two libraries of the Ministry of Health. Internal memos, reports from health facilities, districts, provinces, the strategic plans from different areas were the main documents collected. Sistema de Informação de Saúde (SIS) was the main system analyzed.

2.4.2.1 Review of Documents

The Ministry of Health keeps all organizational information in the library – Documentation Centre. Järvinen (2001) gave a definition of document, saying that

> "A document is something written or printed, and to be used as a record or in evidence. Documents involve a deliberate attempt to preserve evidence for the future, possibly for purposes of self-vindication or reputation enhancement".
>
> (Järvinen, 2001, p.138)

During the research process I decided which type of documents to access. I was authorized to access the following documents: internal memos at the Ministry of Health, reports from health facilities, districts, and provinces, as well as the strategic plan of the Ministry of Health for different areas.

The analysis of the documents was critical. Questions in my mind and answers given by interviewees were clarified at this stage. The analysis of documents was important because it provided the basic elements necessary to the fieldwork in the districts. As an example, a list of different forms and reports used and produced at the district level were provided.

2.4.2.2 Analysis of Existing Systems

As a researcher, I had the opportunity to access only one automated system in the Ministry of Health, the Sistema de Informação de Saúde (SIS). It was not possible to access and analyze existing manual systems. In the visited districts it was possible to analyze the following automated systems: District Health Information System (DHIS), Sistema de Informação de Saúde (SIS) and Computerized Weekly Epidemiological Bulletin (CBES). The analysis provided a general view of how these systems work within the Ministry of Health. Analysis if existing systems was important, as it provided information on how data is stored in the Ministry of Health. From the analysis it was also possible to conclude that the systems were not integrated.

2.4.3 Data Analysis

There is much theoretical and practical advice provided in the extensive literature on qualitative analysis and the different steps in this process. For example, Miles and Huberman (1994) define data analysis as three linked sub-processes: (1) data reduction, (2) data display, and (3) conclusion drawing and verification. The first is described in terms of data selection and condensation. In their words data analysis "refers to the process of selecting, focusing, simplifying, abstracting, and transforming the data that appears in written-up field notes or transcriptions". In data display, the second sub-process, the goal is to represent the data in a more comprehensive way. Displaying the data in diagrammatic, pictorial, or visual forms should make it easier to see what those data imply.

The last analytical sub-process, conclusion drawing and verification, focuses on interpreting and deciding what the, now, intelligibly displayed data actually means. Chapter 5 presents the findings and Chapter 7 presents the discussion of my research findings and research questions. In both chapters I presented different data analysis. The data analyses are related to the information flow, with focus on data quality.

2.5 Research Constraints

The first constraint of the study is the literature review. The literature review shows that several authors address data warehouse more within a business perspective. The second limitation is related to the empirical study. The researcher was not able to interview key staff at the Ministry of Health and access to the different departments with available legacy systems was poor. More interviews would have provided valuable data.

2.6 Ethics

Ethical standards were maintained during the research:

(1) Information was not disclosed without participants approval;
(2) Verbal consent was obtained from all participants;
(3) Names of participants were not used in the report;
(4) Confidentiality was maintained.

The thesis proposal was submitted and approved in the University of Oslo, Norway.

2.7 Summary

To achieve the proposed objectives of the research, I divided the work into four different stages. In the first stage of the research, I learnt about the Ministry of Health through reports and documents and conducted interviews and observations in the Ministry of Health. At the same time, I reviewed the literature on data warehouse and integration topics. In the second stage, I learned the theory about data warehouse implementation project with focus on the data warehouse life cycle.

I learnt about the existing systems in the Ministry of Health at this stage. At the same time I conducted some fieldwork in districts and in the Ministry of Health. In the third stage, the data warehouse architecture model was developed and presented. The last stage presents the conclusions and recommendations of the research. Figure 2.4 presents a summary of what I did during my research.

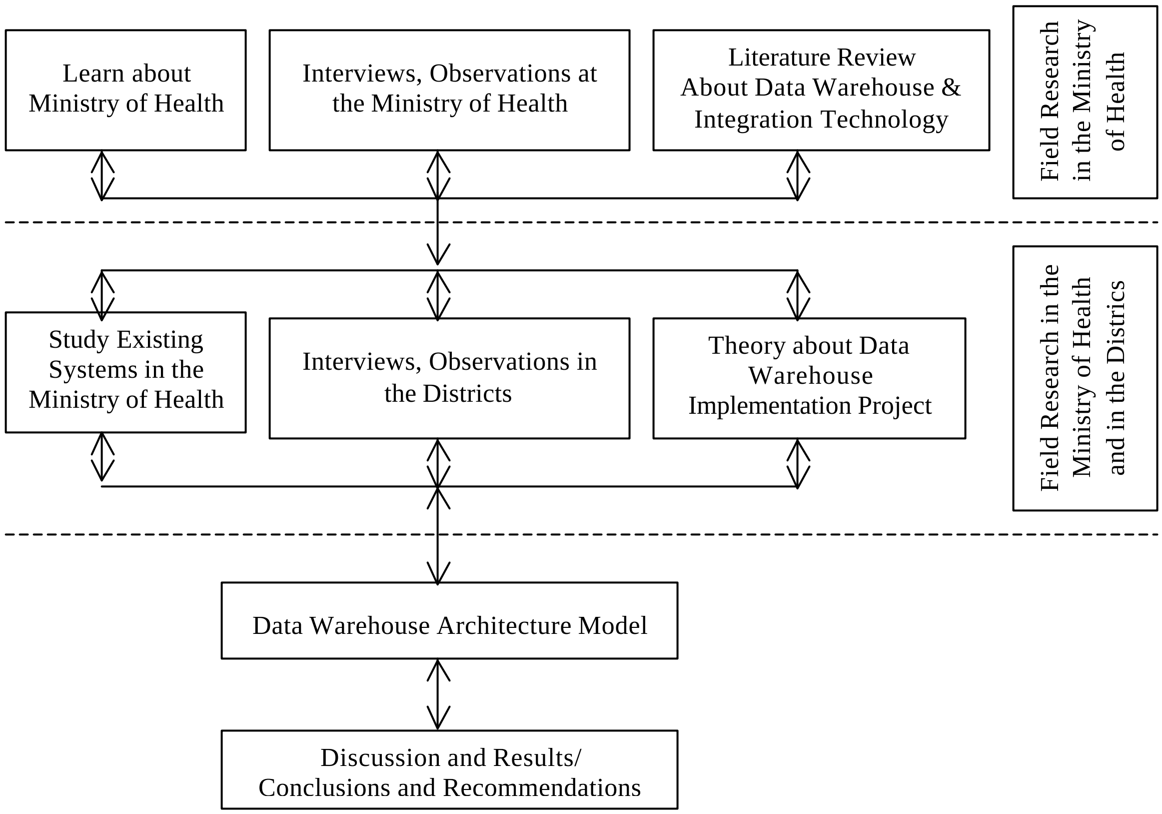


Figure 2.6 – Summary of different stages of the research.

Chapter 3

Literature Review on Data Warehouse

This chapter aims to present the literature review, a study about data warehouse technology, specifying concepts, characteristics and different types of data warehouses and data marts architectures.

Different researchers from different areas (database management, information system design, data and information integration) have come out with their own conclusions. As Mull (1983) observes:

"*We must be prepared to learn more than we can understand.*"

Thus, there are many sources that could be quoted to illustrate the research methods used to understand data warehousing and integration concepts. The work summarized here is based on relevant literature review and on research performed in Norway and Mozambique.

3.1 Data Warehouse Methodology

With the quick evolution of information and communication technologies and dissemination of computer use, most of large and medium size organizations are using Information Systems (IS) to implement their most important processes. As time goes by, these organizations produce a lot of data related to their business, but the data is not integrated. Such data are stored within one or more platforms and constitute the resource for the organizations, but are rarely used for decision-making process.

Traditional information systems are not projected to manage and store strategic information. They are formed by crucial data – operational data – needed for daily transactions. In terms of decisions, data are empty and without any transparent value for the decision process of organizations (Domenico, 2001). Decisions are taken based on administrators experience and sometimes based on historical facts stored in different information systems.

A data warehouse is projected in a way that data can be stored and accessed and is not restricted only to tables and relational lines. As the data warehouse is separated from operational databases, users queries do not cause any impact in these systems. Data warehouse is protected from any non-authorized alteration or loss of data. Data warehouse contemplates the base and the resources needed for a Decision Support System (DSS), supplying historic and integrated data. These data are for top managers, decision makers, partners, donors – who need brief, summarized and integrated information – and for low-level managers, for whom detailed data helps to observe some tactical aspects of the organization.

In this way, data warehouse provides a specialized database that manages information from corporative databases and external data sources.

3.2 Basic Concepts

3.2.1 Data Warehouse

In the bibliography many definitions can be found about data warehouse:

- Inmon (1997) says, that data warehouse is a data collection oriented to a subject, integrated, changeable in time and not volatile, to provide support to the decision-making process.
- Harjinder and Rao (1996) argue, that data warehouse is a running process that agglutinates data from heterogeneous systems, including historic data and external

data to attend the necessity of structured queries, analytical reports and decision support.

- Barquini (1996) defines the data warehouse as a collection of techniques and technologies that together provide a systematic and pragmatic approach to solve the end user problem in accessing information that is distributed in different systems inside organization.
- Kimball et al. (1998) argue that, data warehouse is a source of an organization data, formed by the union of all corresponding data marts.

To better understand the data warehouse concept it is important to make a comparative study between the traditional concept of database (DB) and data warehouse (DW).

"A database is a collection of operational data, stored and used by application systems from a specific organization", (Batini and Lenzerini, 1986). Data kept by an organization is called 'operational' or 'primitive'. Batini and Lenzerini (1986) referred to the data stored in database as 'operational data', distinguishing the input, output and other types of data. Based on the Batini & Lenzerini definition of operational data, I can define data warehouse as a data collection derived from operational data to support the decision-making process. "These derived data are most of the time called 'analytical', 'informational' or 'managerial' data" (Inmon, 1997). Table 3.1 presents the characteristics of two approaches:

Table 3.1: Operational data versus data in the data warehouse (Adapted from Madeira, 2002).

Operational Data	Data in Data Warehouse
• Operational objectives • Read/Write access • Access by predefined transactions • Access to few records at once • Updated data at real-time • Optimized structure for updates • Even-driven: Processes generate data	• Historical register • Read only access • Access by ad hoc queries and periodic reports • A lot of records in each access • Periodic load of more data • Optimized structure of complex queries • Data-driven: Data generate answers

Operational databases store all information needed to make possible the daily operations or transactions of the organization. They are used by employees to register and execute pre-defined operations; therefore their data can suffer changes, as new necessities for the organizations appear. Because redundancy in the data does not occur and historic information is not stored for long time, this type of database does not need a large capacity of storage.

Already a data warehouse stores analytical data that are needed by managers in the decision-making process. Data warehouse stores historic information of many years, it implies a big capacity of data processing and storage. Data are in two ways: detailed and summarized.

We can conclude that data warehouse is a resource that an organization has to analyze historic information. Data warehouse is composed by summarized data extracted from one or various computerized systems, normally used for several years and still working.

3.2.2 Basic Elements of Data Warehouse

The table that follows presents the main components of one data warehouse in an integrated vision of Kimball et al. (1998).

Table 3.2: Definition of basic elements of the data warehouse (Domenico, 2001).

PHASE	BASIC ELEMENT	DEFINITION
Data Sources	**Source systems**	Operational system that the function is to capture business transactions.
Staging Area	**Staging area**	Storage area and set of processes that clear out, transform, combine, remove duplications, archive and prepare the source data to be used in the data warehouse.
Integration	**Data Mart**	A logic subset of complete data warehouse.
	Operational Data Storage	Integration point of operational systems of the organization. Create to integrate the different systems of the organization at operational level.

Data Warehouse	**Presentations server**	Physical target machine where data are organized and stored for end users access, reports generators and applications queries.
	Dimensional model	Specific subject matter for data modeling as alternative for Entity-Relationship model.
	Relational Online Analytical Process (ROLAP)	Set of user and application interfaces that gave multidimensional characteristics to relational databases.
	Multidimensional Online Analytical Process (MOLAP)	Set of user interfaces, applications with a proprietary database that are strongly
	Metadata	All information in the data warehouse environment that is not real data.
Dimension Construction	**Online Analytical Process / Online Analytical Process Cubes**	Generic activity of querying and presenting textual or numeric data from data warehouse, as well as a specific dimensional way to query and present. It is a non-relational technology and usually based on data multidimensional cubes.
Data Analysis Tools and Applications/Users Exploration of Information	**Business process**	Coherent set of organization business activities that make sense to business users of data warehouse.
	Application for end user	Collection of tools that query, analyze and present desired information.
	Data Access Control Tool for End-Users	Data warehouse client. Can be simple as ad-hoc queries systems or complex and sophisticated as data mining or modeling applications.
	Tools for Ad-Hoc Queries	Specific type of data access tool that induces the end user to create his/her own queries, manipulating relational tables and its functions directly.
	Modeling Applications	Type of sophisticated tool with analytical capacity to transform or understand the data warehouse outputs (e.g. Data Mining, Forecast Models, Behavior Models)

The work presented by Kimball et al. presents a general vision of basic elements of a data warehouse. Data warehouse construction need to reflect these approaches and solutions for each element.

Legacy systems, usually human resource, finance, logistics, maintenance, external systems and other systems compose the source systems. Due to operational nature, such systems are inadequate to the production of temporary reports and of difficult use for the managers of the organization. In the following section I describe the topic data warehouse with more detail.

3.3 Data Warehouse Characteristics

The data warehouse always contains data and information, on which management decisions can be reliably tested, analyzed, assessed and monitored using the data and information integration. A data warehouse is that portion of an overall architected data environment that serves as the single integrated source of data of information processing, and is subject oriented, integrated, time variant and non-Volatile (Inmon, 2001). A data warehouse has the following characteristics (Watson, 2001): subject oriented, integrated, time variant, non-volatile.

3.3.1 Subject Oriented

Data warehouse data is organized around specific subjects, such as sales, products or customers. This arrangement is different from transactional systems where data is organized by business processes, such as order entry, inventory control or accounts.

Data warehouse store important data for the organization in accordance with the needs of staff that will use information. The Ministry of Health, for example, has among his objectives, to provide sanitary services to the community. One of the interests of the Ministry of Health is to know the profile of its collaborators, what makes the data warehouse contemplate data of collaborators using the services provided.

When initiating data warehouse constructions it is important to involve end users, because they need to define what the important information is.

3.3.2 Integrated

Integration constitutes one of the main characteristics of data warehouses. Through integration we define a unique representation for data coming from different systems that will compose the database of the data warehouse. The operational systems and data analysis process take a long time because data has no standards in the codification. Each system analyst defines the same data structure in different ways, so, the data that represents certain information is represented in different ways inside the systems used by organizations for long periods of time.

One example of a non-standard problem is the representation of marital status for one health worker. In one system, e.g. human resources, we can define the field of one alphanumeric position, where M='Married', S='Single', O='Other'. In another system, e.g. staff training, 1, 2 and 3 can represent the same information, respectively. With the integration of data, this problem will disappear as table 3.3 shows, because we should use a unique representation for the information.

Table 3.3: Data integration (Domenico, 2001).

System	Operational Environment	Data Warehouse
Application X	Represents marital status as: M, S, O	1, 2, 3
Application Y	Represents marital status as: 1, 2, 3	1, 2, 3

> "Two basic elements of data warehouse are related to integration: data stage area and operational data storage. Clean process, transformation and aggregation occur in the data stage, while the integration in the legacy systems occurs in the operational data storage".
>
> (Kimball et al., 1998)

3.3.3 Time Variant

A data warehouse maintains historical data (i.e. it includes time as a variant). Unlike transactional systems, where only recent data, such as for the last day, week or month are maintained, a warehouse may store years of data. Historical data is needed to detect deviations, trends and long-term relationships. All data in the data warehouse are needed in a specific space of time (Inmon, 1997). The variation in the time can be presented in three ways:

- In a data warehouse it is common for information to be represented in time horizon up to five years. In the operational applications the time period is shorter. It varies between two and three months, because quick answers are necessary to the demands of daily tasks, which can only be obtained through the processing of some data;
- As well as the data, the metadata, that include definitions of items of data, validation routines and derivation algorithms, has temporal elements so that, with eventual changes in the business rules, the organization does not lose historical data;
- Data stored correctly in the data warehouse will not be updated, having a faithful image of the time where they had been generated.

Data can also be separated in two categories:

- Actual detailed data are data of bigger interest because they reflect most recent transactions. They appear in big volumes and are stored in places of quick access, but of difficult management. These data provide a view of recent behavior and allow the use of techniques such as data mining and knowledge discovery.
- Old detailed data are data that are not accessed frequently and are normally stored in low cost devices. Depending on the needs, the data can be loaded on the data warehouse.

3.3.4 Non-volatile

A data warehouse is non-volatile – users cannot change or update the data. Non-volatility makes sure that all users are working with the same data. The warehouse is updated, but through information technology department rather than by users. After the loading process (from operational database) users can only access data using queries.

3.4 Data Warehouse Architecture

One of the data warehouse requisites is to be able to answer to fast queries. For this, data warehouse must have an architecture that allows gathering, manipulation and presentation of data quickly and efficiently.

In a research conducted by Barker (2000), he gives an overview of the data warehouse topic. Barker introduces data warehousing (DWH) and makes a match between data warehouse (DW) and operational database systems (ODS). Figure 3.1 presents a generic three-tier data warehouse architecture proposed by Barker.

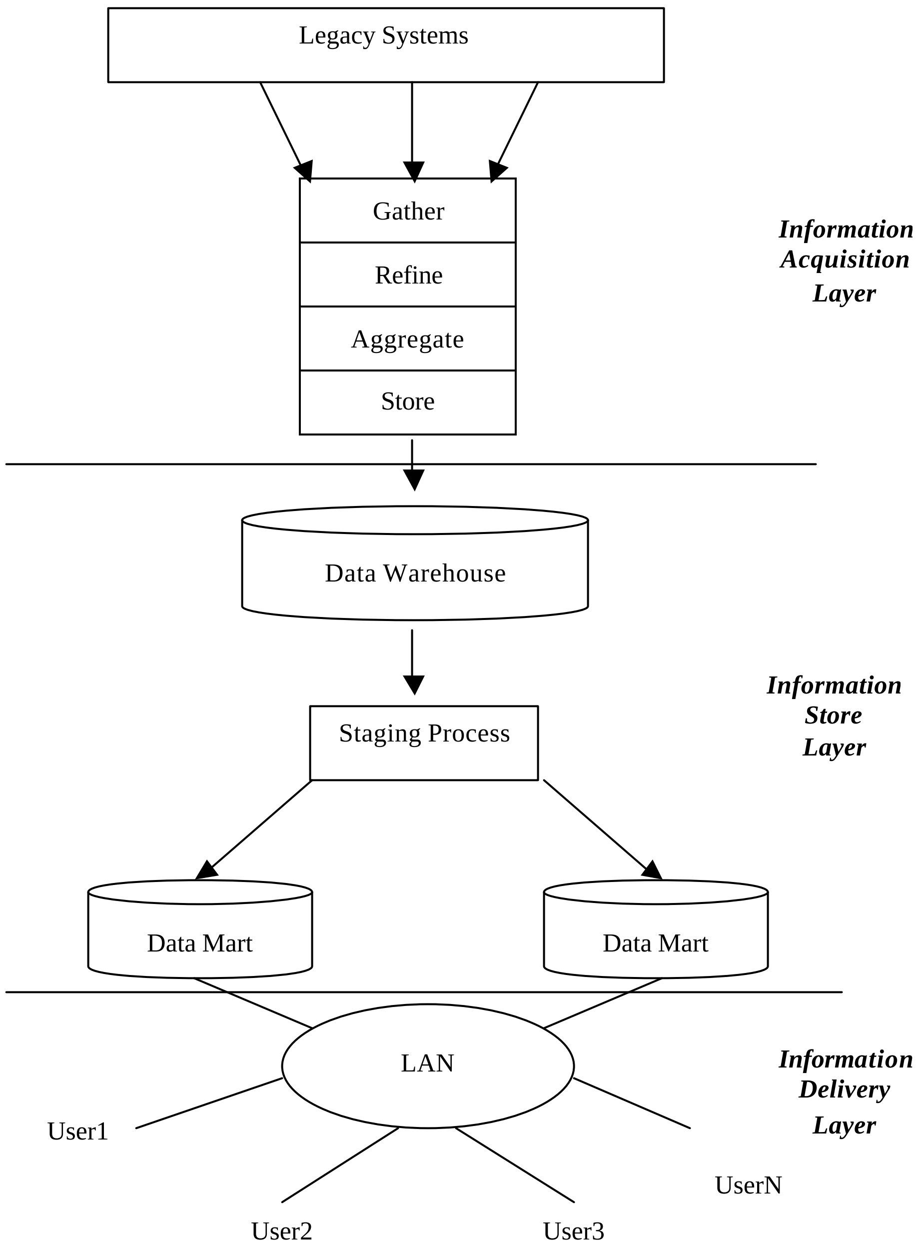


Figure 3.1: Generic three-tier data warehouse architecture (Barker, 2000).

Information Acquisition Layer

At this layer we find data sources (legacy systems, transactional systems and external sources) and the staging area (extraction, clean, transformation, load, data transformation services).

> "Over the years, organizations developed a large number of applications to support business processes. Many of these applications were written in COBOL (Common Business Oriented Language), Smalltalk, and Clipper and continue to be used today. Many of the applications are poorly documented and use cryptic table and attribute names. The data in these applications are often difficult to access for decision-support purposes and even if the data can be accessed, doing so creates performance problems for the applications because accessing the data slows the processing of transactions".
>
> (Watson, 2001)

The Ministry of Health is no exception to this rule and has legacy and operational systems from different vendors such as Dbase, FoxPro and Microsoft Access. These systems are an important source of decision support data, but the software often stores data in complex data structures, and the systems have different formats. As a result, extracting data from them to make integrated analysis is difficult.

Data sources often use different hardware and software, and a mixture of hierarchical, network and relational data models for storing the data. The data are extracted from the sources using custom-written software or commercial extraction, transformation and loading (ETL) software. The data is then fed into the staging, area where it is transformed.

Data extraction, transformation and loading (ETL) take data from source systems, prepare it for decision-support purposes, and places it in the target database. Target databases can be an Oracle database or Microsoft Access database. Watson (2001) argues that,

> "It is the ´plumbing´ work of data warehousing – dirty, complex, time consuming and expensive. More than one data warehousing project failed because appropriate extraction, transformation and loading (ETL) processes could not be put in place, either because of problems working with sources systems, inadequate technology for the task at hand, inadequate data warehousing expertise on the project, and/or organizational issues".
>
> (Watson, 2001, p.6)

The data are transferred from data sources by means of interfaces. The two major options for extracting data from sources systems are (Watson, 2001):

1. ***Custom-write data extraction programs***. Because of the prevalence of COBOL, Smalltalk and Clipper legacy applications, the extraction programs are often written in these languages with embedded structured query language (SQL) queries. Companies that know their source systems well, understand their complexities, have excellent in-house technical skills, and want to avoid the cost of purchasing extraction, transformation and loading (ETL) software, may want to write their own ETL software.
2. ***Purchase commercial ETL software***. This software is available from the major database vendors (e.g., IBM, Teradata or Oracle). The specialized extraction, transformation and loading (ETL) software allows companies to specify source systems relatively easily

Information Store Layer

The data mart strategy is a "*start small, think big*" (Watson, 2001). A data mart is similar to a data warehouse, except that a data mart stores data for a limited number of subject areas. Because it is smaller in scope than a data warehouse, it also tends to contain less data and fewer applications. Data mart can be either independent or dependent. An *independent data mart* is built from source systems and *dependent data mart* is created with data drawn from a data warehouse. A data mart has the following advantages:

simple to implement, provide usable data faster, fast construction, lower cost and less financial risk.

The starting point for the design and development of the data warehouse environment is the data model. Without the data model, it is difficult to contemplate building the data warehouse. The data model acts as the roadmap for development.

Information Delivery Layer

This layer prepares information that can be accessed using dynamic queries with a good performance. In the literature, Domenico (2001) presents a general view of how to implement a data warehouse in a high education organization in a study conducted in Brazilian Universities. He first presents a study about organizational structure, with a focus on decision process in these types of organizations. The author presents the basic concepts, data warehouse architectures and characteristics. He argues that data warehouse construction can be done in two ways: (a) Top-Down architecture, known as Inmon architecture, and (b) Bottom-Up architecture, known as Kimball architecture.

3.4.1 Top-Down Architecture/Inmon Architecture

Introduced by Inmon (1997), it is the first data warehouse architecture. Figure 3.1 presents the top-down architecture. The first step is the extraction, transformation, migration and load of data coming from legacy systems or external sources. In the extraction, transformation and migration process, data are collected from different sources and stored in the data staging area. After that, data and necessary metadata are loaded into data warehouse. Having data warehouse, data marts that is a subset of the data resource, usually oriented to a specific purpose or major data subject, and it can be constructed from summaries of data warehouse and metadata. Data warehouse is formed by atomic data and also historic detailed data. In contrast, data marts contain highly summarized and softly summarized data.

The top-down model can use normalized entity-relationship data model (Domenico, 2001); in contrast, data mart uses star data model. In top-down architecture model, integration between data warehouse and data marts is automatic. We only need to know that data marts are subsets of data warehouse. Major criticisms of top-down model are that implementation has high cost and takes time to obtain partial results. Figure 3.2 presents a top-down model of data warehouse based on the Ministry of Health existing systems.

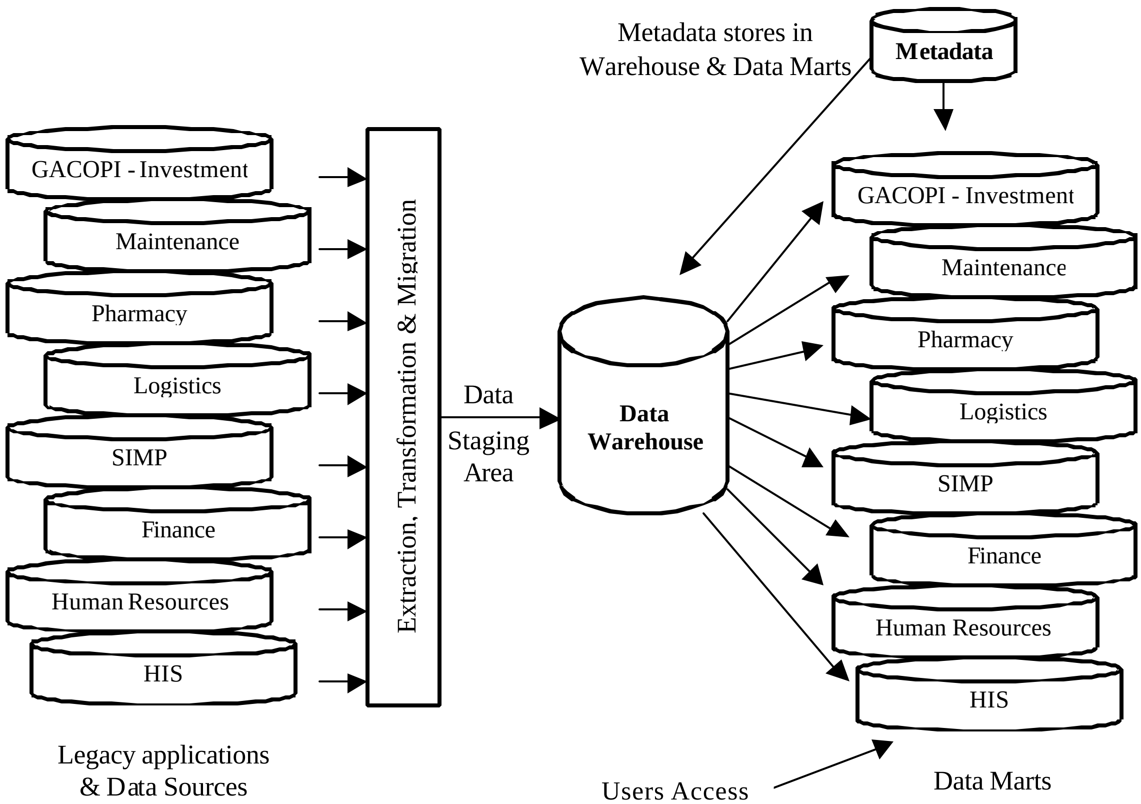


Figure 3.2: Top-Down Architecture for the Ministry of Health (Adapted from Domenico, 2001).

3.4.2 Bottom-Up Architecture/Kimball Architecture

The second data warehouse architecture is the bottom-up architecture. Top-down architecture took long to be implemented. It was politically unacceptable and implementation was very expensive and delayed. Bottom-up architecture offers an alternative in the data warehouse construction, the incremental construction.

The basic idea of this architecture is to construct the data warehouse in an incremental way. In the literature, Ralph Kimball introduces this architecture. The process starts with the construction of one or more data marts. Also, the data staging area exists, and is separated from each data mart. Figure 3.3 presents the bottom-up architecture.

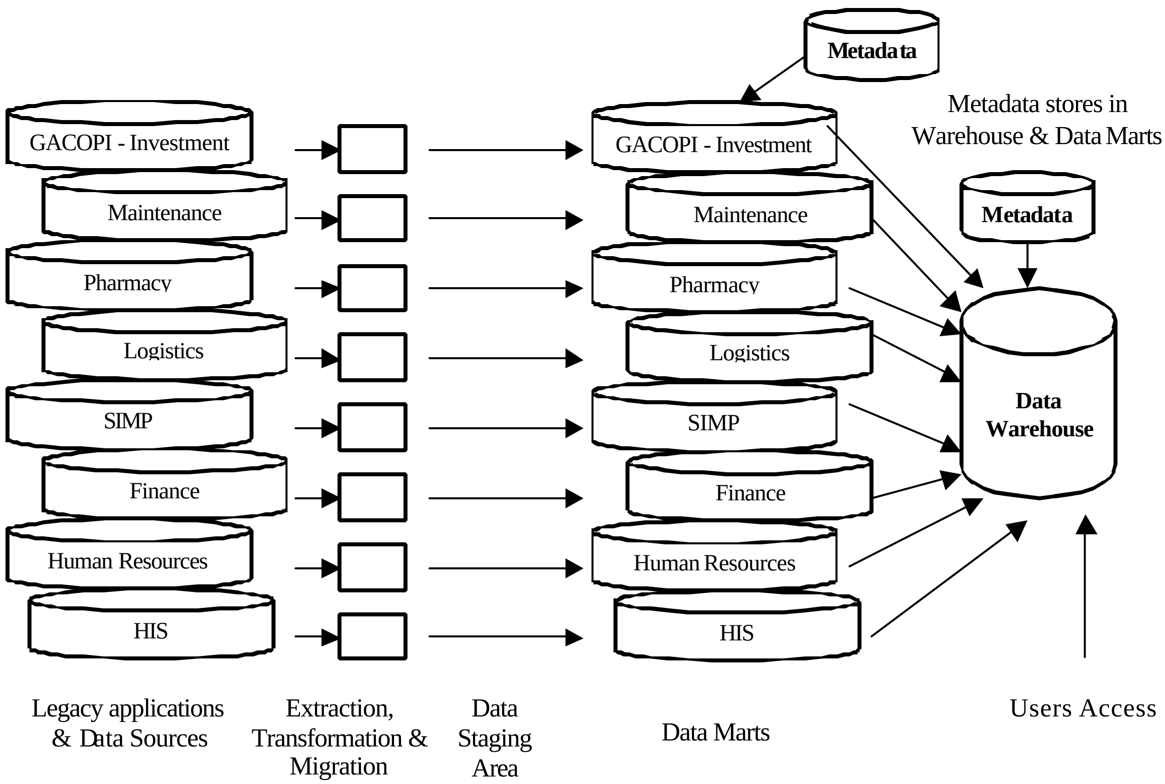


Figure 3.3: Bottom-Up Architecture for the Ministry of Health (Adapted form Domenico, 2001).

Data marts usually do not use entity-relationship data model in normalized form (Domenico, 2001). It is frequently is recognized that when data marts use relational database they use the star schema model, or a multidimensional modeling (Kimball et al., 1998).

In the top-down architecture, data marts use lightly and highly summarized data, but the bottom-up architecture use atomic and detailed data, including historic data. Considering that data marts are subsets that support the data warehouse, they should contain all data that appears in the data warehouse project.

Both approaches provide benefits but involve limitations. When successfully executed, both strategies result in an integrated enterprise data warehouse (Watson, 2001). Domenico conclude based on his research that, Kimball model compared with the others, e.g. Inmon model, is the best one, because it is incremental and it foresees integration.

Another important conclusion of Domenico research is that, the management process is a purely human process. He argues that having all available technology the human beings still take all the decisions in the organizations. The technologies only help managers in their decision-making process.

3.4.3 Different Data Warehouse Implementation Approaches

Different authors (e.g. Inmon, 2000) argue that in order to build a successful data warehouse, a hybrid top-down and bottom-up approach must be used, because this approach keeps the projects simple, focused, and short.

The top-down approach dictates that the central data warehouse is built first. All common elements, attributes, and dimensions are identified and collected. All levels of aggregation and facts are placed in the corporate data warehouse and then exact subsets of the data are created in the distributed, subject area data marts. The problems with this approach are the massive effort and lengthy time required to define the corporate data needs. This often causes data warehousing efforts to stall and ultimately fail (Inmon, 2000). The bottom-up approach establishes distributed, subject-orientated data marts first, using a multi-phase approach, and then combines the data from all data marts into a corporate data store. The benefits of this methodology are four-month turn-around on the data marts, quick return on investment, and a chance to leverage lessons learned as the project team moves from installation to installation.

The downside to using the bottom-up approach in its purist sense is preventing the creation of stovepipe data marts. It is very easy when using a bottom-up approach to make design or quality concessions within the individual data marts that will make future integration into the corporate data warehouse very difficult. It is not uncommon for further data quality, integrity, and business rules being required to populate the data warehouse from these stovepipe data marts. In this way, the data marts become a second staging area for the corporate data warehouse instead of an end user reporting facility (Inmon, 2000).

The hybrid approach ensures that the individual business units' needs are focused upon, answered, and delivered within four-month windows as per the bottom-up approach. The twist comes with defining and then using a framework to facilitate corporate integration (Inmon, 2000). The three major steps to the iterative hybrid approach are planning, implementation, and support as presented in Figure 3.4.

Planning • Define the need for a data warehouse • Design a corporate data warehouse architecture
Implementation • Build data marts
Support • Review the data warehouse to ensure integrity, consistency and usability

Figure 3.4: Summary of the hybrid approach to data warehouse architecture.

Using this hybrid approach solves the integration problem as all data marts will share common business rules, semantics, and the enterprise metadata repository (organization metadata) enforces definitions.

Detailed steps to hybrid approach to data warehouse development, as shown in Figure 3.5, follow an iterative cycle.

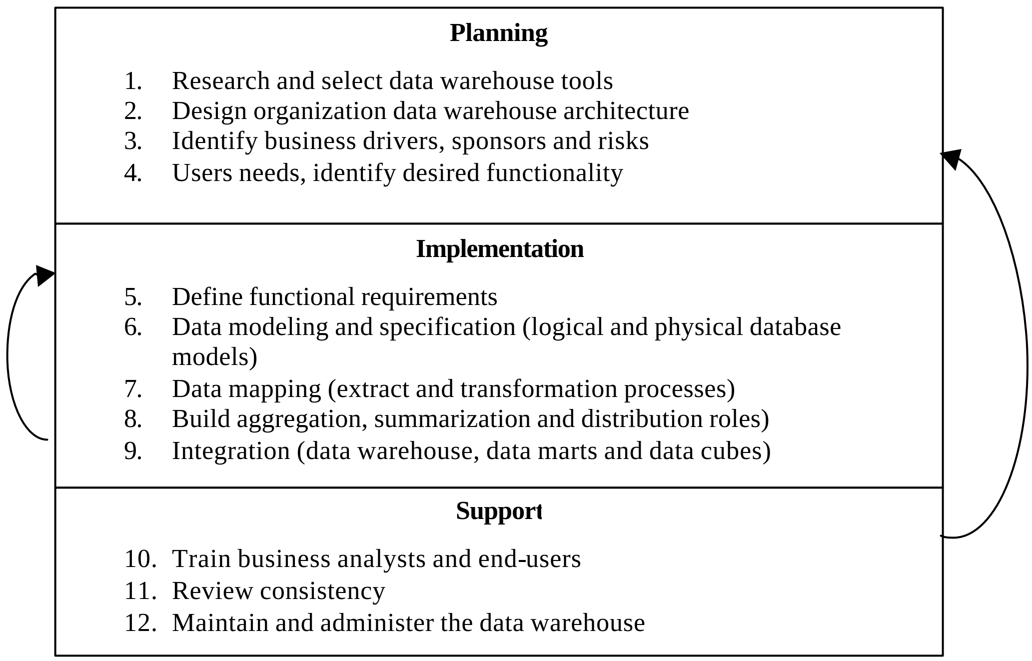


Figure 3.5: Hybrid approach for data warehouse implementation.

3.5 Data Warehouse Life Cycle

The data warehouse is built in an entirely different way from that in which the classical operational environment has been built. The classical operational environment has been built on the system development life cycle – the SDLC (Software Development Life Cycle) – that requires that requirements be identified, that analysis proceeds, followed by design and programming. Then testing occurs, and is completed by implementation. The SDLC is fine where requirements can be known in advance. The DSS (Decision Support System) analyst – who is the ultimate beneficiary of the data warehouse – does not and cannot know their requirements. There is an entirely different mode of development that is required for data warehouse development (Inmon, 2000a).

Flanagan and Safdie (1998) present a data warehouse life cycle with the following stages:

1. **Document Organization legacy systems**
2. **Model of data warehouse**
3. **Clean data**
4. Move data
5. Query against data warehouse
6. Manage the data warehouse environment

From the literature review, different authors suggested that the following steps are an extension of the data warehouse life cycle and should be followed when building a warehouse:

1. Collect & analyze business requirements
2. **Define data sources**
3. **Extract, transform, clean data**
4. **Create a data model & physical design of the data warehouse**
5. **Choose database technology/platform for the data warehouse**
6. Populate data warehouse
7. Choose database access & reporting tools
8. Choose database connectivity software
9. Choose data analysis & presentation software
10. Update the data warehouse

Bold points in above list (1, 2 and 3 for data warehouse life cycle and 2, 3, 4 and 5 for steps to data warehouse construction) represent the points covered during my study and are presented in more detail and discussed in chapter 6.

3.6 Data Warehouse Development Issues

A data warehouse is a database that pulls together information from a variety of data sources to give an integrated view of business activities. A data warehouse is a multidimensional database that optimizes data retrieval in response to analytical queries. Data warehousing is about turning stored data into useful information. The primary motivation for an organization to implement a data warehouse usually centers around improving the accuracy of information used in the decision-making process.

The implementation process is a collaborative process involving many people throughout the organization and often external to the organization. Decisions must be approved and communicated to everyone who must take action. In order to support the whole process of data warehouse implementation, the organizations must define a team and the project.

A data warehouse development requires phased development cycles as a rapid applications development (RAD) project. About data warehouse development phases Oster (1998) write:

> *Justification phase*: The project team is formed – drawn from IT and end user departments. The project team begins by evaluating what the possible impact of a data warehouse for the business could be, and why it is justified to start a data warehouse.
>
> *Requirements Gathering phase*: The project team collects the end user departments' information requirements (what, in what form, how often) and examines the site's IT architecture and strategy.
>
> *Design / Modeling phase*: Based on the information collected in the previous phase, the logical data warehouse is designed, data sources are identified, and transformation and business rules are defined. The logical data model is mapped to a physical data model and a prototype warehouse produced.
>
> *Implementation phase*: All necessary programs and applications are written, the data warehouse is made available to the end user department and all users of the data warehouse are trained.

Review phase: The whole project, its success and implications on the company's business are evaluated, both immediately after the data warehouse has been delivered and after a certain time has elapsed.

Each of the above phases provides a project deliverable as input for the next phase.

3.7 Data Model

In the literature we can find different approaches about data modeling. Although exists more than one model to construct a data warehouse with success, dimensional modeling becomes more effective for a data warehouse project (Domenico, 2001). In this research, I present in summarized way, Kimball dimensional model of data.

Complex questions that involve organization business analysis, usually require a vision of the data from different perspectives. Answers to this type of questions can lead to correct or wrong decisions. Tools based on Structured Query Language (SQL) help in the data search related to this type of queries.

In Mozambique, a good example can be the case of the National Health Directorate; Pharmacy Department at the Ministry of Health wants to increase the performance of sales in the different pharmacies around the country. The Pharmacy Department wants to know information about all pharmacies, if promotions are bringing results. To answer these questions, it is necessary to analyze data about sales in different pharmacies branch offices.

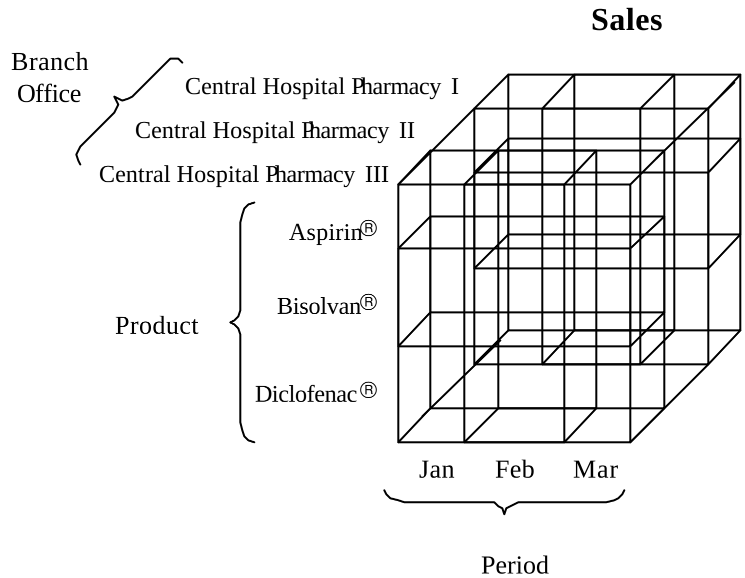


Figure 3.6: Multidimensional model for sales. Adapted from Madeira (2002).

This type of analysis requires a historical vision based on the sales volume, analyzing multiple perspectives, as, for example, quantity of sales per product (medicines), quantity of sales for each branch office or total of sales for a certain period of time.

Dimensions are the different perspectives involved, in the case, product (medicine), branch office and month. Figure 3.6 presents an example of the multidimensional model for sales. Dimensional modeling is the name of one technique of logical project, used frequently for data warehouse. The main objective is to present data in standard architecture, to allow high performance access (Kimball, 1996). Each axle in the multidimensional space corresponds to a field or column of the relational table and each point a value for the intersection of these fields or columns. Dimensional data can be stored and represented in relational structures. For such, it is necessary to use specific forms of modeling, as the star model described in the following section.

3.7.1 Star Model

Valente (1996), in his research, states that, traditionally, data model of relational databases presents tables with complex relationships and with multiple unions. For most users using tools to compose their queries, it is necessary that the access to the database is simple to facilitate the direct access to the database.

The main type of dimensional model is called Star Model, where a dominant table exists at the center of the model. The table in the center is called Facts table. This table has multiple junctions connecting with other tables called Dimension tables. Each secondary table has only one junction with a fact table.

The star model presented in the Figure 3.7 has the advantage of being simple and intuitive. For Kimball (1996), entity-relationship model is not adjusted to data analysis in the management environment. The dimensional model is the most appropriate for this environment.

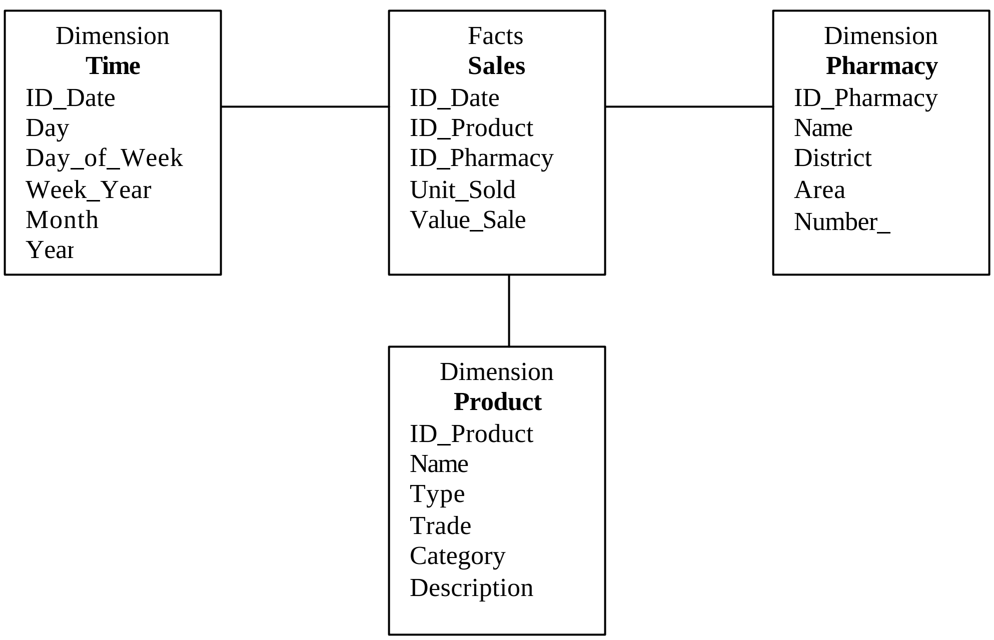


Figure 3.7: Dimensional model of the star type. Adapted from Domenico (2001).

The facts table contains thousands (or millions) of values and measures of business, as transactions of sales or purchases. The most useful facts are numerical and additive. Facts tables represent many-to-one (M:1) relationships with business dimension tables.

Dimensions table store textual descriptions of business dimensions. Each table represents one business dimension, e.g. time and product. One important factor related to facts table is that, as it represents the relationship many-to-many between dimension tables, it has as primary key, a key composed of all foreign keys of dimension tables (Kimball, 1996).

The dimensional model presents several advantages (Kimball, 1996):

- Predictable and standard architecture;
- Dimensions of the model are equivalent;
- It is flexible, because it allows the inclusion of new data elements;
- Easy alteration of facts and dimension tables;
- All the applications that existed before the changes continue operating without problems.

Another type of structure sufficiently common is the snowflake data model. The snowflake model is an extension of the star model presented above. Each “tip of the star” becomes the center of other stars. The model appears from de-normalization and cardinality reduction of star model breaking original table.

3.8 Metadata

Metadata constitutes the nervous system of data warehouse. Without metadata, data warehouse and its associated components in the projected environment become isolated components, functioning independently and with separated objectives (Inmon, 1999). In the operational environment users interact with information through screens and forms, allowing users to be unaware of how information is stored in the database. Metadata is treated later in the process and normally has the same importance as system documentation.

Metadata is typically defined in the literature as “data about data”. For Inmon (1997), metadata can be characterized “as one directory that helps decision support systems analysts to find data warehouse components". It keeps information about data structure, in accordance to the programmer and decision support system analyst. It also keeps information about data model, archive specifications (keys and attributes), description of extractions and access control.

Metadata can be classified as two types (David, 1999), technical metadata, used by development and maintenance personnel and business metadata, used by business analysts. There is also the core metadata, used by database and data administrators. Technical metadata provides developers and database administrators with all technical descriptions about data and its operations. It specifies attributes names, data types, sources from where they had been extracted, and transformation rules. Business metadata are usually used by end users and are derived and inferred for existing specifications. It is a link between data warehouse and business users, who are the executives or business analysts and tends to be less technical, but they need to have a clear vision of business rules. Figure 3.8 presents the different types of metadata.

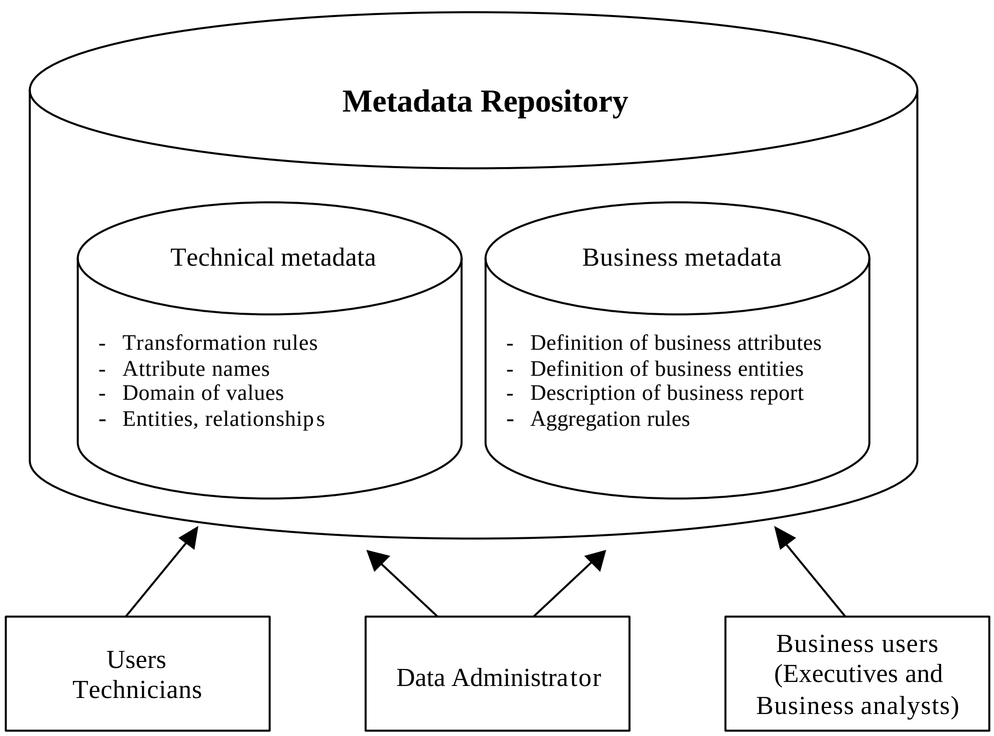


Figure 3.8: Metadata repository (Domenico, 2001).

Central metadata represents the way in which the data are treated in the system. It is related to data warehouse performance, is useful for queries generation in order to have a good performance. Examples are list of accessed tables or statistics of input and output for a query. Technical and business metadata are part of organization existing systems, and therefore are characterized as formal metadata because they are documented and formalized. While metadata that is not documented, it is only in people's head and characterized as informal metadata.

3.9 Data Mart

The data mart strategy is a "*start small, think big*" approach. It typically begins with a specific business need for data (Watson, 2001). In the construction of one integral corporative data warehouse they are factors that affect the complexity, for example, the construction of the project that is delayed and expensive. With the aim of balancing the costs and offer results in short stated periods, it is possible to construct data marts that in fact are small departmental data warehouses. One advantage of data marts construction is that the implementation time can be reduced.

In the literature we can find several definitions of data mart. Domenico (2001) defines data mart as a specialized system that provides the necessary data to the department or a relational application. Inmon (1997) writes, "Data marts are organization data subsets physically stored in more than one place, usually divided by department (departmental data marts)".

There are different alternatives to implement data marts. Originally they were developed from a central data warehouse. Figure 3.9 presents top-down data mart architecture also called dependent data mart. It created with data drawn from a data warehouse. It provides a copy of data extracted from the data warehouse.

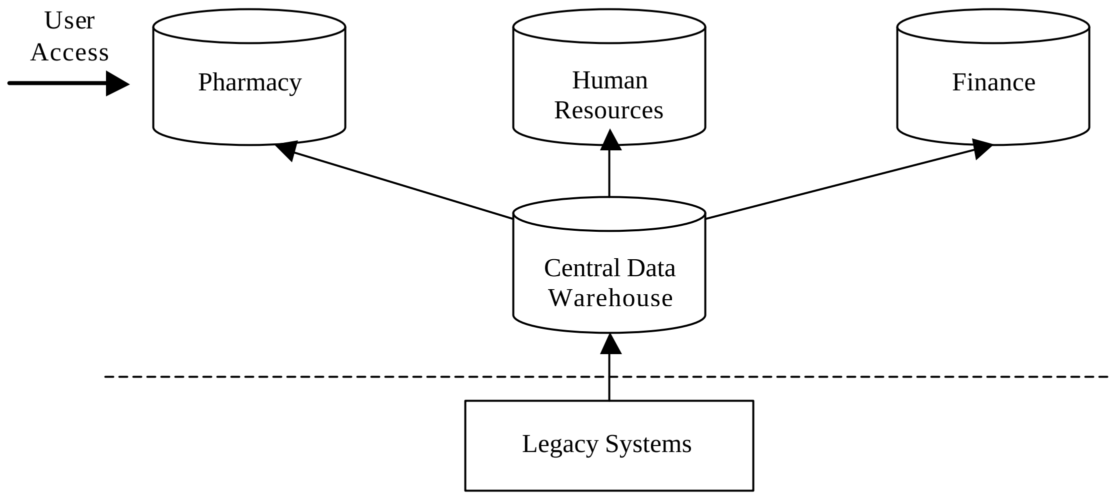


Figure 3.9: Top-down architecture.

In this architecture, groups of users access data marts directly from its respective departments. Data is acquired from legacy system directly to the data warehouse. Only analyses needing a general vision of the organization are made in the data warehouse.

Figure 3.10 illustrates the bottom-up architecture. This architecture is also known as independent data mart.

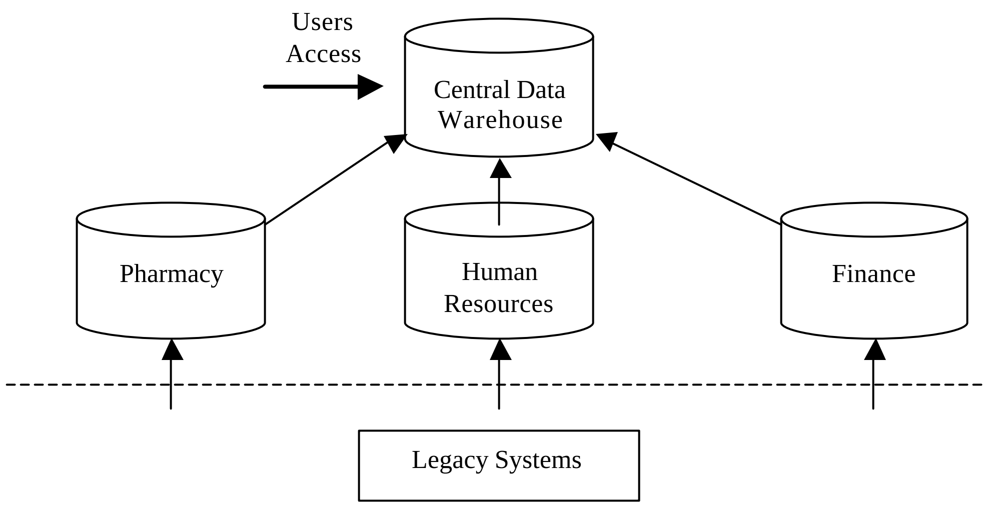


Figure 3.10: Bottom-up architecture.

In this architecture, data is acquired from legacy system and departmental data marts are the first place where data is placed. After that the central data warehouse can be built.

Data marts differ from data warehouse due to the following factors:

- *Can be personalized*: They take care of the needs of a specific department or group of users;
- *Hold smaller volume of data*: Because they take care of a specific department, they store a smaller volume of data;
- *Limited description*: Data marts rarely maintain the same period as data warehouse;
- *Volatile*: Users can change or update the data directly in the data mart; and
- *Summarized data*: Data marts do not usually keep data granularity at the same level as data warehouse.

3.10 Data Warehouse Key Component Areas

Complete data warehouse architecture includes data and technical elements. From the literature review, I found that most of the authors on data warehouse area break down the architecture into three broad areas. The first, *data architecture*, centered on business processes. The next area, *infrastructure*, includes hardware, networking, operating systems, and desktop machines. Finally, the *technical area* encompasses decision-making technologies that will be needed by the users, as well as their supporting structures. Different experienced authors in the field of data warehousing state that it is difficult to develop a data warehouse architecture model and that it involves an organizational effort.

3.11 Data Warehouse Challenges in the Area of Public Health

The construction of a data warehouse for public healthcare data poses major challenges beyond those required for the construction of a commercial data warehouse (e.g., retail sales). Such challenges include:

- Data come from a diverse set of sources. Health care data are published in a wide range of formats with differing semantics. There are currently few standards in the health care field for data. The data integration task to build the data warehouse requires significant effort;
- Reports are disseminated to a diverse and geographically distributed set of stakeholders;
- Data warehouse is required to support the activities concerning public policy formulation. Socio-political issues of health care policy impact on design features such as security, availability, data quality and performance.

Kerkri et al. (2001) argue, that over the past years data integration has become one of the most important issues confronting senior executives in the healthcare industry tasked with planning and executing long-term information technology strategy. This is mainly due to the following

1) ***Exponential increase in information loads***
   The amount of data that healthcare workers process in order to keep patient data is currently scaling at near exponential rates.
2) ***Complex data environment***
   Managing this emerging data environment presents healthcare companies with a number of challenges:

   i) The new information is often highly complex in structure;

ii) Much of the data being loaded into the enterprises' information infrastructure originates from sources outside the enterprise and over which it has little control;

iii) The data sources are in heterogeneous, often proprietary, formats with little or no common structure;

iv) The data sources can be structured, semi-structured and unstructured and in relational databases or flat files;

v) Health workers often need real time or near real time access to large sections of the data;

3) ***Planning is difficult***

The future structure of the healthcare industry is unclear. Regulatory, structural, market and technological changes are converging to create a period of unprecedented change and uncertainty. At the same time, many of the new information based on analytical techniques are not yet fully validated. Today, the role of most individual new technologies in the long-term structure of the different healthcare processes is uncertain.

Senior executives in the healthcare industry tasked with planning and executing long-term information technology strategy need to build sufficient flexibility into their plans to take account of this fact.

All of these factors combine to make the task of planning, managing and integrating data sources and data generation technologies in healthcare organizations extremely challenging.

3.12 Data Warehousing Advantages and Disadvantages

Whereas a data warehouse is a repository of data, data warehousing is the entire process. Data warehousing encompasses a broad range of activities: all the way from extracting data from source systems to the use of the data for decision-making purposes (Watson, 2001).

Sakaguchi and Frolik (1996) conducted a research that presents the advantages and disadvantages that information systems (IS) managers will encounter while developing data warehouses. The research was based on 456 published articles on the data warehousing topic. The referred advantages of data warehousing are as follows:

- ***Simplicity***. The most frequently (59, 13 percent of 456 articles) mentioned advantage of data warehousing is summarized as 'simplicity'. Data warehouses allow existing legacy systems to continue in operation, consolidate inconsistent data from various legacy systems into one coherent set, and reap benefits from vital information about current operations. Data warehouses can also store large amounts of historical data and corporate wide data that companies need to turn into vital business information Data warehouses offer the benefit of a single, centralized data location while maintaining local client/server distribution. Furthermore, data warehouses are company wide systems; therefore, they improve corporate wide communication.
- ***Better quality data; improved productivity***. The second most frequently (53; 12 percent) mentioned advantage is better quality data. Other data quality issues include consistency, accuracy, and documentation. Improved decision-making through OLAP and data mining analysis were mentioned as improvements in productivity.
- ***Fast access***. The next most frequently mentioned (48; 11 percent) advantage is "fast access." Since data warehouses allow users to retrieve necessary data by themselves, the work log of IS can be cut. The necessary data is in one place, so systems response time should be reduced.
- ***Easy to use***. Forty-seven or 10 percent of the articles mentioned "easy to use." Queries from users do not interfere with normal operations, because a data warehouse enables easy access to business data without slowing down the operational database by taking some of the operational data and putting it in a separate database. Data warehouses focus on subjects, support on-time, ad-hoc queries for fast decision-making, as well as the regular reporting; and they are targeted at end users
- ***Separate decision-support operation from production operation***. Another advantage mentioned in 32 articles (7 percent) is that data warehouses are built in order to separate operational, continually updated transaction data from historical, more static data required for business analysis. By doing so, managers and analysts can use historical data for their decision-making activities without slowing down the production operation.

- ***Gives competitive advantage***. Twenty six articles or 6 percent of them mention that data warehouses better manage and utilize organization knowledge, which in turn helps a business become more competitive, better understand customers, and more rapidly meet market demands. Therefore, this benefit can justify the large expense.
- ***Ultimate distributed database***. Fifteen (3 percent) of the articles discuss data warehouses pulling together information from disparate and potentially incompatible locations throughout the organization and putting it to good use. Middleware, data transfer software and other client/server tools are used to link those disparate data sources. A data warehouse is an ultimate distributed database.
- ***Operation cost***. In fourteen (3 percent) articles, it is said that data warehouses provide fertile ground to architect new operational systems. It eliminates paper-based files and once the initial investment is covered, the organization's information-technology group generally requires fewer resources.
- ***Information flow management***. The next highly mentioned topic (13; 3 percent) is that data warehouses handle a large amount of data from various operational data sources, and data warehouses manage the flow of information rather than just collecting data. To respond to changing business needs, production systems are constantly changing along with their data encoding and structures. Data warehouses, especially the metadata, help continuous incremental refinement that must track both production systems and the changing business environment.
- ***Enables parallel processing***. Eleven (2 percent) of these authors indicate that parallel processing helps users perform database tasks more quickly. Users can ask questions that were too process-intensive to answer before and data warehouse can handle more customers, users, transactions, queries, and messages. It supports the higher performance demands in client/server environment, provides unlimited scalability, and thus, better price/performance.
- ***Robust processing engines***. Ten (2 percent) of the articles mention that data warehouses allow users to directly obtain and refine data from different software applications without affecting the operational databases, and to integrate different business tasks into a single, streamlined process supported by real-time information. This provides users with robust processing engines.
- ***Platform independent***. Seven (2 percent) of the articles point out that data warehouses can be built on everything from a high-end PC to a mainframe, although many are choosing Unix servers and running their warehouses in a client/server environment. IBM and other five data warehouse software venders formed alliances to clear the cross-platform hurdles inherent in data warehouse implementation. Other vendors have formed similar partnerships. It is crucial to have such independence, which was not easy in the legacy system.
- ***Computing infrastructure***. Seven (2 percent) of the articles mention data warehousing helps the organization create a computing infrastructure that can support changes in computer systems and business structures.

- ***Downsizing facilitation***. Six (1 percent) articles suggest that data warehouses empower employees to make decentralized decisions since they put information closer to users. They are designed to give end users faster access to the information that is already there without impacting other systems or resources. Therefore, users do not need to ask IS to get needed data and IS managers can concentrate on other tasks. This potentially cuts the information middleman who passes information from one place to another and suggests downsizing.
- ***Quantitative value***. Another advantage, mentioned in six articles (1 percent), is realistic benchmarking. Data warehouses provide the quantitative metrics necessary to establish business process baselines that are derived from historical data and allow business managers to measure progress.
- ***Security***. Three (1 percent) articles talk about the fact that clients of the data warehouses cannot directly query the production databases, thus improving security of the production databases as well as their productivity. Some warehouses also provide management services for handling security.

Data warehousing is not without its disadvantages and they are the following:

- ***Complexity and anticipation in development***. The most frequently (48; 11 percent) mentioned disadvantage is the complexity in development. The information system department cannot just buy a data warehouse; the information system department has to build one because each warehouse has a unique architecture and a set of requirements that spring from the individual needs of the organization. Builders need to pay as much attention to the structure, definitions, and flow of data as they do to choosing hardware and software. Data warehouse construction requires a sense of anticipation about future ways to use the collected records. Developers need to be aware of the constantly changing needs of their organization's business and the capabilities of the available and emerging hardware and software. In summary, developing such a large database requires an expert.
- ***Takes time to build***. Second, 32 (7 percent) articles point out that to build a data warehouse takes time (2 to 3 years). In a situation where there is not strong executive sponsorship, information system directors or others wishing to develop a warehouse may spend an inordinate amount of time justifying the need.
- ***Expensive to build***. Similarly, 17 (4 percent) mentioned that a data warehouse is also expensive to build (2 to 3 million US dollars). One reason data warehouses are so expensive is that data must be moved or copied from existing databases, sometimes manually, and data needs to be translated into a common format.
- ***Lack of API (Application Programming Interfaces)***. Ten (2 percent) articles suggest that data warehousing software still lack a set of application programming interfaces (API) or other standards that shuttle data smoothly through the entire data warehouse process, such as Open Database Connectivity (ODBC) interface

(Microsoft Corp.). However, ODBC API that lets personal computers access data from many different databases is not everywhere.

- ***End-user training***. Seven (2 percent) articles suggest it is necessary to create a new "mind-set" with all employees who must be prepared to capitalize upon the innovative data analysis provided by data warehouses; those end users require extensive training. A communication plan is essential to educate all constituents.
- ***Complexity involved in symmetrical multiprocessing (SMP) and massively parallel processing (MPP)***. Six (1 percent) of the articles point out the complexity of data warehousing, which will be increased if the data warehouses involve symmetrical multiprocessing (SMP) and massively parallel processing (MPP). Synchronization and shared access are difficult.
- ***Difficulty in distributed database environment***. Because the data warehouse is a method of bringing disparate data together, it is centralized by its very nature and this is mentioned in 5 or 1 percent of the articles. While many companies are still in the preliminary stages of putting their data warehouses together, this centralization means only workers located at the same site as the warehouse have access to the data.
- ***Time lag between data warehouses and operation***. Lastly, in 3 (1 percent) of the articles, it is said that the data in data warehouses is extracted from operational databases that are continuously changing. A real-time data warehouse is an oxymoron because it is impossible to have real-time replication while maintaining a full-scale data warehouse. Data warehouses store only a time slice of corporate data that is steadily drifting backward out of relevance until the warehouses are replenished.

From my point of view, I think that the data warehouse technology will bring advantages and disadvantages for the Ministry of Health, so the Ministry of Health need to be aware of the disadvantages.

3.13 Deadly Sins of Data Warehousing

The ‘deadly sins’ of data warehousing is an invaluable map through the minefield for anyone setting out to design, construct, and implement a data warehouse project. Based on Barker extensive consulting experience, helps anyone involved in a data warehouse project to recognize, correct, and prevent some of the most common mistakes and misperceptions. The deadly sins presented by Baker (2000) are the following:

Sin 1: It is important to clearly define the business objectives, sponsor, and champion and develop a modular design / implementation approach. It is important that internally truly understands how difficult the project is, make sure that the data exists for the project and clearly define how the data is to flow into the data warehouse.

Sin 2: Failure to develop and design a data warehouse architectural framework. Factors to consider in determining a framework – the diversity and volume of data, the storage, twenty other things that you can think of and million you can not think of at the moment.

Sin 3: Underestimating the importance of documenting assumptions and conflicts early – How much data should be loaded initially into the data warehouse? What is the level of granularity (detailed or summarized) and how often must data be refreshed (Daily, weekly, etc)?

Sin 4: Abuse of methodology and tools. Methodology – there is no one right answer but some factors to consider: the approach used to acquire the information, corporate culture, operational database systems being included, toolkits available and data model adopted for the warehouse *per se*.

Sin 5: Data warehouse life cycle abuse. A data warehouse does not have an end-point like the system development. The waterfall model is not a good one.

Sin 6: Ignoring the resolution of data conflicts. Identify key files and systems.

Sin 7: Wasting knowledge learned from each project. Phase in the project so smaller can be used to build knowledge for more advanced phases.

3.14 Tools for Data Warehouse Implementation

This section presents a brief description of tools used in the different phases of data warehouse construction. Barker (2000) conducted a research that, as a result presents four groups of data warehousing tools:

***Analysis Tools**:* used to study the current operational database systems, to identify the requirements, primary data source for information during acquisition and building the data model (E.g. Bachman Analyst Tool).

***Development Tools**:* used during code generation for the information acquisition, data cleansing, data integration and loading (E.g. Oracle Tool).

***Implementation Tools**:* used for cleansing, consolidating and loading data, some of these tools can be developed in-house. Data acquisition tools used to gather and clean data (E.g. IBM and Oracle) and information store tools used to load data into data warehouse (E.g. Oracle and Prism)

Delivery Tools: assist in data conversion, data derivation, data loading and reporting on the delivery platform. Data loader converts data from the host computer to the delivery platform (E.g. Oracle Developer and Structured Query Language), data glossary describes in business terms what data is on the data warehouse (E.g. Lotus Notes) and querying and reporting are on-line and batch reporting facilities.

The research presented by Baker shows that it is difficult to choose a tool for data warehouse implementation. Each data warehouse phase uses a specific tool. We need to choose carefully the tool to use; otherwise we will comprise the process inside the phase. Different vendors have different tools for different phases of data warehouse implementation. It seems that, based on the literature review, Oracle constitutes a good tool to use during the data warehouse implementation. Domenico (2001) presents a research on data warehousing area. In the implementation part, he uses Oracle Data Mart Suite®. He chose Oracle tool based on the following reasons:

- It is a tool that is easy to learn (he took three months to implement the project).
- The tool supports the life cycle of data mart, from project until end user results analysis.
- He argues that it is important to look to the infrastructure within the organization where the data warehouse will be implemented. The Oracle environment was already implemented at Universidade do Oeste de Santa Catarina – Brazil, so it was easy to familiarize with the tool.
- *Oracle Data Mart Suite*â provides different software needs to prototype construction and holds the following tools:
    - *Oracle Data Mart Designer*â: Tool used to project the data mart.
    - *Oracle Data Mart Builder*â: To extract and transform data from operational/legacy systems.
    - *Oracle 8 Enterprise Server*â: Database server.
    - *Oracle Web Server*â: Allows Intranet to have access to data mart.

- *Oracle Discover*^{â}: Tool used for queries, reports and data analysis.
- *Oracle Reports*^{â}: Used to produce and deliver reports.

3.15 Medical Data Warehouse

An approach for integrating heterogeneous information sources in a medical data warehouse is a research conducted by Kerkri et al. (2001). The authors present

> "a medical data warehousing methodology that aims to use data semantics to regroup and merge patient medical data from different health information systems, which may be autonomous and heterogeneous. The proposed solution takes into account European laws concerning the security and anonymity of personal data".
>
> (Kerkri et al., 2001, p.167).

The aim of the research was to

> "define a data warehouse framework to regroup patients' medical information from various health structures at a regional level and to integrate them in a comprehensive information system. The data in the warehouse provide access to relevant data concerning the patient such as his/her antecedents and risk factors, in order to enhance diagnosis and medical decisions. This information stored in the warehouse can also be used for epidemiological and medico economic studies. The fusion of medical data is guided by a semantic data warehousing methodology that facilitates integration and migration of data from distributed and heterogeneous systems".
>
> (Kerkri et al., 2001)

To perform semantic integration, Staccini et al. (1998) propose a pragmatic way to describe the semantics of the elements of a database, based on a bottom-up three step process:

1) A back documentation of the elements of the system from their description contained in the data catalog of the database;
2) A first semantic extension to transform a data catalog into a data dictionary;

3) A second semantic extension to create a dictionary of the medical concepts from a data dictionary.

Their approach focused on the design of the structure of a dictionary able to describe the data catalog of a database and to extend it with the semantics expressed by the conceptual level, that is, the semantics of the domain of end users.

3.16 *Epidware*: A Medical Data Warehousing Framework

The *Epidware* project is being implemented in the Burgundy region in France and aims to obtain an economic assessment of medical strategies and developing epidemiological studies such as evaluation of the impact of risk factors on cardio-vascular mortality, taking into account the hospital type.

Epidware is an integrated system for providing access to a collection of heterogeneous medical information systems. It is based on two evolving information design methodologies:

(1) Data warehousing; and
(2) Database integration and inter-operation.

Data warehousing, the implementation process of data warehouse, is a progressive process that can be carried out in two main steps. First, data marts, which are data warehouses related to a given department or activity, are implemented. Secondly, two types of developments are possible according to the organizational choice of the enterprise or institution:

(1) Progressive centralization of strategic data; or
(2) Decentralized implementation of service-specific data marts.

The second methodology used in *Epidware* is the integration or inter-operation of information systems. The first step is to create a target schema that defines the overall structure of the data that are merged to create the data mart or warehouse. To model the target data structure, the designers must establish, with the help of the enterprise or institution, a dictionary of data descriptions (metadata) and specify tools to extract, translate and integrate data from the different data or information sources (the initial databases). This requires integration techniques to reconcile semantic differences among the schema or data structures of the underlying databases. Once the target schema is implemented, the data warehousing must be update to ensure that the format of data corresponds to the user's need, which may change later on. Furthermore, changes in sources of information have to be propagated on the data warehousing.

Finally, data from source databases are extracted, aggregated and filtered to harmonize their formats and to eliminate redundancies. Before loading data in the data warehousing, coherent values are assigned to variables or fields that cannot be initialized from the local systems. The general architecture of the *Epidware* integrated system consists of information systems at the lowest level, a group of components, called wrappers, at the intermediate level, and the integrator and data warehousing at the top level.

As a conclusion, the research presents the general architecture of *Epidware* to build data warehouses for epidemiological and medical studies. The authors' solution could be useful at several levels, first to learn from a running project in the health care area and secondly to follow some steps for data warehousing implementation.

3.17 Summary

In summary, this chapter presents a general vision of data warehouse technology, its basic concepts, the basic elements of data warehouse, data warehouse characteristics, data warehouse architecture, the data model, data mart approach, data warehouse challenges, advantages and disadvantages, tools for data warehouse implementation and a case study of a medical data warehousing framework. The main data warehouse elements include from data sources (operational or legacy systems) to applications in constructed environments. The construction architecture of the data warehouse environment appears either from a top-down approach where the data warehouse is completely constructed and data marts are extracted from data warehouse, or from a bottom-up approach where the data warehouse environment is the result of data marts integration.

It is important to point out that the data warehouse environment has become a necessity for organizations. Data warehouse can manage all historical data and information of one organization. Given that information in our days plays a big role within them, they should use data warehouse as a tool to store and analyze information and then make important decisions based on their own well-organized data.

For the data warehouse construction process the Ministry of Health needs to look at the following key points:

- *Data sources definition*. Select the systems that will provide data to the data warehouse. These systems should play an important role in decision-making process;
- *Documentation and analysis of the existing legacy systems, with focus on the quality data.* This is probably the most important task in the process of the data warehouse construction. Special attention to the quality of the data needs to be taken.
- *Steps of data warehouse life cycle to be followed*. To construct the data warehouse the Ministry of Health needs to follow the life cycle as a methodology to implement the data warehouse adapting the life cycle to the health sector.
- *Data warehouse architecture and model*. The challenge is in the design and development of data warehouse architecture model, because identifying user requirements is a big challenge for analysts and designers of data warehouse. The research shows that bottom-up approach is the most used. From the literature review, some authors stress that this approach is adequate for the organization within an economic, cultural and political point of view.

Chapter 4

The Health Information Systems in Mozambique

This chapter presents a brief view of the Health Information System in Mozambique and is an empirical work based on the literature review and on the fieldwork research conducted in two districts of Gaza Province – Chibuto and Manjacaze and in the Ministry of Health. This chapter, presents an overview of Health Information Systems, describes the decision process and the health structures in the Ministry of Health, describes the different levels of healthcare in Mozambique and presents the information that flows from one level to another. The different automated systems are also described here.

4.1 Profile of Mozambique

General Background

Mozambique is situated in Southern Africa, bordering the Mozambique Channel, between South Africa and Tanzania. The country shares borders with Zambia 419 kilometers, Malawi 1,569 kilometers and Zimbabwe 1,231 kilometers in the West, South Africa 491 kilometers and Swaziland 105 kilometers in the South, Tanzania 756 kilometers in the North, and the Indian Ocean in the East 2,470 kilometers.

Figure 4.1 shows a map of Mozambique.

Figure 4.1: Map of Mozambique.

Mozambique covers a total area of about 802 thousand square kilometers. In the year 2000 the population estimate was 17.2 million, distributed in 11 provinces. The majority of the population lives in the cities. About 50 % of the population is under 15 years old. The population growth rate is 1.13 % and the life expectancy is 35.46 years.

Infant mortality is 101 deaths/1,000 live births. In the Southern Africa, other countries have the following rates per 1,000 live births: Madagascar (68), Malawi (130), Tanzania (72) Zambia (90) and Zimbabwe (60). GDP per capita is US$900 dollars in 2002, and the literacy rate is 43.2 %. Mozambique is a country that has been plagued with instability in recent years. For nearly 25 years Mozambique struggled a war and for a little more than 10 years now it has wrestled with peace. The years of war destroyed health posts, area clinics and rural hospitals. It disrupted communication infrastructures in different provinces and it forced people to leave the areas they once called home (Gomes and Johnson, 1994).

In October 1992, RENAMO (*Resistência Nacional Moçambicana*) and the Government of FRELIMO (*Frente de Libertação de Moçambique*) signed up a peace agreement in Rome, Italy, that led to the first democratic elections in October 1994. The transition from war to peace (initiated before the signing of the peace agreement in 1992) focused on country reconstruction. The health sector, supported by many international donor agencies, has been particularly active in this process, which is still under way.

After the peace agreement, management tasks took over planning ones. Officials were totally absorbed by the new pressing demands of reconstruction, reconciliation, and service expansion. No manager had the time or energy to concentrate on planning issues. The result was that the plans finalized by 1992 were rather successfully implemented, whereas areas lacking a clear Ministry of Health policy were left open to many uncoordinated and inconclusive donor-led initiatives. For instance, many initiatives, in turn promoted by UNICEF, USAID, the World Bank, were launched in the management reform area, without significant success. The Ministry of Health was hesitant, while external agencies were competing for influence (Pavignani, 2002).

In terms of health problems the epidemiological profile of Mozambique is similar to any other tropical developing country. Malaria, cholera, tuberculosis, AIDS and malnutrition are all of great concern. In Mozambique, malaria infection rates were estimated at 4.5 per cent in the cities, and 7.4 per cent in the rural areas, 12,317 cases of cholera reported in 2000 and 168 deaths. In terms of tuberculosis, Mozambique has a prevalence of tuberculosis greater than 3.5 % (1996-1999), AIDS prevalence in 2001, show a 5.7 per cent prevalence rate in the north (Nampula, Cabo Delgado, Niassa), 13.2 per cent in the south (Maputo city, Maputo province, Gaza and Inhambane), and 16.5 per cent in the centre (Sofala, Manica, Tete and Zambezia). Mozambique has 36% of children under five chronically malnourished.

Despite rapid economic growth in recent years, Mozambique remains a very poor country, with approximately 70 per cent of the population living in poverty (Govern of Mozambique, 1998, hereafter GoM, 1998). These expenditure-based poverty measures are reflected in widespread food insecurity and poor health indicators. Sixty-four percent of all Mozambicans (over 10 million people) live in food insecure households (GoM, 1998). The prevalence of malnutrition implies higher risk of mortality and morbidity, and of retarded physical and cognitive development.

Poor environmental conditions and extremely limited access to water, sanitation, and health services also contribute to poor health status. Sixty-six percent of urban dwellers have access to a generally "safe" water source, like piped water, while only 12 per cent of rural dwellers do. A large percentage of the population gets water from a public or private well, but lakes and rivers remain an important source of water, particularly in the central region of the country. Only 21 per cent of rural households have latrines. There is, however, considerable regional variation, and latrines are generally more common in urban areas (GoM, 1998).

“In the health sector, the war exacerbated the existing problems and brought additional ones. The insurgent army consistently targeted the health system where it was more vulnerable, i.e. in rural areas, thereby wiping out past gains" (Cliff and Noormahomed, 1988).

Health workers were murdered and kidnapped. "Endangered supply lines became unreliable and prohibitively expensive. The already weak referral and supervision systems broke down. The health network shrunk to a cluster of facilities, mainly large, located in secure areas and defended by the army, and overstaffed, as health workers, like the ordinary population, concentrated there in search of safe havens. By the mid-eighties, the National Health System had become totally dependent on external aid" (Noormahomed and Segall, 1992).

Whereas the support provided by the Eastern block decreased, western donors became prominent. Besides the genuine humanitarian motivations, a quite explicit political agenda explains the substantial involvement of many donor agencies in the Mozambican context. Consequently, the government reconsidered its political and economic choices, switched alliances and endorsed western economic models. This culminated in the launching of a structural adjustment programme in 1987 (Hanlon 1991).

The last years of war (from the mid-1980s to 1992) were emergency-dominated: many agencies, largely uncoordinated, started operations in the country. The health sector was flooded by "easy" money, which stimulated inefficient approaches to service delivery. Large agencies, such as the European Union (EU) and the UNHCR (United Nations High Commissariat for Refugees), managed significant amounts of relief funds and became the main financiers of non-governmental organizations' (NGOs) projects. Under the pressure of speeding operations up, criteria for funding relaxed. Summary proposals were approved rapidly and after only superficial scrutiny. Monitoring and evaluation of funded projects were even scantier, if carried out at all.

4.2. The Ministry of Health

The Ministry of Health is the Government's main agent and advisor on health. It develops policy advice for the Government on health, administers health regulations and legislation, funds health and support services, plans and maintains nationwide frameworks and specifications of services, monitors sector performance and provides information to the wider health sector and the public. The Ministry of Health plays a major role in educating and providing information to the public on how they can maintain good healthy lifestyles. The Ministry also plays a key role in reducing illness in Mozambique through the control and prevention of diseases and ensuring that the resources are allocated appropriately to do this.

4.2.1 Health Structures in the Ministry of Health

The Ministry of Health is composed of the Minister's Office, assisted by a Vice-Minister and advised by the Health Consultative Council and Technical Council. The Ministry of Health contains four national directorates. Each directorate has two or more departments, and sections. These national directorates are as follows: Administration and Management, Planning and Cooperation, Human Resources, and National Health. The National Directorate of Health is responsible for managing health programmes, community health, medical care and epidemiology. The Planning Directorate has the responsibility of making long-term planning, statistical collection and analysis of health data. This directorate is subdivided into various departments, of which the Health Information Department (HID) is one of them. The Health Information System is managed by Health Information Department through its Health Information Directorate (Macome, 1992).

Each province has a provincial health authority and the provinces are divided into districts, each with a director of health. At provincial level, there are different sections, one of them being a nucleus which deals with health information system and statistic analysis, and it is called Núcleo de Estatística (NE). Figure 4.2 presents the health structures inside the Ministry of Health (MoH).

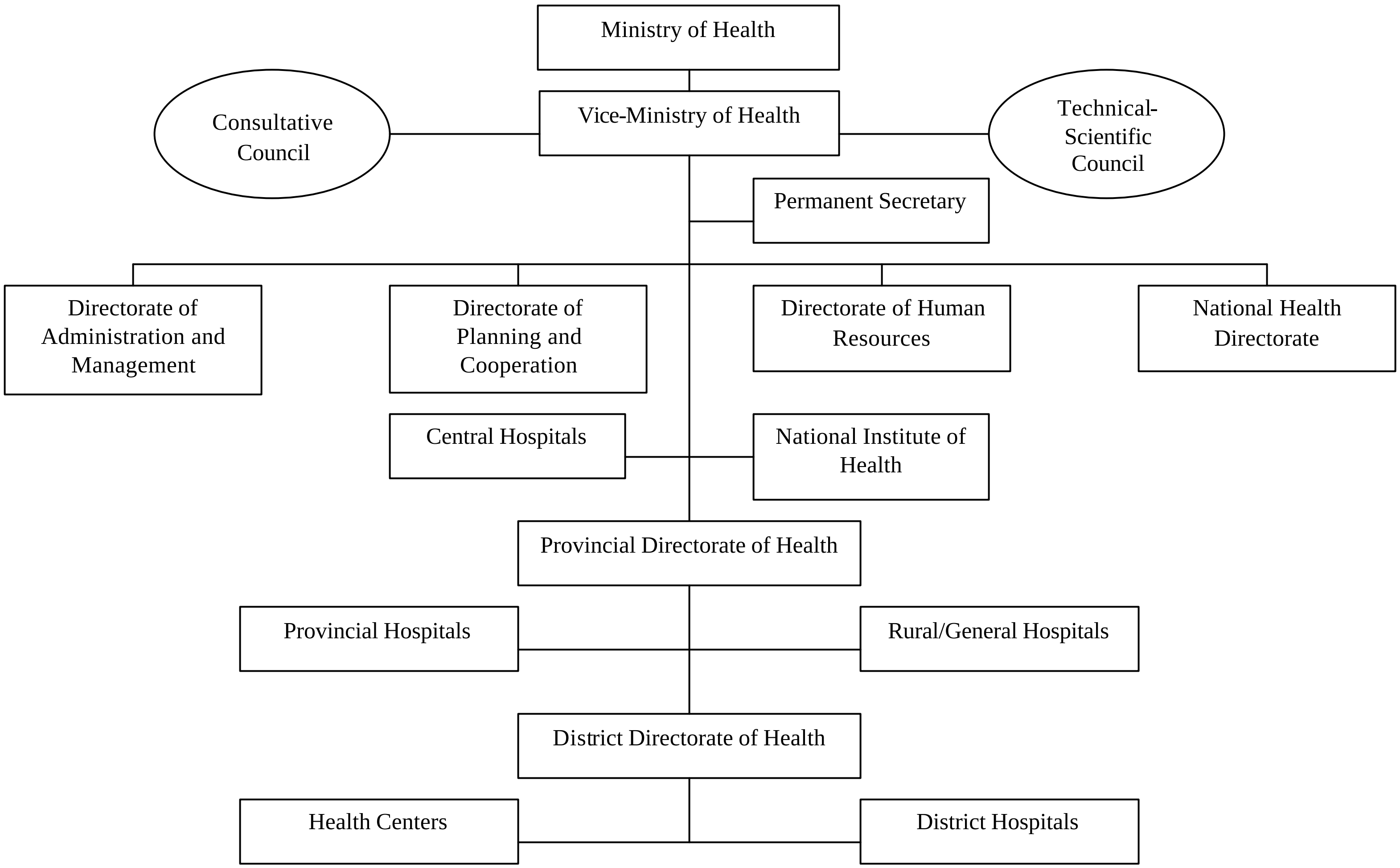


Figure 4.2: Health structures within the Ministry of Health in Mozambique

(Source: Ministry of Health/Directorate of Planning and Cooperation, May 2002).

4.2.2 Decision Processes at the Ministry of Health

The Ministry of Health has four levels of health system and three types of management, as shown in Table 4.1.

The Ministry of Health has three management functions: Sector/System, Health Facilities/Health National Service and Patient/Client Management. These management functions are implemented at four levels of concentration: primary, secondary, tertiary and quaternary.

Table 4.1: Types of management.

	Management Functions		
Concentration Level	**Sector/System Management**	**Health Facilities/Health National Service Management**	**Patient/Client Management**
QUATERNARY	National (Central Office)	Hospital (Central Hospital Level IV, Teaching Institution)	Referred or urban Patients/Clients
TERTIARY	Provincial (Provincial Directorate of Health)	Hospital (Provincial Hospital Level III)	Referred or urban Patients/Clients
SECONDARY	District (District Directorate of Health)	District Hospital (Rural/General Hospital Level II	Referred or urban Patients/Clients or from rural/urban district
PRIMARY	-	Health Facilities (Health Center, Health Post Level I)	Patients/Clients from Community
	Strategic Plan, Operational and Control Decisions.	Operational and Executive Decisions.	

Source: Ministry of Health/Directorate of Planning and Cooperation, February 2003).

4.3 The Sistema de Informação de Saúde (SIS)

Overview

The Ministry of Health, recognizing that one correct, complete and useful information system constitutes an essential element for an efficient allocation of different (human, material and financial) resources, planning and management of health services decided to define, for the first time in 1978/80, i.e., the Sistema de Informação de Saúde (SIS) in Mozambique. Between 1978 and 1980 the first uniform, nationwide health information system was instituted (Gomes and Johnson, 1994). Subsequently, using World Health Organization (WHO) technical assistance, a number of changes were made which resulted in the introduction of an operational national health information system in 1982. The intention of the National Health Information System was to assist with, and improve, the planning and management of health care services at all levels (Gomes and Johnson, 1994 and Macome, 1992). In a recent study, Braa *et al.* (2000), defined the Sistema de Informação de Saúde (SIS), the Mozambican health information system as:

> "a collection of data collection forms, procedures for reporting and aggregating data from the health facility, to the district, province and national department of health, and a database application being used at province and national levels".
>
> (Braa *et al.,* 2000, p.6)

The Sistema de Informação de Saúde (SIS) serves the following purposes within the Ministry of Health (Gomes and Johnson, 1994):

- Collection of routine data on health services;
- Processing data at an appropriate intermediate level in the system and generating the required management reports, performance indicators and statistical series needed by management personnel at each level;
- Transmitting the report or data both upward and downward in the system at the required frequency; and

- Maintaining a facility level deposit files consisting of summary reports from the low level and summary reports sent to the higher level.

Macome (1992) argue that the Sistema de Informação de Saúde (SIS) has several problems in the functioning:

> "The health data have been collected by health workers untrained in their collection, purpose or use. Despite the longevity of the Sistema de Informação de Saúde, it was widely viewed within the Ministry of Health as being inefficient as it produce data which were unresponsive to management needs, inaccurate, untimely and under-utilized by management personnel at all levels especially at the low levels of the health system".
>
> (Macome, 1992, p.14)

These were some factors that contributed to the unsatisfactory state of the Sistema de Informação de Saúde. As a result of this, since 1989 a number of steps have been taken to improve the Sistema de Informação de Saúde and in 1991, the Ministry of Health concluded the revision and improvement of the Sistema de Informação de Saúde. During revision of the Sistema de Informação de Saúde, the number of data collecting forms was reduced from 60 to 12 and some basic indicators were included in the forms for data collection, for use at the district and health facility levels. Most of the data collecting forms were based on the health programs and services. The introduction of the new version in all provinces concluded in November 1992.

The Sistema de Informação de Saúde encloses the primary (health post and health center) and secondary (general and rural hospitals) levels and covers the epidemiological information, consumable resources (essential medicines, contraceptives and vaccines) and some non-consumable resources (personnel, infrastructure, beds and equipment for the Expanded Program of Immunization (EPI)). Data are collected at health units, aggregated and collated at district level and transmitted to provincial and national levels.

I conducted a research in Gaza Province and at the Ministry of Health. Interviews, observations and analysis of documents such as data collection forms, reports and analysis of the existing technological and medical infrastructure were the research methods used. In the next section, I describe the structure of the Sistema de Informação de Saúde in Mozambique.

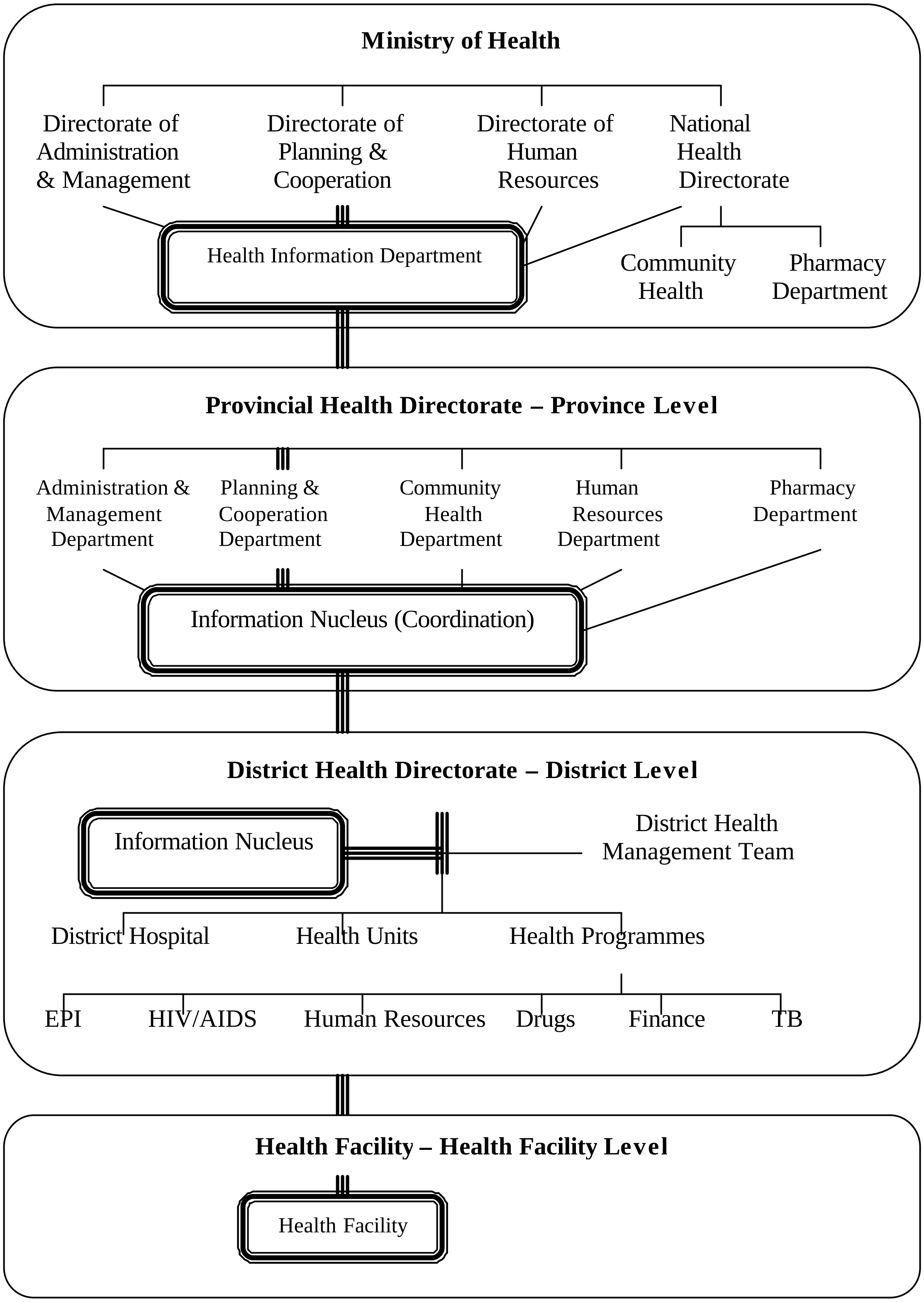


Figure 4.3: Organizational structure and main information flow.

(Adapted from Braa *et al*., 2000).

Key: ≡ - Represents the downward and upward main information flows.

The triple line means that the data goes from the health facilities at health facility level to the Núcleo de Estatística (Statistics Team) at the district level, then to the Núcleo de Estatística (Statistics Team) at the province level, and finally to the Health Information Department at national level, the Ministry of Health.

Figure 4.3 gives an overview of co-ordination and the main flow of information within the health facility, district, province and national levels of the health system in Mozambique. The information flow as presented in Figure 4.3 can be summarized as follows:

1. Health workers (doctors, nurses and administrative personnel) collect data at health facilities using SIS forms. They aggregate data and then send to district level.
2. Health workers at the Núcleo de Estatística (NE) in the district receive forms from health facilities; they aggregate in new forms and a report with a district overview is produced.
3. At the provincial level, data from the district are entered into the database and then transferred electronically to the national level quarterly.
4. At the national level data from all provinces are stored in the database.

4.4 Information Flow Among Different Levels of Sistema de Informação de Saúde

There are four levels of health care: primary (health facility level), secondary (district level), tertiary (province level) and quaternary (national level).

> "The primary level cares for about 60 to 65% of the entire population through the health centres and health posts. The secondary level corresponds to the districts (rural and general) hospital. The tertiary level corresponds the provincial hospitals and

> the quaternary level concerns units at national reference level (central hospitals)".
>
> (Macome, 1992, p.11)

In the following section, I present a description of each level of health care.

4.4.1 Health Facility Level

The health facilities are the origin of the health care data and statistics being generated in the health system. Health workers at the health facilities as well as in the community collect data on Immunization, Mother and Child Health, Family Planning, Medicines and outpatient statistics. In each health facility we can find staff, e.g. nurses that are responsible for all data generated at the health facility.

To maintain a record of services provided, workers in the SIS use:

- **Tick registers –** where there is a need for patient identification with services provided;
- **Tally sheets** – where important conditions and services not needing follow-up provided are recorded with a mark, with no patient identification;
- **Program registers** – for expanded programs such as Immunization, antenatal care, family planning, tuberculosis;
- **Reports** – which contain the selection of services provided, sent to the district each month.

The data on Mother and Child Health (MCH) and Sistema de Informação de Saúde (SIS) are aggregated into a monthly report and then sent to the district, while the data on Expanded Program of Immunization (EPI) are sent to the district on a daily basis (without being aggregated) and data on Weekly Epidemiological Bulletin (BES) are sent to the district on a weekly basis.

Different forms, such as A03 (the monthly summary of EPI for health facilities) or B04 (Nutritional Vigilance, monthly summary of Mother and Child Health for health facilities) are being used at the health facilities for data collection and then the report is produced in a different reporting form. For example, for Mother and Child Health, there are three different forms for data collection which are then combined into a single reporting form

At health facility level, health facilities have paper-based record keeping. During the aggregation process at the end of the month, the nurse acquires data from the forms filled on a daily basis and produces the report of the health facility. Then he submits the report to the responsible of the health facility who sends the report, after analysis, to the next level, the District Directorate of Health.

Figure 4.4 presents an overview of the health information system at health facility level:

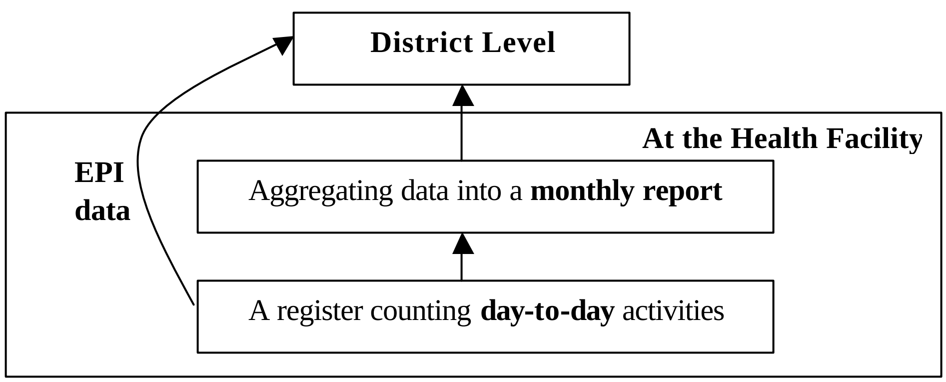


Figure 4.4: Overview of the Sistema de Informação de Saúde at Health Facility Level.

Table 4.2 summarizes the information that flows from the health facilities to District Directorate of Health on a daily, weekly and monthly basis.

Table 4.2: Detailed forms submitted by health facilities to the District Directorate of Health in different period basis.

Period	Activities
Monthly basis	SIS Summaries to District Directorate of Health (DDH): • EPI (A03). Mother and Child Health visits, Epidemiological Surveillance, Family Planning and Maternity, Outpatient, medicine kits A and B (C04 and C05); • Laboratory services (E02), Oral Health Services (G02); • Anti-conceptive stock (B05); • Inpatient for health centers and posts, pediatrics, hospitals summaries (D04). Summaries from other subsystems: • Three (3) forms of monthly notification (STD/HIV); • One (1) Monthly Epidemiological Bulletin (BEM) with list of diseases (not SIS).
Weekly basis	• Form to count new cases of illnesses of obligatory declaration; • Epidemiological Weekly Bulletin – BES (C03) (Malaria, measles, convulse cough, acute limp paralysis, cholera, dysentery, anger, plague, meningitis).
Daily basis	• SIS registration book: outpatient, Inpatient, chirurgic Inpatient and maternities; • STD/HIV Book (not SIS); • Register forms and control forms of SIS: EPI (BCG, DPT, measles and hepatitis), Tetanus (VAT), form to control the temperature of cold chain, pre-natal consultation, child consultation of 0-4 months and nutritional vigilance (35 months).

Source: Ministry of Health/Directorate of Planning and Cooperation/2002

Fragmentation of information at health facility level can be a major problem, not only for the health facilities, but also for the Ministry of Health. Health information sent to the district level is not integrated with other types of information, such as material, human or finance resources used or budget allocated for a specific activity.

4.4.2 District Level

In order to understand district level, the study was conducted in two districts of Gaza province: Chibuto and Manjacaze.

Chibuto District

Chibuto district consists of six administrative posts – Chibuto, Malehice, Chaimite, Godide, Alto-Changane and Changanine – and 25 localities. The district government is composed of directorates of Agriculture and Fisheries, Industry, Commerce and Tourism, Education, Planning and Finance, Public Works and Housing, Health, Labor, Culture, Youth and Sport, and Social Welfare. Apart from these institutions, there are also the police, the telecommunications company, the civil registry office, the court, the post office and the state housing department. The coordination of government activity is undertaken through regular meetings. According to the numbers provided by the district administration, the population of Chibuto district is estimated at 250,120 inhabitants (Ministério da Administração Estatal (State Administration Ministry), 2002b).

The district has a rural hospital with 69 beds and a rehabilitated health centre, which has no beds. Both are in Chibuto city. The health centre has a fixed vaccination post. There are also five health posts, in Chibuto city and in the localities of Malehice, Chipadja, Alto Changane and Chaimite. The Malehice, Chibuto, Alto Changane and Chipadja health posts have maternity units, but only the Chipadja post has three beds. All the health posts have fixed vaccination posts. Chibuto health post was built with funds from the general state budget, and the one in Malehice was built by the Catholic Church. The other three health posts have been rehabilitated by MSF-Switzerland (Alto Changane) and by the micro projects programme of Rural Development Institute – INDER (Chaimite) (Ministério da Administração Estatal (State Administration Ministry), 2002b).

The Provincial Health Directorate distributes medicines and medical materials to rural hospitals, health centres, and health posts on a monthly basis. They receive A, B and C type kits in unspecified quantities. Health workers in Chibuto told the researcher that they regard these supplies as insufficient for local needs, particularly for the fight against malaria and diarrhea diseases.

The District Directorate's Health annual budget covers expenditure on maintaining and repairing infrastructures and vehicles, but not wages or the purchase of essential medicines and fuel.

Table 4.3: Health sector staff in Chibuto.

Staff	Number
Doctor	2
Medical Assistant – Middle	15
Basic	20
Elementary	6
Community health Worker	N.A.
Traditional Midwifes	N.A.
Orderly	22
Total	65

Source: Ministry of Health/Chibuto district, 2002.

As we can see Chibuto has a serious problem of Human Resources for the healthcare sector. This is a country problem.

Manjacaze District

Manjacaze (also known as Mandlakazi) district consists of seven administrative posts and seventeen localities. The various state institutions of state present in the district include District Directorates of Education, Agriculture and Fisheries, Commerce, Industry and Tourism, Health and Social Action, as well as the judiciary, the police, the civil registry and the security service. The district administration calculated the population in 1996 at 214,952 inhabitants in this district. The area of the district is 3,748 square kilometers, giving a population density of 57 inhabitants per square kilometer (Ministério da Administração Estatal, 2002a).

Manjacaze has a rural hospital with 78 beds. There are two health centres, one with no beds and the other in Tavane with 24 beds. They also function as fixed vaccination posts. In addition, there are 10 health posts, at Mausse, Macuacua, Chidenguele, Dingane, Macupulane, Mangonde, Matsinhane, Manhique, Mussengue and Cucuine. Medicines, equipment and medical supplies from the Provincial Directorate of Health are distributed to the hospital, health posts and centres, although the district authorities note that the quantities received have not been enough to meet demand and transport problems delay deliveries. The annual budget for the District Directorate of Health in 1996 was MTS 153,000,000,00 (US $13,734 in 1996) (Ministério da Administração Estatal, 2002b).

Table 4.4: Health sector staff in Manjacaze

Staff	Number
Doctor	2
Medical Assistant – Middle	15
Basic	20
Elementary	6
Community Health Worker	N.A.
Traditional Midwife	N.A.
Orderly	22
Total	65

Source: Ministry of Health/Manjacaze district, 2002.

As we can see, Manjacaze also has a serious problem of Human Resources for the healthcare sector. This is a country problem.

The Núcleo de Estatísticas is present in every district, consisting of one group of health workers, who were responsible for data collection from the health units, collation to get an overall district picture and analysis of basic indicators before sending them to the province level. Only one person is responsible for each program in the district (e.g.:

Mother and Child Health, Family Planning and Medicines) reporting to the Núcleo de Estatísticas.

At district level, data from the health facilities are aggregated into single reporting forms and sometimes separated into different reporting forms. For example, a single reporting form received from the health facilities on Mother and Child Health (MCH) is being reported into three different reporting forms from the district to the Province. The reporting on mother and child health data from the district to the province is done per health facilities, i.e. data are not aggregated, while for Expanded Program of Immunization and Sistema de Informação de Saúde (SIS) data are aggregated. The person responsible for the SIS receives all the reports (for all the programs) from the health facilities and then he/she distributes them to each individual responsible person (for Mother and Child Health and Expanded Program of Immunization).

Each member of the Núcleo de Estatística team prepares the report on a particular program and then at the end of each month, they sit together to produce a district report, which has to be checked by the district health director before being sent to the province. After checking the report the District Directorate of Health Manager sends it back to the person responsible for SIS who then sends the report to the province.

Braa *et al*. (2002) conclude that, the information handling at district level consists basically of aggregating the forms from the health facilities into 'district forms', which are then submitted to the province.

In Gaza province, Manjacaze, Chibuto, Biline, Macia, Chokwé and Xai-Xai are District Directorates of Health that have computers. In these districts the Núcleo de Estatística personnel enter data in the computers. In Manjacaze and Chibuto, health workers introduce data in the District Health Information System (DHIS). "DHIS is a computer database that uses a combination of forms, procedures and analytical tools to convert routine anonymous data into useful management information that can be used by local program and facility managers" (Heywood and Rohde, 2002). In Mozambique, the

Ministry of Health started to use this tool in 1999 in Niassa province. With the District Health Information System they only need to introduce data and the application will generate different types of reports.

The flows within the district level can be presented as represented in Figure 4.5:

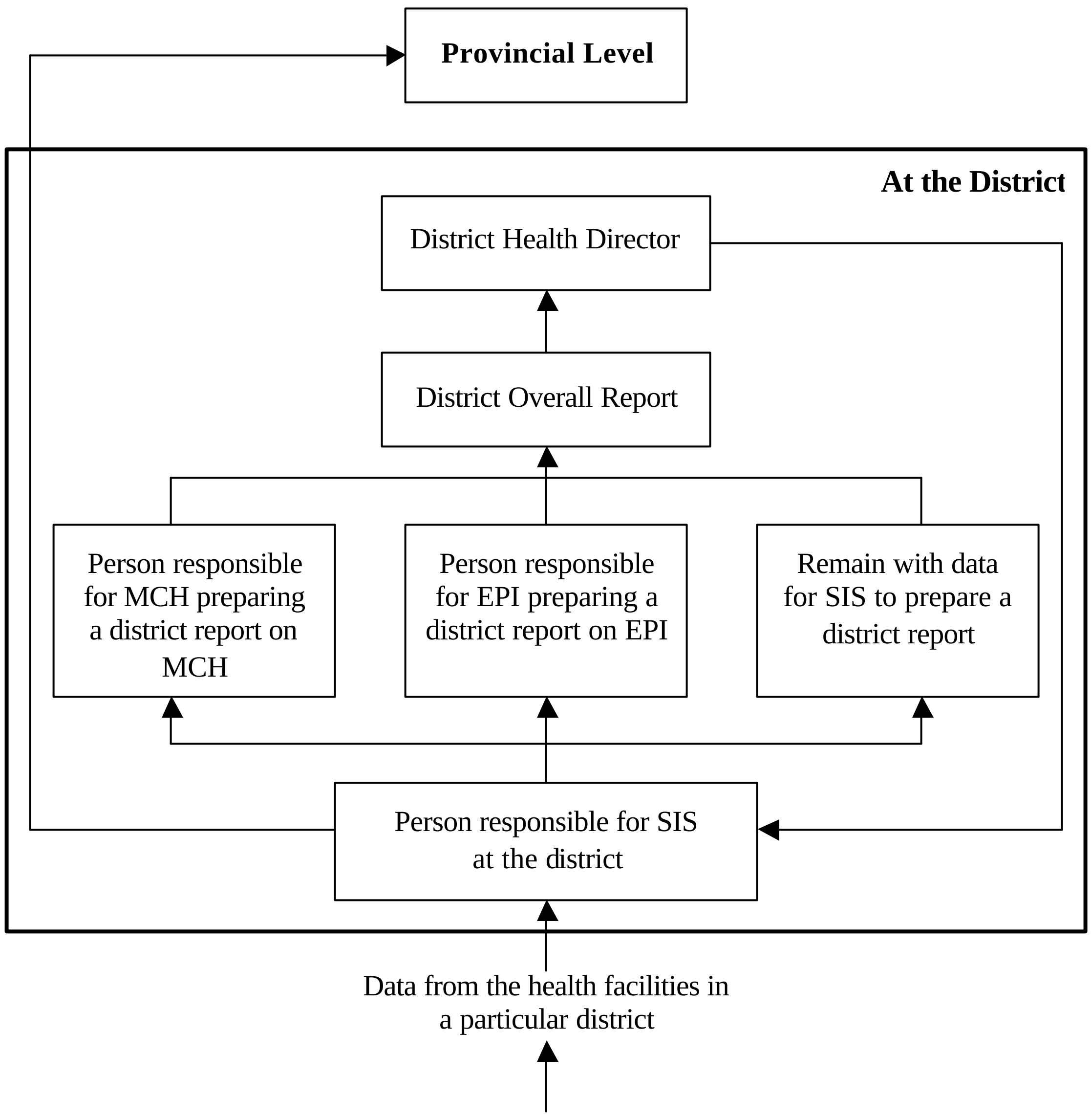


Figure 4.5: Overview of the Sistema de Informação de Saúde at District Level.

Fragmentation of information at district level occurs because different reports are sent separately to the province level. The main report does not integrate data from all areas or services, it is sent vertically without integration of information.

Table 4.5 summarizes the information that flows from the districts department of health to the province level on a monthly, quarterly and annual basis.

Table 4.5: Detailed forms submitted by District Directorate of Health to the Provincial Directorate of Health in different period basis.

Period	Activities
Annual basis	• Summaries of Sistema de Informação de Saúde (SIS) with data of EPI (A05); • Sanitary network and beds (J01); • Primary and secondary level Human Resources (K01)
Quarterly basis	• Summary of cases of tuberculosis and leprosy.
Monthly basis	**Summaries of SIS aggregated from all health facilities**: • EPI is aggregated per district; • Monthly summary of Mother and Child Health (MCH) (B04) is divided in three forms: MCH and Nutritional Vigilance (B06), Maternity (B07) and Antenatal visits, Postpartum visits and Family Planning (B08); • Medicines Kits A and B (C07); **Separated flow of data (not SIS):** • Finance; • Logistics; • Administration; • Management; • Pharmaceutics deposit.

Source: Ministry of Health/Directorate of Planning and Coopertion/2002.

4.4.3 Province Level

To better understand the province level, I conducted a study in Gaza province. The province has 12 districts: Massangena, Chicualacuala, Chigubo, Massingir, Mabalane, Guijá, Chokwé, Chibuto, Manjacaze, Bilene and Xai-Xai. Gaza has an area of 75,450 square kilometers.

The borders are: Manica to the north, Maputo to the south, South Africa to the west, Inhambane to the east and the Indian Ocean to the southeast. The main products are cashew, cotton, rice and maize. The population density is 16-inhabitants/square kilometer. The most represented ethnic group is Changana.

At province level, district reports are aggregated every three months whereby the three-months report is cumulative, i.e. three month, six months, nine months and annually.

The person responsible for the Sistema de Informação de Saúde (SIS) receives all the reports (for all the programs) from the districts, enters the data into the SIS electronic database, and then distributes the reports to each responsible person (for Mother and Child Health and Expanded Program of Immunization). Each health worker prepares the report on the particular program, while the SIS person also prepares an overall report for all programs (SIS, Expanded Program of Immunization and Mother and Child Health) and then they give the reports to the SIS person who then coordinates the reports by making comparisons with what was produced using computer tools. After coordination, the overall province report is sent to the national level, the Ministry of Health. When the provinces send the reports to the Ministry of Health they also send a diskette with data contained in the reports. The data in the diskette are captured to the SIS electronic database. Each quarter the province is to make a backup disk of its data and send this disk to the national level. In a recent study, Braa et al. (2000) show that this system of feeding aggregated data to the national level is functioning in all eleven provinces. The flows of information within the province are described in Figure 4.6.

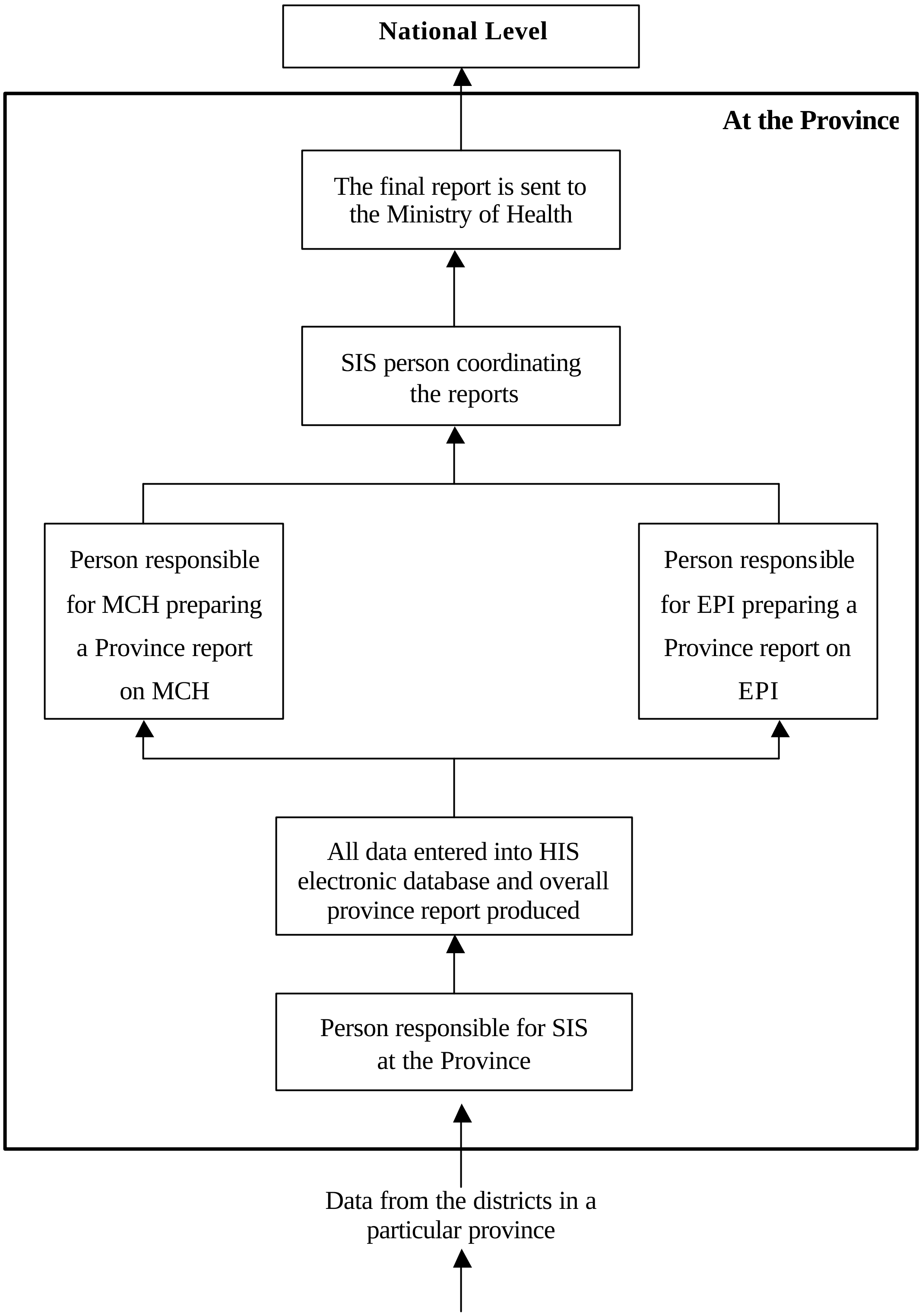


Figure 4.6: Overview of the Sistema de Informação de Saúde at Province Level.

Table 4.6 summarizes the information that flows from the Provincial Directorate of Health to the national level, in a weekly, monthly, quarterly, semester and annual basis.

Table 4.6: Detailed forms submitted by Provincial Directorate of Health to the national level (the Ministry of Health) in different period basis.

Period	Activities
Yearly basis	• Twelve (12) summaries (quarterly/semester) + 3 summaries (EPI resources (A05), sanitary network and beds (J05) and primary (rural and urban) and secondary level Human Resources (K02).
Semester basis	• Eleven (11) quarterly summaries + 1 summary of inpatients in health post and center (D05).
Quarterly basis	• Eleven (11) provincial summaries (five (5) summaries A06 to A10; four (4) summaries: B09, B10, B11 e B12; one (1) summary: D04 and one (1) summary: G04).
Monthly basis	• Compilation of archive forms coming from district monthly summaries. The compilations are kept at provincial level.
Weekly basis	• Weekly compilation of Weekly Epidemiological Bulletin (BES). Aggregated from all districts.

Source: Ministry of Health/Directorate of Planning and Coopertion/2002

Fragmentation of information at province level occurs because different reports are sent in a vertical way and are not shared in a horizontal communication.

4.4.4 National Level: the Ministry of Health

At the Ministry of Health, the Health Information Department (HID) coordinates the whole process started at health facility level.

After receiving the reports and diskettes from different provinces, the Sistema de Informação de Saúde electronic database operator at the Ministry of Health, uploads the data from the diskette to the Computerized Sistema de Informação de Saúde database and stores the reports.

All reports and summaries of information from every province will be collected at the Health Information Department and other central health departments or programmes are going to gather the information (before and after processing) from the Health Information Department (Macome, 1992). The information will be processed for the purpose of planning, monitoring and evaluation of all health development projects and distributed to concerned agencies, for example, United Nations Children's Fund (UNICEF), Save the Children, and other agencies. The Health Information Department is also responsible for coordinating with other information generating institutions, for example, the INE – Instituto Nacional de Estatísticas (National Institute of Statistics) that is responsible of all statistics in the country.

At national level, fragmentation occurs in the data collection systems, too often vertically organized.

4.5 Brief Description of Existing Computer Systems

"Governments and private institutions are all interested in having data collection systems that enable them to have reliable data in a timely basis. It is in the sense that computer assisted data collection systems are developed as ways to answer the problem of data collection" (Macome, 1992).

Presently, the Ministry of Health has several computerized systems running. The following systems are running in different departments.

- Computerized Sistema de Informação de Saúde (CSIS)
- Computerized Weekly Epidemiological Bulletin (CBES)
- District Health Information System (DHIS)
- Tuberculosis
- Malaria
- Pharmacy
- Human Resources
- Investment (GACOPI)
- Maintenance (Vehicles, Technology and Infrastructure)
- Integrated Plan and Management (SIMP)
- PESS (Strategic Plan for the Health Sector)
- Finance (Administration and Management)
- Logistics

In the following section I describe and analyze the following systems: Computerized Sistema de Informação de Saúde (CSIS), Computerized Weekly Epidemiological Bulletin (CBES) and DHIS (District Health Information System). I selected these systems because I had the opportunity to see and receive an explanation of how they function.

4.5.1 Computerized Sistema de Informação de Saúde (CSIS)

The Computerized Sistema de Informação de Saúde program was developed and deployed in 1992 – 1994 in Mozambique by a British engineer. The program was made with Clipper and data are stored in Dbase database version III. The program is a MS-DOS (Microsoft Disk Operating System) based application and does not have support for mouse device. Even though the program is DOS-based it can run in a DOS windows in Windows 95, but not Windows 98 or later.

The Computerized Sistema de Infomação de Saúde is an automated health information system part of National Health Information System and is being used at province and national levels. The purpose of the system is:

- Reporting and analyzing monthly health information;
- Storing of infrastructure information (all the health units, districts, provinces);

Important parts of Computerized Sistema de Infomação de Saúde are the paper-based forms. The structure of the database is build up around the paper forms. Furthermore, all the information in Computerized Sistema de Infomação de Saúde comes from the paper forms reported and aggregated in the lower levels. The reporting from health units to districts is done on specific forms; these forms are then collected and some of them are entered into new forms at district level to further be sent to the province level each month, where the values are entered into Computerized Sistema de Infomação de Saúde. The hospitals also enter data on specific forms.

The communication between the Computerized Sistema de Infomação de Saúde at the province and national levels is on a quarterly basis with diskettes containing parts of the province database. The Computerized Sistema de Infomação de Saúde also contains information about the infrastructure that is data that are not being reported monthly, but are updated as the health infrastructure changes. For each health unit the Computerized Sistema de Infomação de Saúde stores:

- The unique identification code and name;
- Parent district;
- Type of unit;
- Urban or rural status;

- Type of data to be delivered (Expanded Program of Immunization, Mother and Child Health, Maternity); and
- Infrastructure (power, water, cars).

For each district Computerized Sistema de Infomação de Saúde stores:

- The unique identification code and name for the district;
- Population and geographic area;

The system also stores several columns that I have been unable to decode. The Computerized Sistema de Infomação de Saúde comes in two versions: national and provincial versions.

4.5.1.1 Data entry

In the Computerized Sistema de Infomação de Saúde, health workers or users can enter data in two ways:

- Computerized Sistema de Infomação de Saúde provincial data are entered manually from paper Sistema de Infomação de Saúde forms;
- Computerized Sistema de Infomação de Saúde national data are imported from a diskette; in the national version the system has no support for manual data entry.

4.5.1.2 Data storage

The Computerized Sistema de Infomação de Saúde provincial version only stores information about the province for all years. Data are stored in the following format: a folder for each year and a database for that year. The Computerized Sistema de Infomação de Saúde national version stores the information for all provinces for all years. A folder for each province, a folder for each year and a database for that year for that province are kept.

The database for a province at national level is a subset of the provincial version of that database, that is, the Gaza database at national level is a subset of the database of province level.

The Computerized Sistema de Infomação de Saúde program consists primarily of three parts: data entry, analysis and control.

4.5.1.3 Data entry in the Computerized Sistema de Infomação de Saúde

The person responsible for the Computerized Sistema de Infomação de Saúde at province level carries out data entry at province level. The person enters data from paper Sistema de Infomação de Saúde forms received from the districts. The procedures are the following: 1) selection of the form, 2) selection of the year, month and the health facility or district and 3) choose which data to enter. The following section describes each of the above procedures.

1 – *Select which form to work on*:

Figure 4.7 appears on the screen of the Computerized Sistema de Infomação de Saúde application and users can make the desired selection.

```
22/04/02        Ministério da Saúde - Sistema de Informaçäo        GAZA 2002

            +-------------[  Introduçäo de Dados  ]--------------+
            |                                                    |
            | A04  : PAV Resumo Mensal para Distritos            |
            |                                                    |
            | B05  : Stocks/Métodos Anticonceptivos              |
            | B06  : SMI Cons 0-4 Anos/Vig.Nutric.               |
            | B07  : SMI Maternidades e Parts. Trad              |
            | B08  : SMI Cons Pre Nat/Post Parto/PF              |
            |                                                    |
            | C04  : Consultas Externas e Stocks de kit A – CS   |
            | C04H : Consultas Externas dos Hospitais            |
            | C05  : Consultas Externas e Stocks de kit B – PS   |
            |                                                    |
            | J01  : Recursos - Rede Sanitaria/Camas             |
            | K01  : Recursos - Pessoal Niveis 1/2 (K02)         |
            |                                                    |
            |      Retorno ao Menu Principal                     |
            |                                                    |
            +-------------------[  Ajuda - F1   ]----------------+
```

Figure 4.7: "*Select form to enter*" screen capture.

It is important to remember that the program is DOS-based and the support for graphics is very low. It seems, that this is a problem for the health workers because they are accustomed to using Windows applications and application with support to mouse device.

2 – Choose year, month and place (district or health unit);

3 – Choose which data to enter in the system, move around entry boxes (similar to the paper form).

At national level to enter data in the system users only need to select the disk that contains data to import. The program recognizes which province and year the data is for.

4.5.1.4 Analysis

The embedded graphical analyzing utility is very poor compared to tools that the market offers. The users of the Computerized Sistema de Infomação de Saúde database prefer to export data to other graphical tools such as Microsoft Excel or Lotus 123. The provincial and national versions differ in a way that:

- The provincial version has the ability to make reports and analyze data per month;
- The national version has the ability to make reports and analyze per quarter.

4.5.1.5 Control

To control the data entry, the Computerized Sistema de Infomação de Saúde program is able to automatically find missing reports and gives its user a list of missing reports. The province can use this list to request the forms from the districts that have not delivered the forms. A tool called *veri.exe* creates text files containing information about duplicate records.

The Computerized Sistema de Infomação de Saúde national version has the function of checking the data that generates a report on empty databases and checks for faulty tables. It also checks for missing reports from districts, but this is not perfect due to the fact that it only reports up to which month data is reported, it does not tell the user if it is one or ten units that have reported.

According to Heywood and Rohde (2002, p.18), health information systems should support the following processes:

- Collect information (Essential data sets, tools, quality);
- Process information (Computerize, collate, report, target population, ensure validity, consistency and accuracy);
- Analyze (Compute indicators, targets, goals, coverage);

- Present (Tables, graphs, presenting the information, feedback);
- Interpret (Comparison, trends, meetings, supervision, epidemiological thinking);
- Use (Planning, decision-making, action).

Figure 4.7 shows which processes of the Computerized Sistema de Informação de Saúde are supported, and which are not supported with evaluative comments.

Table 4.7: Evaluation of Computerized Sistema de Informação de Saúde supported and not-supported processes.

Process	Explanation
Collect information	Partly supported by Computerized Sistema de Infomação de Saúde. Health workers are able to use the system to collect some information, but they are only allowed to collect information that the Ministry of Health wants, they have no ability to customize the system for local use and new data elements, i.e. they cannot add local information they want to track or add new vaccinations.
Process information	Not supported. Computerized Sistema de Infomação de Saúde is unable to assist the user with ensuring the validity and quality of data. The system is also making the data worse due to its inability to search for missing reports, forces 0 (zero) when a value is unknown.
Analysis of information	Not supported. Computing coverage factors is difficult when the infrastructure data is not updated. Low quality data leads to low quality analysis.
Tools	Not supported. I found that no one is using the Computerized SIS graphics tool; health workers told the research that this occurs because of the bad functionality of the tool.
Information interpretation	Not supported by the Computerized Sistema de Infomação de Saúde. Districts are unable to compare their data with other districts and facilities.
Use of information	Partly supported. At the national level some reports are produced from the Computerized Sistema de Infomação de Saúde.

Source: Ministry of Health, Provincial Directorate of Health, 2002.

Validity is very important, as everything is dependent upon the data quality.

4.5.2 Computerized Weekly Epidemiological Bulletin (CBES)

The Computerized Weekly Epidemiological Bulletin (CBES) is a computerized health information system part of the Sistema de Infomação de Saúde and is being used at national and provincial levels, while in the districts and health units paper-based forms are used. The purpose of the CBES is to:

- Report and analyze epidemiological information about critical diseases;
- Report weekly to allow quick reactions from the Managers to changes occurred;

Using the Computerized Weekly Epidemiological Bulletin, Managers can have control of the different diseases once they receive information weekly.

4.5.2.1 Data reporting and infrastructure

Health units report on a weekly basis using Sistema de Informação de Saúde form C03 (Weekly Epidemiological Bulletin). Form C03 contains data entry boxes for several diseases. Some are broken into age groups. This form is then sent all the way up to the province, where they are furthermore analyzed and then sent to the national level, the Ministry of Health.

The Computerized Weekly Epidemiological Bulletin (CBES) program is a MS-DOS based application and was built using FoxPro. At the national level the latest version is quite new (since 2000, version 1.6A), indicating that development of the application is still taking place. Also at the national level the program was configured to run on new computers, even if the program is MS-DOS based. A shortcut was added and configured to start the software in DOS. At the Provincial Directorate of Health the program is running on an old computer.

From the research at Provincial Directorate of Health, the researcher found that data in Computerized Weekly Epidemiological Bulletin has been stored in several text files and in a large Microsoft Excel file. Data are aggregated to district per week. The text files are laid out as a spreadsheet (form C03), the first column presents the name of the district that is reporting. The name of the file, according to health workers, determines the source of the file. For example, BEM96_9, means, year is 1996 and week is 9. This way of data storage is complicated and produces a lot of files. Taking into account the process of data gathering during the data warehouse construction this will add more work.

4.5.3 District Health Information System (DHIS)

The Health Information System Program (HISP) in collaboration with the Ministry of Health in Mozambique is implementing new computer database software at district level to support collection, compilation and analysis of health data. The computer database called District Health Information System (DHIS) is being used as a tool for health data collection, process and analysis. District Health Information System is closely related to Computerized Sistema de Informação de Saúde.

The District Health Information System has two main data categories: routine monthly data and semi-permanent data that are subject to change annually or within several months such as population data, health facilities infrastructure and staff list. In order to calculate indicators, routine monthly data are used as numerators and the semi-permanent data are the denominators.

The District Health Information System uses Microsoft Excel spreadsheet pivot tables that allow users to generate reports and graphs. At the time of thesis writing, District Health Information System was being introduced and used in Gaza districts with computers.

4.6 Summary

Figure 4.8 presents an overview of the existing health information system at the Ministry of Health. Several types of flows that occur among the systems are presented.

Key for Figure 4.8:

- Ideal flow ( ) ;
- Flow with problems of indicators of coordination (
- Sporadic flows ( );
- Sistema de Informação de Saúde and its different components;
- Different departments and systems inside Ministry of Health;
- Governmental or external institutions.

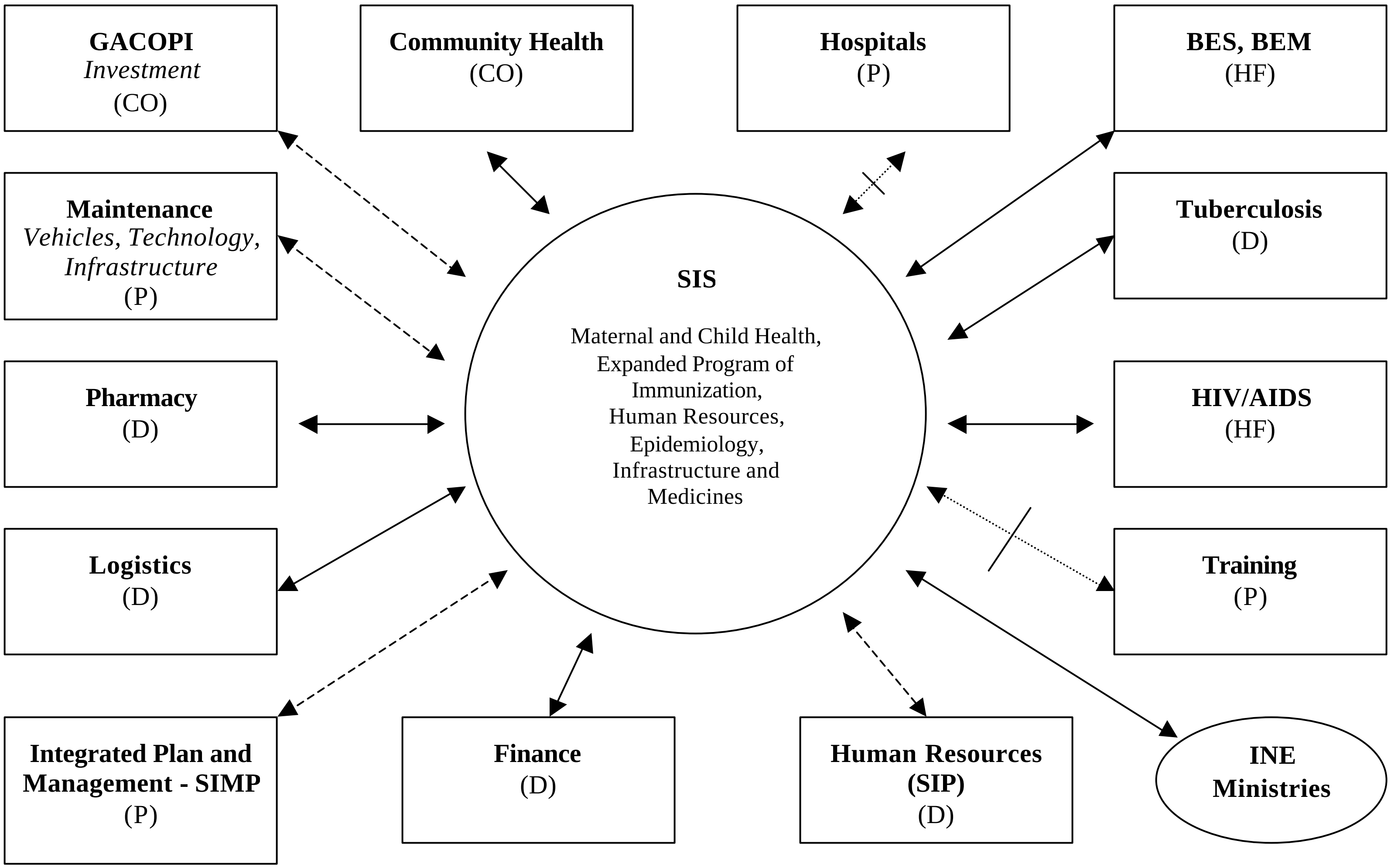


Figure 4.8: Data flow diagram of existing information systems and flow types between them and the Sistema de Informação de Saúde.

Key: (CO) – Central Offices, (P) – Provinces, (D) – Districts and (HF) – Health Facilities.

Figure 4.8 present the different existing systems in the Ministry of Health, as we can see the flows between the main system, the Sistema de Informação de Saúde (SIS), and other systems do not follow a unique standard, some are ideal flows others sporadic or have problems, so the fragmentation of the systems will make decision-making process, which should take all data into account, difficult. In summary, Figure 4.9 represents the Sistema de Informação de Saúde information flow.

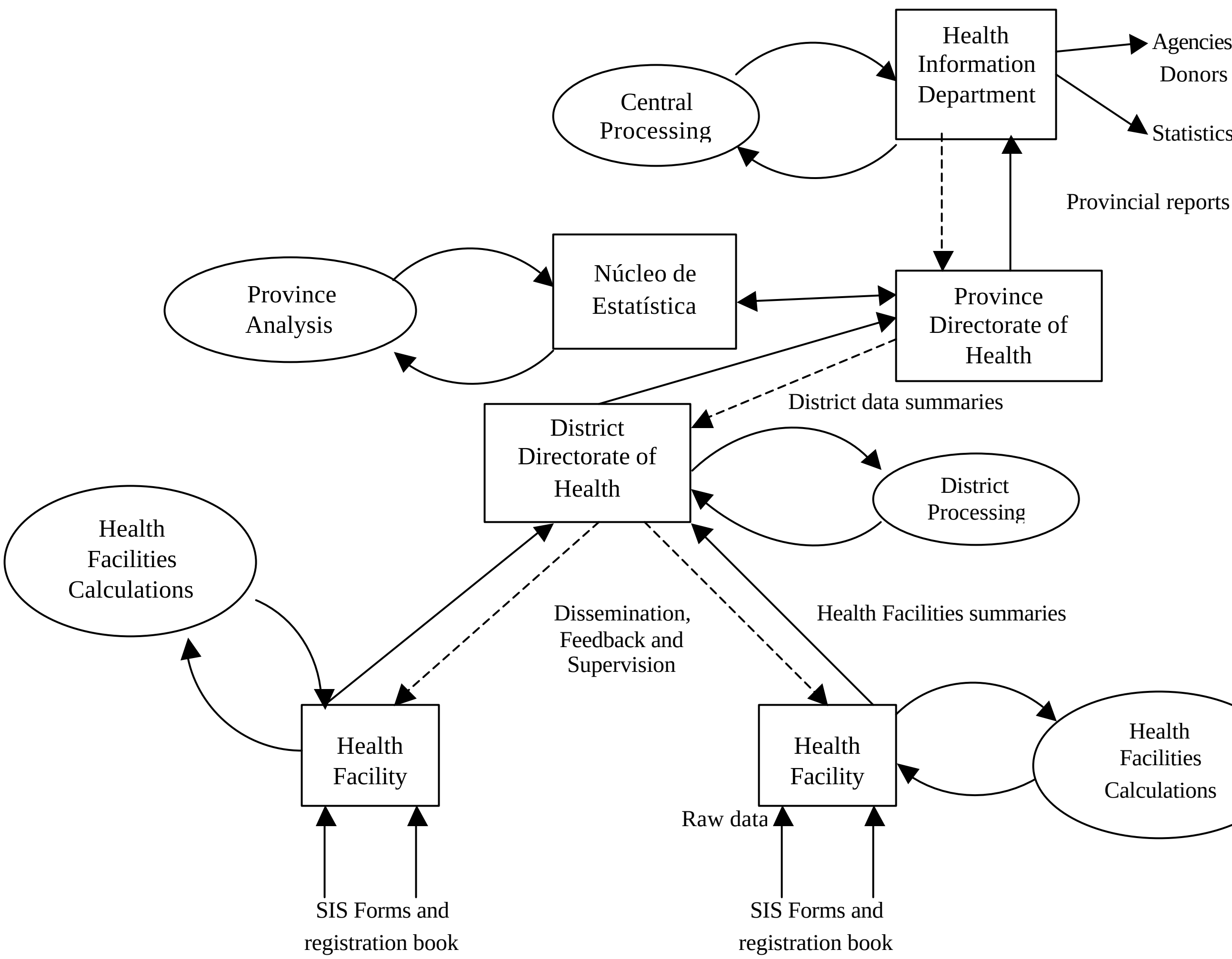


Figure 4.9: Summary of the Sistema de Informação de Saúde information flow.

(Adapted from Macome, 1992).

The Sistema de Informação de Saúde, has not changed much since 1992. The big change occurred in the number of Sistema de Informação de Saúde forms that were collected before. Flows continue the same. One problem since 1992 that the Ministry of Health did not solve is related to the dissemination and the feedback from the different levels.

Chapter 5

Information Flow at District, Province and National Levels

This chapter presents the main findings of the empirical study. The study was conducted in the Ministry of Health in Maputo, and Manjacaze and Chibuto, in Gaza Province. The findings are based on the qualitative methods and research techniques described in chapter 2.

5.1 Introduction

In chapter 2, I described the field sites where my research was conducted, as well as the qualitative methods and research techniques. In the field sites I was able to contact health workers working in different areas. Access to all types of health workers, e.g. managers was difficult, due to the fact they had other tasks to perform. The research techniques and qualitative methods were applied in different circumstances being the main purpose the collection data.

The empirical findings presented here follow the information cycle model proposed by Heywood and Rohde (2002) for the Health Information System Project (HISP). The model guided me during the field research and analysis.

5.2 The Information Cycle

A health information system is a reporting system that has the health facility in the community as the primary source of data. A health facility uses its inputs for its own local management. About health information system information cycle, Heywood and Rohde (2002), wrote:

> "The information cycle is a diagrammatic way of looking at information that enables us to see the links between the different phases of collecting, processing, analyzing, interpreting and using information. Each of these phases has a number of sub-phases that give more practical details of how to implement the cycle".
>
> (Heywood and Rohde, 2002, p.18)

All steps in the information cycle have the same importance, but it is important to notice that the use stage gives meaning to the whole cycle. The information that flows in the cycle is oriented towards action.

Figure 5.1 presents the information cycle, a general model suggested by Heywood and Rohde (2002, p.18).

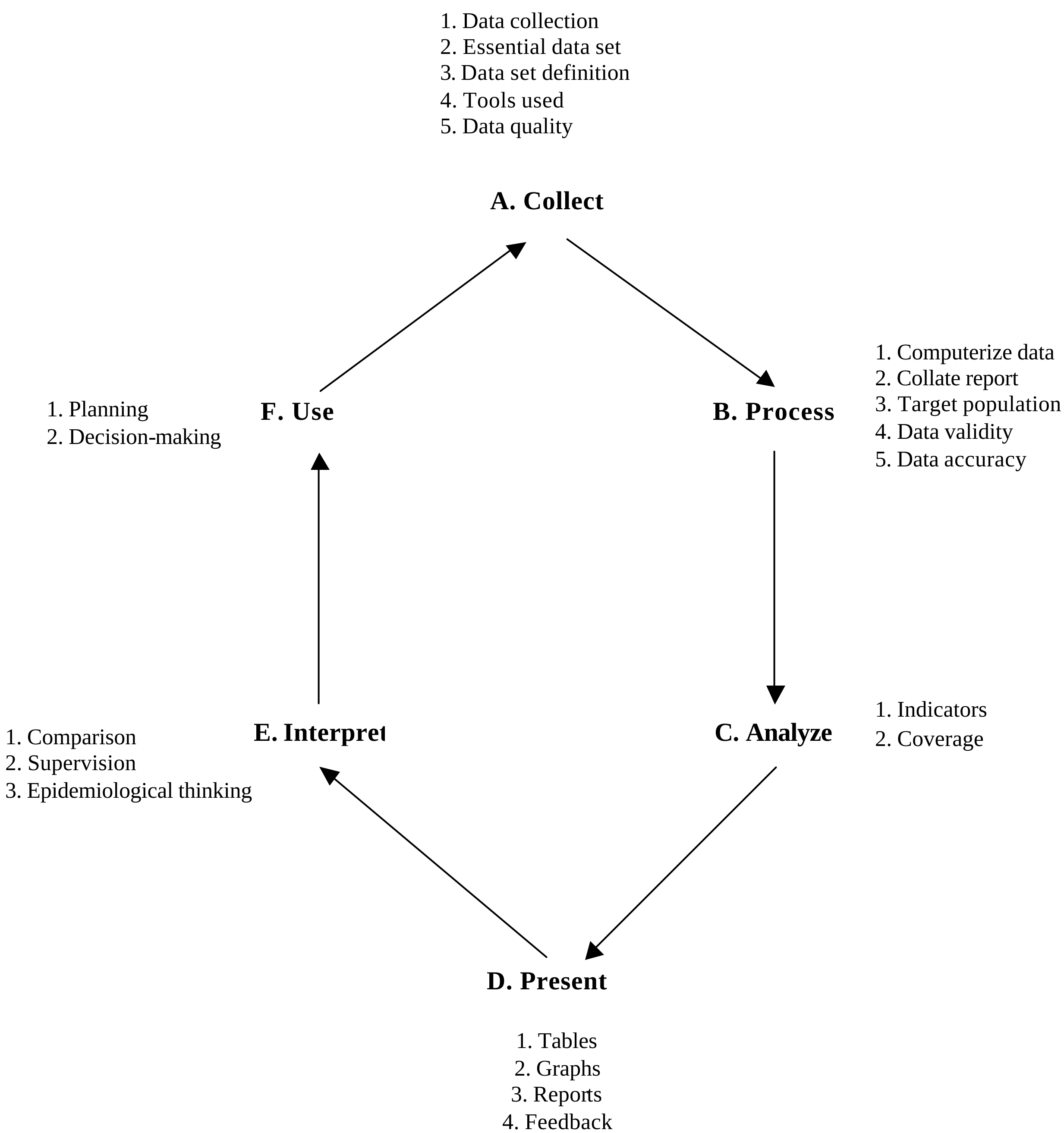


Figure 5.1: Information Cycle Proposed by Heywood and Rohde.

Adapted from Heywood and Rohde (2002, p.18).

5.3 Findings

The findings presented here are related to the following topics:

- Data collection;
- Data and Information processing;
- Information analysis;
- Information presentation;
- Information interpretation; and
- Information use.

5.3.1 Data Collection Phase

Heywood and Rohde (2002) state that, all data collected must have a purpose and the purpose of data at the facility level is to be used to calculate indicators that measure how well the various programs are performing.

5.3.1.1 Data Collection: Sources of Data

The sources of information are the data registered in a daily routine at the health facility. All health facilities in the districts collect data using the Sistema de Infomação de Saúde (SIS) forms for Expanded Program of Vaccination (EPI), Mother and Child Health (MCH), Weekly Epidemiological Bulletin (BES), hospitals and their infrastructure, human resources and population data. The tally sheets and registers represent the largest manual database of the Sistema de Informação de Saúde in Mozambique. There is a lot of registered data and not all of it can be processed. With data collection, the data are selected and the Ministry of Health has to decide and select which data to process.

In a recent study, Braa et al. (2000) discovered that, data collection is based on the needs of the health programmes and not on the needs of local health facilities. One nurse at the Manjacaze Rural Hospital, working at the Núcleo de Estatística (NE) in Manjacaze said:

> *"We have 17 health facilities. Until the first of each month we should receive the following forms of Sistema de Informação de Saúde: B03, B04, C04, C05, D03, process and then send them until 4th of each month to the Provincial Health Directorate. Not all data that comes in the forms are processed … ".*
>
> (Nurse, Núcleo de Estatística, Manjacaze Rural Hospital)

Errors

During the data collection process, different types of errors occur. Most of these errors occur at the health facility level, but are detected only at district level. From the research, I found that the typical and most important errors made by health workers in data collection are:

1) *Gap between the real and the reported data.* Real data means the data that are in the books and forms at health facility or the right data confirmed by supervisors. Reported data are the data that appear in periodic reports produced by health workers at different levels. – The data presented by health facilities in the reports are not the real data. This problem arises when supervisors go to health facilities and districts to supervise tasks. The supervisors check all Sistema de Informação de Saúde forms and then they compare the results with different reports submitted by the health facilities and districts.

Tables 5.1, 5.2 and 5.3 present the problem. The tables presented here are from the report of supervision conducted by the Provincial Directorate of Health in 1999. The first example comes from Gaza province in the district of Mabalane:

Table 5.1: Gap between real and reported data in the district of Mabalane.

Data Element	Real	Reported
Outpatients	2.579	2.629
Total admission days	378	342

Source: Supervision Report /1999/Provincial Directorate of Health.

The second case comes from the district of Chicualacuala.

Table 5.2: Gap between real and reported data in the district of Chicualacuala.

Data Element	Real	Reported
Outpatients	1.637	1.558
Discharges	79	74
Total admission days	542	462

Source: Supervision Report /1999/ Provincial Directorate of Health.

Table 5.3 shows an example from district of Massagena.

Table 5.3: Gap between real and reported data in the district of Massagena.

Data Element	Real	Reported
Total admission days	180	175
Bacillus Calmette-Guerin vaccine (BCG)	37	76
Diphtheria, Tetanus and Pertussis vaccine (DPT1)	61	136
Diphtheria, Tetanus and Pertussis vaccine (DPT2)	42	130
Anti-Tetanus Vaccine (VAT)	29	62

Source: Supervision Report /1999/ Provincial Directorate of Health.

Analyzing the tables above, I can conclude that the district of Massagena is the one presenting poor data. As we can see, the discrepancies for BCG, DPT1, DPT2 and VAT are big, for example, for DPT2, the district reported to the province 130 cases, but the real value or the values that the supervisors found during the supervision was 42. Another example, for DPT1 the district reported to the province 136 cases, but the real value or the values that the supervisors found during the supervision was 61.

By understanding the words of one supervisor working at the Provincial Directorate Health, we can understand the results presented in above tables.

> "*I usually participate in the supervision here in Gaza province and I think that people at health facilities sometimes give a likely looking number in their reports to give us, at province level, the image that they are working hard … and the supervisions prove the opposite*".
>
> (Supervisor, Provincial Directorate of Health, Gaza)

Other stakeholders at Province and National levels argue that this is an intentional error used by health workers to give the impression of huge workload.

2) *Calculation errors*. Calculation errors usually occur in health facility that does lackse calculators or other tools. This error is common in Sistema de Informação de Saúde forms that have calculated fields. One example is the Sistema de Informação de Saúde form A04, which has approximately ten calculated fields. In this form, health workers are supposed to sum two fields: *'0-11 months'* and *'12 months or more'* in order to obtain the calculated field "*total applied*". The problem is that health workers calculate the sum only one time. The first field has 833 and the second 11; therefore the sum should be 844 and not 864 as presented in the form.

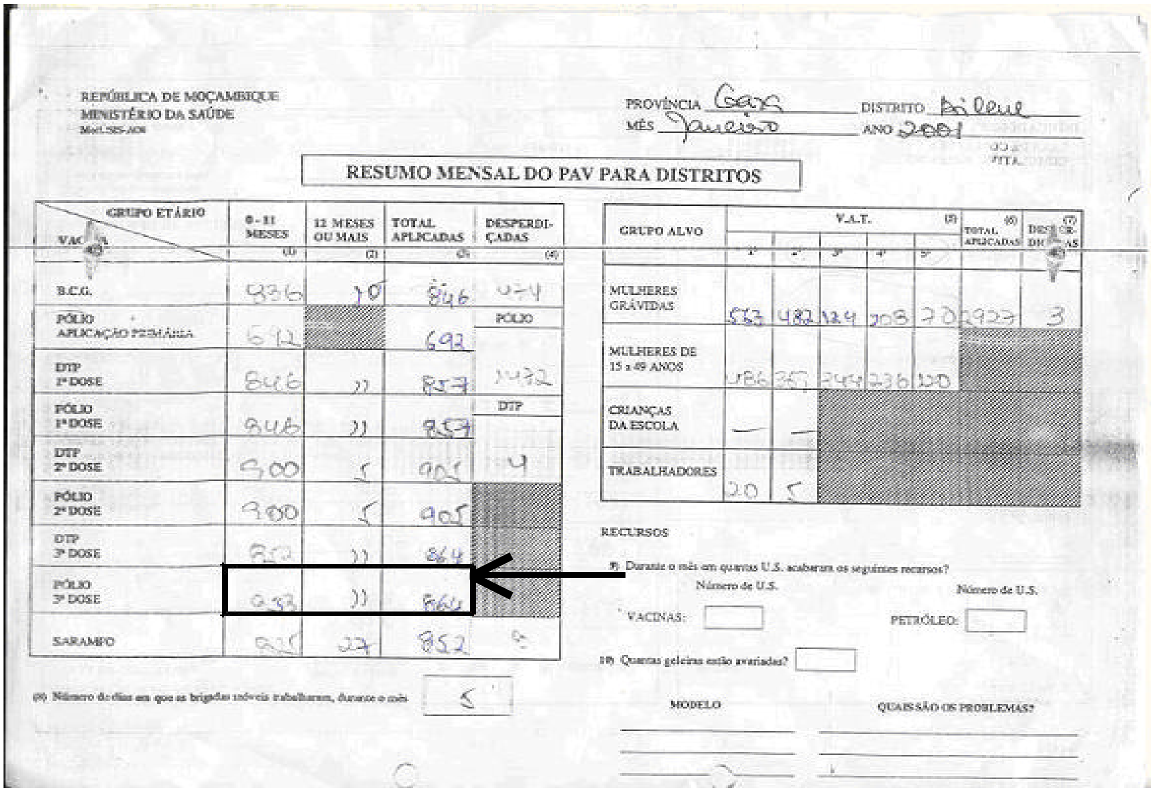


Figure 5.2: Sistema de Informação de Saúde A04 form with errors in the calculated field.

In Chibuto, a nurse and head of the Núcleo de Estatísticas pointed out:

> *"… health workers always forget the Ministry of Health instructions. In the meetings, our managers always ask us to sum and check the values two or three times for every field or establish a control using two colleagues from the Nucleus of Health Information … on the opposite side of some forms there are also some explanations on how to use or fill the form correctly …".*
>
> (Nurse, Head of Núcleo de Estatística, Chibuto)

From an analysis of different forms of Sistema de Informação de Saúde, I found that:

- Forms that have calculated fields such as A04 (EPI Monthly summary for districts) or A05 (EPI District resources annual summary) come with an explanation overleaf (*See Annex E*). The explanations are only for indicators in the A04 form and Expanded Program of Vaccination annual resources for the district in the A05 form;

- It seems that when health facilities receive the forms they also receive attached a text with explanation on how to fill in the forms (See Annex F) but health workers do not follow the instructions in the text.

3) *Duplicate data* – One example of duplication comes from the district of Chibuto, March 2002. *'Preventive consultations 0-4 years'* are collected in three different forms: B02 (Daily record for ante-natal consultations/postpartum consultations and family planning), B03 (record of consultations with children 0-4 years/Nutritional vigilance 0-35 months) and B04 (Health facilities monthly summary for Mother and Child Health/nutritional vigilance). There were different values for the same month on each form in Chibuto for March 2002.

4) *Missing data* – Several health units did not report. I started the analysis of Computerized Health Information System (CHIS) database, year 2000 and I discovered that for Mother and Child Health (MCH) several health facilities did not report.

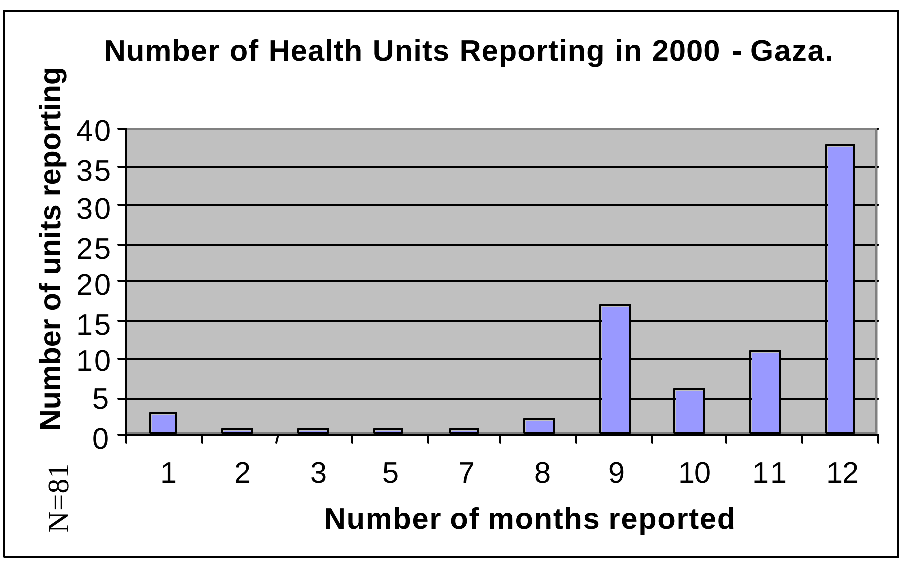


Figure 5.3: Number of health facilities reporting in Gaza in 2000.

Source: District Health Information System/2002.

The above graph shows that for B06 form (District monthly summary – Mother and Child Health (MCH), consultations with children 0-4 years/*Nutritional Vigilance*), 38 health facilities reported for all 12 months, 11 health facilities reported for 11 months. Some health facilities were very bad at reporting. Three health facilities only reported for one month through the whole year 2000.

Analyzing the same data, but this time checking the number of health facilities reporting each month:

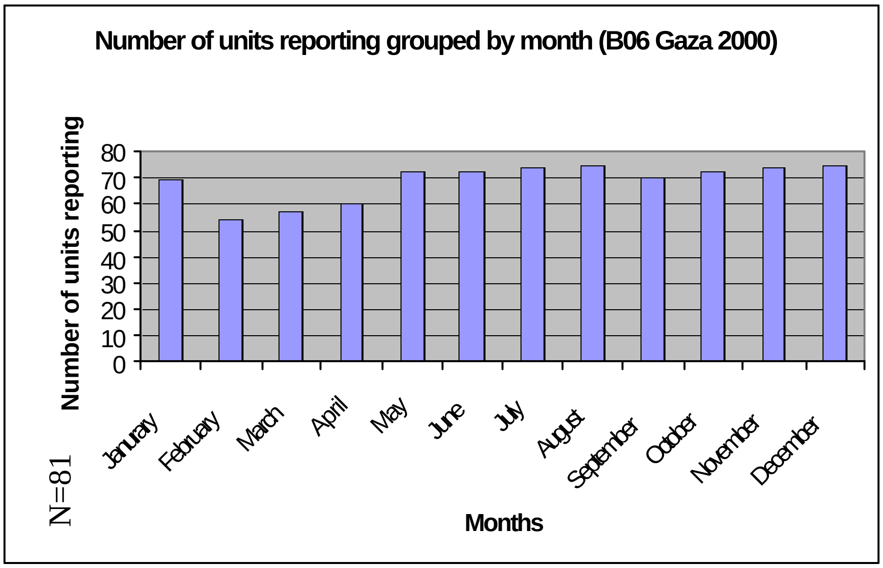


Figure 5.4: Number of health facilities reporting grouped by month in 2000.

The above graph shows that in February, March and April several health facilities failed to report, but we can also see that there is not a single month whereby all health facilities report (should be 81 for B06 form for a complete month).

Computerized Sistema de Informação de Saúde has a table called *'unids'*. Using this table that contains information of services that the health facility should provide, I found that: 81 health facilities report for one or several months in table B06 (district monthly summary – MCH, Consultations 0-4 years/Vigilance nutritional), 69 in table B07 (district monthly summary for maternities and traditional midwives), and 83 in table B08 (district monthly summary Maternal and Child Health, antenatal/postpartum consultations and family planning).

If all health facilities reporting on the tables had been reporting regularly we should have for each table the following number of records: *Number of different health facilities reporting * Reporting once a month = Calculated reports in the table below*. This table shows that health facilities are not reporting every month.

Table 5.4: Coverage of health facilities that reported in 2000.

Table	Number of reports that health facilities should report	Reported	Coverage %
B06	81*12 = 972	826	85 %
B07	69*12 = 828	694	84 %
B08	83*12 = 144	819	82 %

Source: Ministry of Health/Health Information Department.

The following table shows a poor example of missing data in the district of Chibuto, where several health facilities did not report for three months. Chaimite and Chipadja are the health facilities. Chaimite and Chipadja are special cases because these health facilities were new in 2002, so the reporting process is new for them. Having health facilities that are not reporting complicates management in different ways. Managers will not have a global view of all services provided in a certain period of time.

Table 5.5: Missing BES data and reports in Chibuto district.

Health Facility	BES	BES Received	BES Missing	Coverage
Chibuto-Sede	13	13	0	100 %
Malehice	13	11	2	85 %
Muxaxane	13	9	4	69 %
Maievene	13	13	0	100 %
Alto Changane	13	11	2	85 %
Chaimite	13	0	13	-
Chipadja	13	0	13	-

Source: District Directorate of Health, March 2002, Chibuto.

One problem of data collection task is that it consumes a lot of resources. Braa *et al.* (2000) conclude that, health workers spend most of their time collecting and reporting data taking them away from their primary task of providing health care.

Lack of staff is one of the issues that also influences data collection because one person performs many activities and it is important to have all these activities done rather than filing endless forms. So, what is filled in the form is sometimes what the worker thinks was done during the day. This is also influenced by the lack of information culture on filling correct data and making analysis of entered data. Forms are filled mechanically, without knowing the meaning of the data or at least what the end of it will be.

I think that once all data entered in the Sistema de Informação de Saúde at national level are generated at health facilities, districts and province levels, the quality of data at the health facilities, districts and province entry points becomes crucial.

5.3.1.2 Essential data set

To ensure that the same basic information is available from all facilities, the Ministry of Health has determined the essential data set that should be collected by everyone delivering primary health care services. At health facility level we can find the following type of data collected routinely by health workers:

- Monthly routine data:
    - Patients seen;
    - Different programs (Expanded Program of Vaccination, Mother and Child Health, HIV/AIDS and Tuberculosis);
    - Routine services;
    - Epidemiological surveillance (Notifiable diseases).
- Semi-permanent data about:
    - Population (Census, Births and Death registration);

- Administrative systems (Human resources, equipment, transport, equipment, budgets).

Activity data comprises data that changes every day, while the semi-permanent data changes annually or within several months. Examples of activity and semi-permanent data extracted from the District Health Information System are presented in Annex G.1 and Annex G.2.

5.3.1.3 Essential Data Set Definition

At the end of 2002, there were approximately 271 data elements that Sistema de Informação de Saúde collected through the Sistema de Informação de Saúde forms. The data sets were not well defined at that time because each health facility, district or province collected what they thought was most important for them, so the Ministry of Health initiated discussions about the correct data set with the provincial and district directorates of health. These will help the Ministry of Health in the definition of new data set that fits well in each health facility.

The new software District Health Information System addresses this problem of the common data set for all health facilities, districts and provinces. The designers of the software are considering an integrated summary that includes the required and most important data set for each health facility.

In Gaza, a team of Health Information System Project constituted by a researcher of University of Oslo (Norway) and four informatics students (1 PhD and 3 Masters) at Eduardo Mondlane University (Mozambique) and University of Oslo had a meeting with the Manager of the Provincial Directorate of Health, who said that his Directorate was in the process of establishing teams to work on the definition of the new data set for Gaza.

Figure 5.5: Meeting with the Head Manager of the Provincial Directorate of Health, discussing the definition of essential data set for Gaza.

5.3.1.4 Tools used

All levels of Sistema de Informação de Saúde are using forms (See Annex E and H) to collect routine health data at the end of a certain period of time, but for daily data collection they use patient records card, tally sheets and registers. Data collection tools are normally standardized for each province, district and health facility, according to an essential data set.

Figure 5.6: Nurse assisting a patient in Manjacaze Rural Hospital.

At health facility level, after sending the forms for the next level, all forms are stored in archive folders. Districts also keep paper based archive folders, but in districts like Chibuto, Manjacaze or Chokwé, where computers are available, they enter data from the forms into the District Health Information System. At province level data in forms are entered into District Health Information System database and the forms are stored in archive folders.

At national level, the Directorate of Planning and Cooperation stores the paper reports from each province and the files in the diskette, in the Computerized Sistema de Informação de Saúde database. The problem is that not all paper reports come with a diskette and the Computerized Sistema de Informação de Saúde does not allow data entry, but only the reading of data from the diskettes. In case the report comes without the diskette, the person in charge for the Computerized Sistema de Informação de Saúde should enter the data in the system.

From the research[1] conducted in April and May in the Ministry of Health, I noticed that the incompleteness of data is due to the fact that most provinces sent the paper report without any diskette and the person in charge for the Computerized Sistema de Informação de Saúde in the Ministry of Health does not enter the data in the system.

Figure 5.7: Sistema de Informação de Saúde forms archived at Provincial Directorate of Health office in Gaza.

5.3.1.5 Quality

Heywood and Rohde (2002), referring to the problem of data quality, wrote:

> "No data set is ever perfect and data need to be cleaned before it is useful to local managers. The goal of cleaning data is to ensure that data errors are small enough so they do not bias decision-making".
>
> (Heywood and Rohde, 2002, p.45)

[1] As master's students we conducted a research called *"Analysis of health information flows in the Ministry of Health"*.

Quality control is an obligatory exercise to be carried out before data processing. According to Heywood and Rohde (2002), good quality data should be:

- *Available on time and at all levels* – All the forms have their own periodicity, e.g. Weekly Epidemiological Bulletin has to arrive within the following week at the district in order to be useful, otherwise it will not be monitoring the routine data.
- *Correct, complete and consistent* – Checking the correctness of data is to check and observe carefully the reports of all health facilities. Checking completeness is to check if all health facilities are reporting what they are supposed to report.
- *Comprehensive and represent all data gatherers*.
- *Comparable* – Use data to compare data horizontally, e.g. compare data of health facilities in the same district.

From the research results presented in the above sections, I can conclude that collected data have poor quality. Data are not correct (see tables 5.1, 5.2 and 5.3 and figure 5.2), or not available, nor complete (see figure 5.3 and tables 5.4 and 5.5).

Various reasons contribute to this problem of data being collected without quality, including lack of use and control at the point of data entry, aggregation of data having errors and incompleteness, use of forms that are not well understood by health workers, too many forms to fill, inconsistency in the forms used, unavailability of data and the same data collected multiple times. Non-use of collected data to make decisions and poor feedback are also factors that influence health workers to not worry about the quality of data.

Figure 5.8: Graphs at Provincial Directorate of Health in Gaza.

5.3.2 Process Phase

The Ministry of Health has a variety of uncoordinated information subsystems that result in fragmented processing and partial analysis of data. A certain degree of institutional capacity for collecting and analyzing information for management has already been established.

Having collected data, it is necessary to convert it into information. "*Raw data needs to be processed to ensure quality, consistency and accuracy*" (Heywood and Rohde, 2002). Data need to be aggregated and tabulated in a way that aids future analysis so that it can be reported both vertically (sent up to higher levels) and horizontally (shared with all staff, districts or health facilities at the same level) and fed into a computer to be communicated.

5.3.2.1 Computerize data

Data are sent by health facilities to the District Núcleo de Estatística (NE) at the end of every month.

The District Núcleo de Estatística processes data manually or by using computers. Chibuto, Manjacaze, Chokwé, Bilene and Chicumbane districts have computers and have District Health Information System software installed, so these districts can process data using computers and then send the reports to the province.

Figure 5.9: Members of Núcleo de Estatística (statistics team) in Chicumbane working with District Health Information System.

The data computerization process starts when districts enter the data that are in the Sistema de Informação de Saúde forms into the District Health Information System.

Figure 5.10: Graphs on the walls of the Provincial Directorate of Health in Gaza.

5.3.2.2 Collate Report

Collation (also known as aggregation) is the gathering together of data from various sources. The objective is to put data into one comprehensive and representative report. About collation, one health worker said:

> "*Aggregation is a statistic operation that allows health workers to collate and summarize data from all health facilities from one district in a summary form*".
>
> (Nurse, Chibuto district)

Collation usually means the gathering of reports from all different health facilities. Different reports need to be collated into one report for the district. The same occurs at province level. Reports from all districts are collated in one report that presents an overview of the province. It represents a valuable summary of different activities. The advantage of the aggregation is that it allows calculating the district indicators, providing useful information for planning, and also draws the epidemiological profile of the most common diseases in the district. The report forms are standardized for each health facility in the Province Directorate of Health according to the services provided. In terms of disadvantages, data aggregation can turn aggregated data into non-useful data because it hides important details.

For example, when at district level, data from all health facilities are aggregated for all programmes except the Expanded Program of Vaccination; all detailed information about health facilities becomes lost. The managers of District and Province Directorates' of Health only realize this problem when the Ministry of Health sends, for example, a stock of medicines. They have problems in delivering the medicines to the districts and health facilities because the information is aggregated. If Chibuto needs 500 tablets of Aspirin® and Manjacaze needs 100 tablets, the Province Directorate of Health will aggregate the information. The request to the Ministry of Health will be 600 tablets. The Ministry of Health will provide 600 tablets of Aspirin® to the Province Directorate of Health without knowing exactly which district needs the tablets.

5.3.2.3 Target population

In order to calculate indicators, population data need to be processed in order to get information about the population receiving these services. The raw population data need to be processed in a way that health facility staff know the target population to be served by each program. The target population refers to those sections of the total population needing specific services. Each priority health program has a different target group for which it provides services. In order to quantify these groups, the population is divided in to age, sex and certain risk factors.

The target populations for each health facility for antenatal, deliveries and immunization are calculated at the provincial level and sent down to the health facilities. For each target population group, activities are planned and the number of expected interventions for the health facility can be calculated for the year and for each month. The province also calculates the operational targets for each facility according to past performance but these operational targets appear to be unrealistic and are often different from similar target populations. Table 5.6 presents the target populations suggested by the Ministry of Health.

Table 5.6: Target population and percentage of population suggested by Ministry of Health.

Target population	Percentage of total population
Children 0-11 months (Hepatitis B, Anti-Measles vaccines)	3.9 %
Children 0-11 months (BCG and Preventive consultations)	4 %
Children 0-4 years	17.1 %
Foreseen pregnant women	5 %
Foreseen delivers	4.5 %
Women between 15 and 45 who are fertile	24.9 %

Source: Health Information Department, Ministry of Health/January 2001.

The percentage presented in the table above is part of the whole population that represents a specific population. These percentages are uniform across the country. In this research I found that in Chibuto and Manjacaze districts they have problems calculating the target populations. Table 5.7 presents the case of Chibuto district:

Table 5.7: Target population problems in Chibuto.

Vaccine	Target population	Carried through	Coverage
BCG	1.424	4.805	337 %
DPT 1	1.388	4.835	348 %

Source: EPI – Chibuto / 2001.

Using table 5.6, health workers complete the second column. In this case the values should be the same, because for BCG and DPT 1 the percentage is 3.9 % of the population.

Table 5.7 presents the following problems:

- The *target population* for BCG and DTP 1 is different, whereas it should be the same, 3.9 per cent of total population;
- The *coverage* of more than 300 per cent is presented at data analysis workshops as if it is a good achievement, instead of being presented as a problem.

The problems of calculating populations for different programs and the over coverage are found in different districts of Gaza. Using different percentages health workers waste time producing their reports, because the reports do not correspond to the norms established by the Ministry of Health. For percentages, the data presented should be for 100 % and not for more than that, but health workers think that if they present values above 100 % the Ministry will think that they are doing well.

Related to these problems, it is important to notice that in a data analysis meeting in Chibuto, a nurse presented similar values and the managers and other health personnel did nothing. Continuing the research about target populations, I found that in Manjacaze district, three health facilities have the same total population of 5.335 habitants to serve. This induces managers to work with incorrect data.

5.3.2.4 Data Validity

At health facility and district levels there is no evidence of use of validation rules. In the validations rules, health workers would include for example, in the maternity, the checking of number of the discharges minus or equal to the number of births, number of following consultations minus the number of first consultations 0-4 years. Districts with computers and at province level are now making use of validation rules automatically because the District Health Information System has a function to control the validation rules.

In districts without computers, when they produce reports the only control that they make is to check that the values they enter in the summary are the same as the values on the forms received from health facilities. They do not check the validity of the data contained in the forms in terms of consistency, e.g. number of discharges greater than number of births in the maternity.

The District Health Information System software allows district health workers to formulate some validation rules; for example, the Computerized SIS at province level contains few validation rules that allow the Provincial Directorate of Health to define some validation rules. I found in the Computerized SIS a validation rule that checks whether the health facility that is currently reporting, has the ability to report the kind of data that it is reporting. In contrast, District Health Information System provides full information about the reporting process, who should report and what type of report should be provided.

From the research, I found that some health facilities report data that are not valid. On the form B07 (district monthly summary for maternities and traditional midwives), we have "*deliveries*" ("*Partos na Maternidade*" in the form) and "*discharged children*" ("*Partos na Maternidade*" in the form) fields. One validity rule would be *signed out children* less or equal to *deliveries*, that is, a health facility cannot sign out more children than they have delivered. When checking the data for 2000, we can see that Gaza has 145 reports (B07) of a total of 694 where the number of "*discharged children*" is greater than the number of "*deliveries*", 21 per cent of health facilities report impossible values. Figure 5.11 shows that there are few health facilities that systematically report poor data (report invalid data for 12 months):

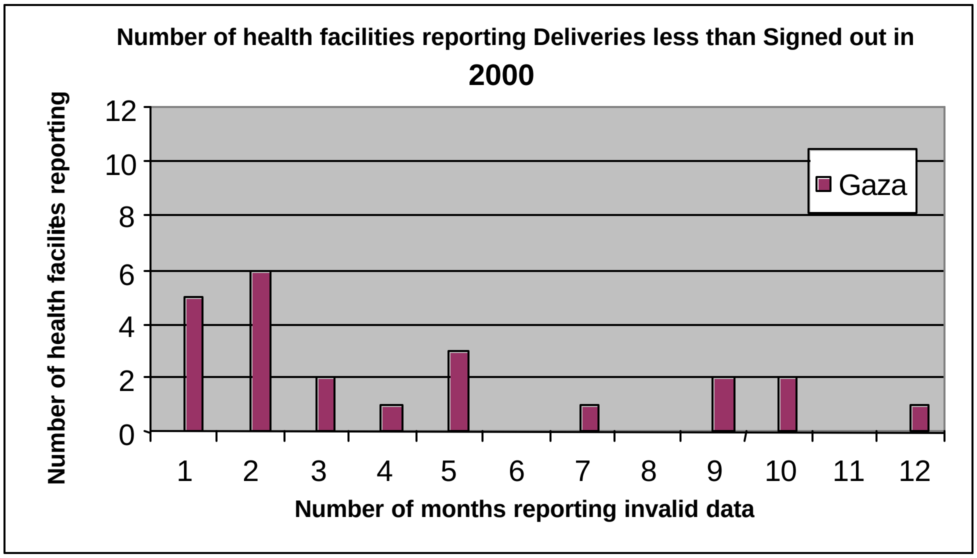


Figure 5.11: Number of health facilities reporting invalid data for Maternity grouped by the number of months they have reported bad data.
Source: District Health Information System/2002.

5.3.2.5 Data Accuracy

The goal of cleaning data is to ensure that errors are small enough so they do not disturb the decision-making in order to ensure data quality. "The worldwide experience is that the more that information is used by people who collect it, the more accurate it will became" (Heywood and Rohde, 2002). In various districts of Gaza province, data has no quality control. Health workers in Manjacaze and Chibuto agreed that, with the introduction of new software, the District Health Information System, the quality of data has improved since 2002. They argue that because the system controls the data entry, poor data are not entered in the system.

5.3.3 Analysis Phase

A large number of health facilities do not analyze the data and information produced, they only send the reports to the district without any analysis. The majority of the analyses are based on the calculation of mathematical operations that measure the frequency of qualitative data. Simple divisions between data in the numerator and data in the denominator result in rates, proportions and ratios.

5.3.3.1 Indicators

Indicators are tools that health workers use to convert raw data into useful information and enable comparison between different health facilities. "While the information system collects data, these data have to be made into information in the form of indicators, which relate the data to standardized populations or sub-groups or items" (Heywood and Rohde, 2002, p.56).

In Gaza province, the Provincial Directorate of Health has sent out a table with 15 indicators of various types (activities, epidemiological, resources, impact, efficiency, utilization, etc.) and norms for each of them at district level. At district level, there is a minimal analysis of indicators and at other levels where they are analyzed, they are generally not interpreted and used for local self-assessment.

I was able to attend two different meetings in Chibuto and Manjacaze conducted by each district health manager:

- In Chibuto, it was a quarterly meeting of all health facility head, district programme managers and some provincial officers. In this meeting health workers made only a formal presentation of data and information to a group of about 50 people, with minimal interaction. The presenters only read the values in the tables or in some cases they read sentences about data, instead of presenting summarized tables or graphs.
- In Manjacaze, I had the opportunity of participating in a monthly meeting of district program heads consisting of five local managers. There was more interaction, but the health workers did not present any graphs or tables. Also here, health workers only read the aggregated forms.

I noticed that in both meetings, there was a great deal of staff effort put into preparing comparison of raw data from 2002 and 2001. Another perception is that not all health workers are able to calculate routine indicators; in Chibuto and Manjacaze just the managers and program managers were able to calculate or work with indicators.

> *"… managers are responsibles for producing indicators, due to the fact that most health workers do know not how to calculate it. A manager receives the reports and produces the necessary indicators. Maybe now with new software that allows automatic calculations of indicators we will receive reports with calculated indicators".*
>
> (Program Manager, Gaza Province, 2002)

It seems that this happens because this essential activity concentrates only in a few hands and ordinary health workers do not have the opportunity to produce reports or show their skills in this activity. However, in the group of ordinary health workers we can find skilled workers to work with health data.

5.3.3.2 Coverage

Analysis of coverage means looking at who actually received services and comparing that to the population that should have benefited from services. From the research, I found that health facilities have been over-reporting. This means that the values that they have been presenting in the reports are not accurate values. The first example comes from Chibuto where health facilities report more than 100 per cent of coverage for immunization. In this case they reported about more than 300 per cent of coverage for two vaccines: BCG and DPT 1.

5.3.4 Presentation Phase

Data need to be analyzed and displayed to be functional for managers. The main instruments that health workers use to present health information are the tables, graphs and reports. In an ideal situation all health facilities should use these instruments to present their summarized data to the next level of management.

5.3.4.1 Tables

According to Heywood and Rohde (2002), tables show figures as cells in spreadsheets format and enable us to compare the facility over time or against other health facilities. A table is one basic instrument to preset information and to organize data to facilitate the calculations and analysis. District summaries are table particular form that serves to compile data from different health facilities. Every month when health facilities or districts send the reports, major part of it is composed of tables. The tables align the data in a way that health workers can easily calculate and make analysis.

5.3.4.2 Graphs

Graphs are the pictorial representation of the information. Linear and histogram are the most common graphs that I found in different health offices. Bar and pie graphs are also in the list of graphs types that health workers use to present the information.

Table 5.8: Types of Graphs used in Manjacaze Núcleo de Estatística Office.

Graph type	Content
Activities growth	Antenatal consultations, Family Plan, Activities, Consultation 0-4 years, consultation 0-11 months
Indicators comparison	Mother and Child Health
Deaths	Medicine, pediatrics, maternal
Expanded Program of Vaccination	Accomplishments 2000-2001, coverage breaks, tax of use
Weekly Epidemiological Bulletin	Weekly Epidemiological Bulletin sent

Both districts in Gaza have large numbers of graphs on the walls of district health directorate. Most of these graphs only compared two years, with detailed calculation of increases or decreases over the period without comparing to the norms. Even the few that made graphs of indicators did not compare them with provincial guidelines. Graph design involves considerable graph-making skills when health workers work without computers. Few health workers are able to interpret graphs and explain their meaning.

5.3.4.3 Reports

Reports are documents that contain analyzed data, comments, observations and indicators to communicate and disseminate information to managers of National Health Service and other users. Some reports are predefined and uniform and some are *ad hoc* depending on the circumstances, type of information that they are reporting and conditions at district or province levels.

In districts where the information is manual the tendency is to use standard reports. The introduction of computers in some districts and conversion of manual database to automated database, allow the elaboration of standard and periodic reports. An essential characteristic of the reports is to cross information and indicators about events, diseases, and services provided with the information about resources used.

The structure of the reports proposed by the Ministry of Health is composed by the following topics: Recital, objectives, description, facts, interpretation and conclusions and recommendations.

From the research I found that not all health facilities report in the same way. Different health facilities present different models of the same report. This way of work causes several problems for the supervisors working at province level. The quality of the reports presented is reasonable and acceptable, taking into account the working conditions of the health workers.

For districts with computers, the process is easy, but for districts without computers sometimes they present hand-written reports. Another problem in the reporting process is time issue. Some health facilities took long time to deliver their reports for the appropriate level. One problem is transport because the roads are in bad conditions.

5.3.4.4 Feedback

Data are collected daily, weekly and monthly and sent to upper levels of management, but the reverse way is too scarce. No one sends any answer about a well-done job; they just complain about what was done wrong or what went missing. This does not encourage health workers to collect data. There are mistakes in the collected data that nobody cares about. For example, in the issue of immunizations, why is it that in a certain month the coverage rates are higher than the target population? Why are they more than 100 per cent?

Feedback is virtually non-existent, unless there are glaring gaps or big mistakes in data sent to higher levels. “The feedback routines from province to district and from district to health facility are poor” (Braa et al., 2000). There is no evidence of written feedback from either province or district levels, but f the feedback works, comments on reports are sent back to district or health facilities.

5.3.5 Interpretation Phase

Interpretation of the information is one of the most important and difficult phases of the entire cycle of information. This phase is important for the whole process of the information cycle. In this phase based on the presented information, managers together with workers start to give some meaning to the results presented.

5.3.5.1 Comparison

Health workers at Provincial Directorate of Health in Gaza interpret by doing comparisons and they use two types of comparisons:

- Data of one health facility or service with another to detect similarities and differences;
- Data of the same area or service analyzed through time to detect oscillations and tendencies;

Health workers interpret to establish a priority scale between disease cases, programs and service results or personnel performance. Calculated indicators are also used to compare the objectives established at the beginning of each year. Health workers use comparison to determine the best performing health facility, district or province in a specific program or service, at the end of the year.

One problem found during the research is that health workers do not follow the national norms or guidelines when comparing the performance of the health facility with the national performance.

5.3.5.2 Supervision

Supervision is the most important form of feedback on information and health personnel working at provincial and national levels makes it. The main focus of the supervision visits is on information analysis, interpretation and use. While considerable data is collected and sent upwards, very little data comes back after analysis to support local action.

In some cases, data are primarily used to criticize or to reprimand local staff for not achieving targets, and not motivate and support their actions.

Table 5.9: Results of 1999 supervision in different areas in Gaza districts.

District	Finance	Pharmacy	Provision	Infirmary	EPI	MCH	Laboratory	Total
Chibuto	1	1	2	2	2	2	2	12
Mangueze	4	4	2	3	2	2	1	18
Bilene	3	1	2	2	1	1	1	11
Guijá	3	3	2	3	2	1	2	16
Chokwé	3	3	2	1	1	1	1	12
Total	14	12	10	11	8	7	7	

Source: provincial directorate of health, 1999 supervision report, Gaza.

Key: scale 0-5 (no problem – very problematic).

Table 5.9 shows that the supervision found that Mangueze and Guijá are the districts that in general have more problems, and Finance and Pharmacy are problematic areas.

One important conclusion of the strategic plan for the health sector is that, "from a technical point of view, supervision has been implemented inefficiently. Monitoring, retro-information and incorporating supervision into management systems as a way of resolving problems have all been poor" (MoH, 2001, p. 27).

5.3.5.3 Epidemiological thinking

Epidemiological thinking is the process of answering questions about the population covered by health facilities. "The key issue of epidemiological thinking is that it always relate data to known population" (Heywood and Rohde, 2002). From the research I can conclude that epidemiological thinking only occurs at province and national levels.

5.3.6 Use Phase

This is the last phase of the information cycle proposed by Heywood and Rohde (2002). All information generated in the previous phases is presented here and managers should use the information for planning, decision-making and to take actions.

5.3.6.1 Planning

The planning process starts at the national level at the Ministry of Health. The Ministry has a plan for different areas of services such as human resources, finance, administration, pharmacy, or laboratory. The plan is usually from three to five years and it is called "*Plano Director*" (i.e., Master Plan).

Planning is still inadequate due to the lack of integrated information for managers at the different levels of the National Health Information System (NHIS). Existing systems, such as Sistema de Informação de Saúde, lack an integrated vision (MoH, 2001), so the planning process is affected and managers cannot plan based only on the data contained in the Sistema de Informação de Saúde.

5.3.6.2 Decision-making

During the research, different Managers recognized that relevant decisions could only be taken based on data coming from different existing systems. Managers interviewed during the research argued that decision makers need to integrate different information sources in order to make correct decision.

Service managers at all levels should use the information collected at different levels for decision-making process. The problem is that, for each level of the National Health Service (health facility, district, province and national) there is a deadline for sending reports to the upper levels, according to managers at the Provincial Directorate of Health in Gaza and at the District Directorates of Health in Manjacaze and Chibuto.

What happens is that sometimes these deadlines are not followed, according to the interviewees. They argue that, if someone wants to take a decision, he/she will take it without any information coming from that level. It leads me to conclude that the data collected are several times not used for decision-making process.

5.4 Summary

The information cycle suggested by Heywood and Rohde (2002, p.18) proved to be a good tool during the field research. By using this tool it was possible to discover the main problems of the National Health Information System. Related to the data warehouse in the Ministry of Health, the definition and introduction of the essential data set will ease the task of data integration and poor data quality will make the integration and cleaning of data processes into a difficult and laborious task.

The main findings of my empirical research can be summarized as follows:

- Data are collected methodically by health facility level and systematically sent up to district and province levels.
- There is an effort by the province to introduce the use of indicators, but indicators are not effectively used in the analysis of data at any level;
- Data quality is poor due to:
  - Excessive amounts of data being collected;
  - Lack of clear definition of data elements;
  - Aggregation at district level in such way that health facility detail is lost;
  - Lack of feedback from higher levels.
- Data analysis meetings do not carry out useful analysis;
- Graphs are widely used and displayed but most of them are meaningless;
- Feedback is non-existent at any level;
- District Health Information Software is not used at district level by health workers;
- There is evidence that information is not used for planning, supervision, monitoring or evaluation of services and;
- Work conditions aspects affect the performance of health workers.

Chapter 6

The Data Warehouse Architecture Model

This chapter starts with a summary of the existing problems in the Ministry of Health and then presents the data warehouse architecture model and a description of its different components.

In general, we can summarize the problems of existing systems in the Ministry of Health as follow (Ministério da Saúde, 2003):

- All systems lack integration within the Ministry of Health. The same happens with the external systems, donors, and other institutions;
- Lack of human resources to support the systems;
- Poor data quality and poor information produced by the different systems;
- Data quality is related to problems related to inefficiency of means of communications;
- Insufficiency of equipment to support the use of the systems;
- Lack of equipment maintenance and users training; and
- Lack of procedures and information management process.

Data warehouse for the healthcare area is not a recent topic. The revolutions that information and communication technologies had originated provide new perspectives to the management in the healthcare sector. As seen in chapter 3, data warehouse is a technology in which data are next to the users. The objective of this chapter is to present the proposed data warehouse architecture model to the Ministry of Health. I hope that this model will help the Ministry in the definition of the right platform to the data integration of different legacy systems. The focus is on the different components of the model.

During the development of a data warehouse model to the Ministry of Health the construction of a decision support system in the healthcare sector was tested. The proposed model was developed through the following stages:

1) Documentation of existing legacy systems;
2) Follow the life cycle and steps for building data warehouse; and,
3) Proposal of data warehouse architecture model.

6.1 Initial Developments of the Model

Based on the theory presented in chapter 3, section 3.5, theory related to the data warehouse life cycle in data warehouse development, this section presents the points of data warehouse life cycle that had been covered during the study.

The phases of data warehouse life cycle presented by Flanagan and Sadie (1998) that I covered in the present study are the following:

1) Document the organization legacy systems;
2) Data warehouse model; and,
3) Clean data.

The following steps were covered:

1) Define data sources;
2) Extract, transform and clean data;
3) Create the data warehouse model and design the data warehouse architecture model; and
4) Choose the database technology or the platform for the data warehouse.

The following section presents the four steps of data warehouse life cycle that I covered. The section is a combination of all theories related to data warehouse presented in chapter

3. Theory presented in chapter 4, the Health Information System in Mozambique also supports the following section.

6.1.1 Define Data Sources

The definition of data sources for the proposed data warehouse architecture model was based on figure 4.2, section 4.2.1 and figure 4.3, section 4.3.1 (see chapter 4). These figures presents the health structures within the Ministry of Health in Mozambique and the main flow of information within district, province and national levels of Health Information System in Mozambique.

The selected data sources for the study are the following: GACOPI-Investment, Maintenance, Logistics and Finance from Directorate of Administration and Management, Health Information System and Sistema Integrado de Planificação (SIMP) systems from Directorate of Planning and Cooperation, Human Resources System from Directorate of Human Resources, Pharmacy system from National Health Directorate and External Sources.

The following section presents a description of the systems.

6.1.2 Document the Existing Legacy Systems

Table 6.1 presents the main legacy systems available in the Ministry of Health. The table presents a brief description in terms of platform used, type of database implemented, type of language used to build the application, description of the system, the main outputs of the system and the location inside the Ministry of Health. Registering the information flow among different systems was difficult. It is important to notice that I was not able to collect data for all systems.

The process was based on the document revision and interviews. The process identified the main and relevant systems in use in different departments of the Ministry of Health.

Table 6.1: Existing legacy systems in the Ministry of Health.

System	Platform	Database	Language	Description	Outputs	Location
GACOPI - Investment	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
Maintenance	N.A.	N.A.	N.A.	N.A.	N.A.	Dept. of Maintenance
Pharmacy	N.A.	N.A.	N.A.	N.A.	N.A.	Pharmacy Dept.
Logistics	N.A.	N.A.	N.A.	N.A.	N.A.	Logistics Dept.
SIMP	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
Finance	N.A.	N.A.	N.A.	N.A.	N.A.	Dept. Administration and Management
Human Resources	N.A.	N.A.	N.A.	Details of human resources distributions	N.A.	Plan and Management Dept.
Health Information System	MS-DOS	DBase	DBase	Detail of health data from different levels	N.A.	Plan and Cooperation Dept.
External Sources	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.

Source: Ministry of Health/ Directorate of Planning and Cooperation/March 2003.

Key: N.A – Not Available.

6.1.2.1 Analysis of Existing Systems

An important analysis related to use of the system should be made at this stage. From the research it was very difficult to prove if the managers are using the system for decision-making process. One example comes from the Computerized Sistema de Informação de Saúde (CSIS). This system contains incomplete data for several years. In this situation it is difficult to conclude that the system is not used.

6.1.3 Extract, Transform and Clean Data

The extraction, transformation and cleansing process was done by a group of Health Information System Program members and by a Norwegian student. We collected data from the Computerized Sistema de Informação de Saúde (CSIS) in the Ministry of Health and we started the process.

6.1.3.1 Extraction Process

For the extraction process we developed our own data extraction software, the software is a Microsoft Access database. For us it was easy to develop the software, instead of purchasing a commercial program. The software developed is easy to use, from the point of view of the developers of the software. Annex I, presents a description of *Database Filer 1.0* software. This software is able to read data directly from the data sources and store data extracted in the software relational database management system. In this staging area we started the transformation and cleansing process.

During the extraction process we faced different problems. The problems were related to data in the Computerized Sistema de Informação de Saúde (CSIS). Data fields with strange names, tables with no primary keys, strange data in the fields, were the main problems encountered.

6.1.3.2 Cleansing and Data Transformation Processes

This stage consisted basically in data cleansing and transformation. We had to go through different records in the extracted database to search for errors in data. The process was conducted using correct data provided by the Ministry of Health.

It is important to notice that, only four tables (forms) were selected from the Computerized Sistema de Informação de Saúde (CSIS), these being: EPI data per district, per month (A04), Mother and Child Health data, children consultations per health unit, per month (B06), Mother and Child Health data, deliveries (Maternity) per health unit, per month (B07) and Mother and Child Health data, pregnant consultancies per health unit, per month (B08). These forms were in a DBase format. We exported all forms to Microsoft Excel and then imported from the *Database Filer 1.0* software.

The main focus of the cleansing process was in the number, location and names of health facilities, the health infrastructure. What we did in the cleansing process was to correct errors and correct inconsistencies. We also dealt with existing duplicated data in the system. Finally, we put data in the correct standard format. In the transformation process we eliminated selected useless fields and we created primary keys.

Watson (2001) argues that poor data in the data sources systems is the result of poor data quality practices. From the cleansing process we found different reasons for dirty data in the Computerized Sistema de Informação de Saúde (CSIS). The main findings are the following:

- ***Absence of data or missing data***. Where data are expected but lacking. In this case, we found that certain fields do not benefit from data.
- ***Dummy values***. Inappropriate values in the fields
- ***Violation of business rules***. Some basic rules are violated, e.g. number of discharged children greater than deliveries.

- ***Calculation errors***. In fields that support sums, the values presented are not correct.
- ***Data analysis meetings do not perform useful analysis***. Health workers do not use meetings to perform better analysis and check for errors on the data.
- ***Feedback is non-existent at any level***. National, province and district level are not using the feedback as a mechanism to correct the errors in the reports.
- ***District Health Information Software (DHIS) is not used at district level***. Health workers at districts that have computers and the DHIS are not using this powerful tool to help them in the control of the quality of the data.

6.1.4 Data Warehouse Model and Data Warehouse Architecture

6.1.4.1 The Data Warehouse Model

The data warehouse data model is built from the Ministry of Health data model.

> "A data model is a graphical representation of the data for a specific area of interest. That area of interest may be as broad as all the integrated data requirements of an entire business organization, such as an 'Enterprise Data Model', or as focused as a single business area or application".
>
> (Sanchez, 1998)

The data warehouse data model consists of at least two major components – a high level model and a mid level model. The high-level data model consists of the major subject areas of the organization. The mid level data model data model consists of keys and attributes (Inmon, 2000).

In this study, I was able to built only the high level component because I did not have access to all the different systems in the Ministry of Health. The diagram presented in Figure 6.1 consists of a model with the main subject areas.

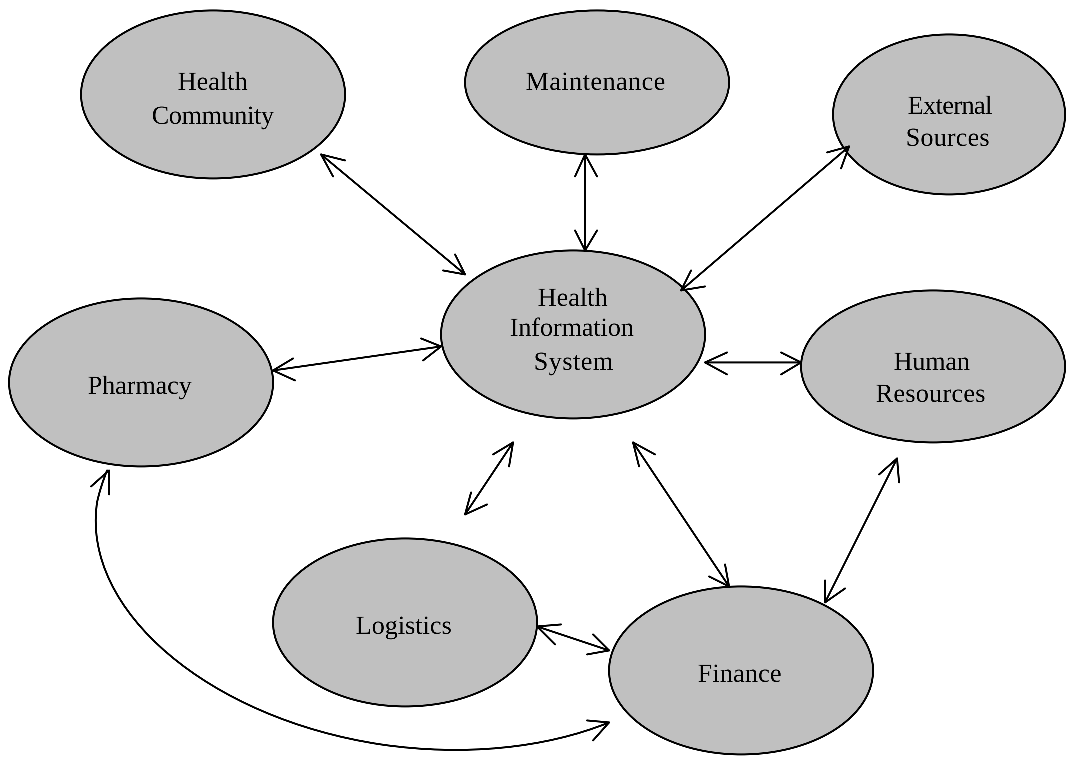


Figure 6.1: Ministry of Health data model showing different subject areas.

The arrows in the figure represent the data flow between systems and the ellipses represent the functional business areas. The model provides a viewpoint of integration for the Ministry of Health.

After selecting the entities for each subject area we need to design the dimensional model called star model (see chapter 3, section 3.7.1). In the star model we have the dominant, or fact table, and the dimensions that are secondary tables. I suggest the star model due to the fact that different authors have tested successfully the star model.

Domenico (2001) presents an example of the use of the star model. In the example, he tested the model in a case study conducted in different Brazilians Universities.

6.1.4.2 The Data Warehouse Architecture Model

Based on the theory presented in chapter 3, I propose the construction of a data warehouse architecture model using the bottom-up architecture. In the proposed model, first we construct the data marts from the legacy systems and then the data warehouse from the data marts. Figure 6.2 presents the proposed data warehouse architecture model.

THE PROPOSED DATA WAREHOUSE

MODEL HERE! PLEASE!!

Figure 6.2 - The Proposed Data Warehouse Architecture Model for the Ministry of Health

PRINTED SEPARATELY

File 85 1 PAGE 149

6.2 Description of the Proposed Data Warehouse Model

This section describes the different stages and components of the proposed data warehouse architecture model for the Ministry of Health. The model presents different stages of the implementation.

6.2.1 The different stages of the data warehouse model

a) Data Sources

Much of data comes from operational systems inside the Ministry of Health such as Human Resources, Logistics, GACOPI-Investment, Maintenance, Pharmacy, Health Information System and Finance. External data such as Mozambique census, data provided by INE[1], Ministry of Health partners, government organizations may be included.

All systems are in different formats, platforms, and are written in different languages and produce different outputs (See Table 6.1, Section 6.1.2). The most important at this stage is the definition of the data sources for the data warehouse project. The Ministry of Health should select the data sources that are important for the decision-making process. After the extraction process using software for extraction data are stored in the staging area.

b) Staging Area

The second stage of the model is the staging area (which can be held in memory, on disk, or even on a separate machine). It is used for consolidating data as it arrives – the staging area should be used as an area to perform all of the data transformations. The staging area is a storage area and a set of processes that clean, transform, combine, archive, and

[1] Statistics National Institute provides different types of data in different areas of development.

prepare source data for use in the data marts. It is the interface between the source system and the data marts.

At data staging the Ministry of Health should design the Wrappers. Wrappers are small programs written in a specific language for each type of data source and it converts data from one data model to another. Wrappers can be written for many data sources types (e.g. Common Business Oriented Language (COBOL), Microsoft Access, DBase, flat files, etc.). Data staging is also called copy management or replication management, but in fact, it includes all of the processes necessary to select, edit, summarize, combine and load data marts and information access data from operational and/or external databases. Data staging often involves complex programming, but data mart tools are increasingly being created, and thus help in this process. Data staging may also involve data quality analysis programs and filters that identify patterns and data structures within existing operational data.

The main tasks in this stage will be: Find and remove duplicate tuples, detect inconsistent and wrong data, attribute values that do not match, patch missing and unreadable data, insert default values and notify sources of errors found.

The next task is to load the resulting data in the data mart. For each data source system we need to have one data mart. During this process the Ministry of Health needs to decide about the type of data load to perform. The Ministry can select, from incremental versus refresh, off-line versus on-line and the frequency of loading (at night, once a week/month or continuously).

c) Integration

Different authors (Murlrow et al., 1997; SILICO, 2001; Staccini et al., 1998) described the integration process as a mechanism to quickly integrate heterogeneous systems to create a single view of health information (clinical and financial). Murlrow et al. (1997) argue, that heterogeneity of data sources complicates the integration process. The heterogeneity of the information systems in use and the slow adoption of open standards are major obstacles to achieving functional integration of the healthcare processes.

At integration stage we have cleaned data in the data marts. A Data mart is a "small" data warehouse designed to support a specific activity or department. Data marts are often used to provide information to functional segments of the organization. Typical examples are data marts for the pharmacy department, the maintenance and logistics department, the finance department, and so on.

Different data marts for the Ministry of Health should be designed from the perspective that they are components of the data warehouse regardless of their individual functionality or construction. This will provide consistency and usability of information throughout the Ministry of Health. At the end of the integration stage there is an integrator that receives data from multiple data marts, then monitors data and integrates into data warehouse. To load information in the data warehouse, the integrator must filter, refine, integrate and combine information coming from different data marts sources. The integrator also solves inconsistencies in data, eliminates duplicated data and summarizes data. It provides the decision-making with clear, relevant and non-redundant information.

d) Data Warehouse Construction

Building a data warehouse system is different from developing transaction systems, whereby the requirements analysis process is supported by numerous methods. Many authors (Davis, 1993; Sommerville and Kotonya, 1998) have written publications that survey requirements management methods and requirement tools. Little has been done to compile the best practices of requirements specification independently of tools or methods. Up to now, the data warehouse design process has not been supported by a formal requirement analysis method, although there are some approaches to requirements gathering. List et al. (2002) introduced an iterative and incremental requirements development process for data warehouse systems, but they do not include an approach for managing the requirements. Inmon (1996) argues that a data warehouse environment is data driven, in comparison to classical systems, which are requirement driven and the requirements are understood after being populated with data and being used by the decision support analyst.

At this level we have data loaded from data marts in the data warehouse. In a physical data warehouse, copies, in some cases many copies, of data loaded from data marts are actually stored in a way that it is easy to access. It is important to notice that this stage will keep communication with the integration stage to reconcile the data loaded in the data warehouse. Access and performance issues are addressed at the physical level with the appropriate database designs.

At this stage developers of data warehouse could select the appropriate data warehouse in terms of distributions: a central data warehouse or a distributed data warehouse. Central data warehouses are what most people think of when they first are introduced to the concept of data warehouse. The central data warehouse is a single physical database that contains all of the data for a specific functional area, department, or organization. Central data warehouses are often selected where there is a common need for informational data and there are large numbers of end-users already connected to a central computer or network. A central data warehouse may contain data for any specific period of time. Usually, central data warehouses contain data from multiple operational systems.

Distributed data warehouses are exactly what their name implies. They are data warehouses in which certain components of the data warehouse are distributed across a number of different physical databases. Increasingly, large organizations are pushing decision-making down to lower and lower levels of the organization and in turn pushing the data needed for decision making down (or out) to the Local Area Network (LAN) or local computer serving the local decision-maker. Distributed data warehouses usually involve the most redundant data and, as a consequence, most complex loading and updating processes.

An important aspect at this stage is the definition of metadata. Metadata is "data about data", a catalog of information about the primary data that defines access to the data warehouse. It is the key to providing users and developers with a road map to the information in the data warehouse. Metadata comes in two different forms: Business and Technical (see chapter 3, section 3.8).

Business metadata serves a business purpose; it translates a cryptic name code that represents a data element into a meaningful description of the data element so that end-users can recognize and use the data. For example, metadata would clarify that the data element "WEB_W32" represents "Weekly Epidemiological Bulletin for Week 32". Transformational metadata serves a technical purpose for development and maintenance of the data warehouse. It maps the data element from its source system to the data warehouse, identifying it by source field name, destination field code, transformation routine, format, key, size, index and other relevant transformational and structural information. Good metadata is essential to the effective operation of a data warehouse and it is used in data acquisition/collection, data transformation, and data access.

The subject areas and the model that will support the data warehouse should be designed at this stage. At the end of the process we define the presentation server. Presentation server is the target physical machine on which the data warehouse data is organized and stored for direct querying by end users, report writers, and other applications.

e) Dimension Construction

This stage provides a front-end to the data warehouse to enable fast querying of information in the data warehouse. It is at this point that the data is presented and stored in a dimensional framework. If the presentation server is based on a relational database, then the tables will be organized as star or snowflake schemas (see chapter 3, section 3.7.1). If the presentation server is based on non-relational on-line analytic processing (OLAP) technology, then the data will still have recognizable dimensions. At this stage we construct the OLAP system. OLAP systems use data in multidimensional formats ("data cubes") to answer queries faster. The main OLAP data cube operations are the following:

- *Roll-up* **–** Refers to increasing the level of granularity along a dimension (with a hierarchy) (choosing "province" instead of "district" as a dimension). It is the same as re-aggregating data.
- Drill-down - Refers to decreasing the level of granularity along a dimension (choosing "district" instead of "province" as a dimension). It is the same as dis-aggregating data.
- *Pivoting* **–** Refers to creating cross-tab type data cubes by using two or more dimensions and creating a new view with an attribute for each grouping dimension and an additional attribute for the aggregate measure.

Fact table and dimension tables are created at this stage.

f) <u>Data Analysis and Use of Information</u>

The data analysis is the process of extracting mathematical patterns (i.e., analysis) from large sets of data to address "why" types of questions. Three levels of data mining have been defined in the data warehouse literature: simple analysis, intermediate analysis, and complex analysis (Linden, 1999). Simple analysis is the use of Online Analytical Processing (OLAP) cubes that present end-users with summarized data (such as count, sum, and average) that can be aggregated, sliced and diced, and drilled down to more detailed levels of granularity. Simple analysis also includes the use of SQL models to generate descriptive statistics related to simple patterns of data stored in relational databases. Intermediate analysis is the use of statistical analysis techniques to understand patterns and relationships in data. Complex analysis includes the use of artificial intelligence and machine-learning techniques.

In the same sense that there are lots of different ways to organize a data warehouse, it is important to note that there are increasingly wide ranges of end-users as well. In general we tend to think in terms of three broad categories of end-users: Executives and managers, "power" users (business and financial analysts, and engineers), and support users (clerical, administrative, etc.). Each of these different categories of user has its own set of requirements for data, access, flexibility and ease of use.

Decision Support Systems (DSS) is the utilization of the data in the data warehouse so as to facilitate and improve the decision-making process. DSS includes query and reporting, online analytical processing (OLAP), analytic applications, enterprise reporting, and data mining (hypothesis generation). This stage includes tools that allow users to select, manipulate and present data in different formats.

In terms of users, data warehouse users are made up of technical users, end-users, and data miners (Haley, 1998; Linden, 1999; Sakaguchi and Frolick, 1997). Data miners are skilled in analyzing data in the data warehouse using complex mathematical, statistical techniques, and artificial intelligence. Their goal is to discover the patterns and relationships to gain insight related to business problems, they also build predictive models to support the decision-making process. The data warehouse usage is defined in two levels that have been recognized as levels of decision support systems for many years: reporting and data analysis (Silver, 1991). Reporting includes the extraction and summarization of data to support "what or how much" types of questions of interest primarily to end-users. Reporting can be standardized where end-users have access to standard weekly report (e.g., Weekly Epidemiological Bulletin report), or *ad hoc* report where end-users or data miners may run a Structured Query Language (SQL) query (van den Hoven, 1998).

6.2.2 Choosing Database Technology for the Data Warehouse

The selection of data warehouse technology – both hardware and software – depends on many factors, such as (Inmon, 2000):

- Scalability – The volume of data to be accommodated;
- The speed with which data is needed;

- The history of the organization;
- Which level of data is being built;
- How many users there will be;
- What kind of analysis is to be performed; and
- The cost of technology.

The data warehouse hardware is typically mainframe[2], or client/server[3] hardware. The software that must be selected is for basic database manipulation of the data as it resides on the hardware (Inmon, 2000).

For the Ministry of Health in the proposed model I suggest both, mainframes or client/server hardware, for the following reasons: mainframes provide a central maintenance and control of data, business rules and operations, low cost departmental support and equipment.

The client/server architecture presents the following advantages: Centralization – access, resources, and data security are controlled through the server, Scalability – any element can be upgraded when needed, Flexibility – new technology can be easily integrated into the system and Interoperability – all components (clients, network, servers) work together.

In terms of software, there is a combination of different software, e.g. the database management system like Oracle or the SQL Server (see chapter 3, section 3.14). From the literature on data warehouse, different authors suggested that Oracle Data Mart Suite® is a good tool to be used during the data warehouse implementation project, in a way that this tool has the following preferable characteristics (Domenico, 2001):

- Provide a quick start of the project with focus on the data marts;

[2] A computer created for large corporations' one-way computing processes. Mainframes are associated with centralized rather than distributed computing.
[3] An interconnected network environment in which servers distribute processing power and software applications to workstations.

- Is simple to learn (Domenico spent no more than three months to learn); and
- The tool support the Data Mart and Data Warehouse life cycle.

I think that the most important aspect in the selection of hardware and software for data warehouse is the combination between the Relational Database Management System (RDBMS) and the hardware. Because the RDBMS physically sits on the hardware platform, there will be certain parts of the code that is hardware platform-dependent. As a result, bugs are often hardware dependent.

Popular relational databases are Oracle, Microsoft SQL Server, IBM DB2 and Sybase, and popular hardware platforms are Sun, Hewlett-Packard, Microsoft and IBM.

6.2.2.1 Data Warehouse Software Tools

The purpose of this stage is to identify candidate tools for developing and implementing the data warehouse data model and application architectures, and for performing technical and support architecture functions where appropriate. It is important to select candidate tools that best meet the technical requirements as defined by the data warehouse architecture.

It is important to stress that the process of selecting tools is often dependent on the existing technical infrastructure of the organization. It is recommended that a thorough evaluation of existing tools and the feasibility of their reuse be done in the context of all tool evaluation activities. From the literature review, I can conclude that in some cases, existing tools can be form-fitted to the data warehouse; in other cases, the organization may need to be convinced that new tools would better serve their needs. From the study in the Ministry of Health, in this case the second option will be the better selection. Using Oracle tools to implement the data warehouse in this case will be a better option and will give several advantages for the Ministry of Health (see chapter 3, section 3.14).

As a guide, it is important to notice that the tools may be categorized according to the following data, technical, application, or support functions: Source data extraction and transformation, data cleansing, data load, data access, security, configuration management, backup and recovery, disaster recovery, performance monitoring, database management, platform, data modeling and metadata management.

6.2.2.2 Data Warehouse Infrastructure

Data Warehouses are stored on client/server platforms, but they are often stored on mainframes as well. It should be pointed out that designs based on application-specific requirements and tactical technology selections are quickly invalidated, leading to substantial rework of the data warehouse infrastructure at greatly increased cost and time to delivery. When the inevitable (change in requirements) occurs, much of the infrastructure and design collapses, unable to meet user demands. Often, the entire structure must be demolished and replaced because the "damage" is so extensive. To void this situation, Ministry of Health must focus on building a data warehouse infrastructure that is flexible.

An application-neutral approach to the data warehouse design and selection of the data warehouse architecture that do not limit future flexibility are critical in achieving this.

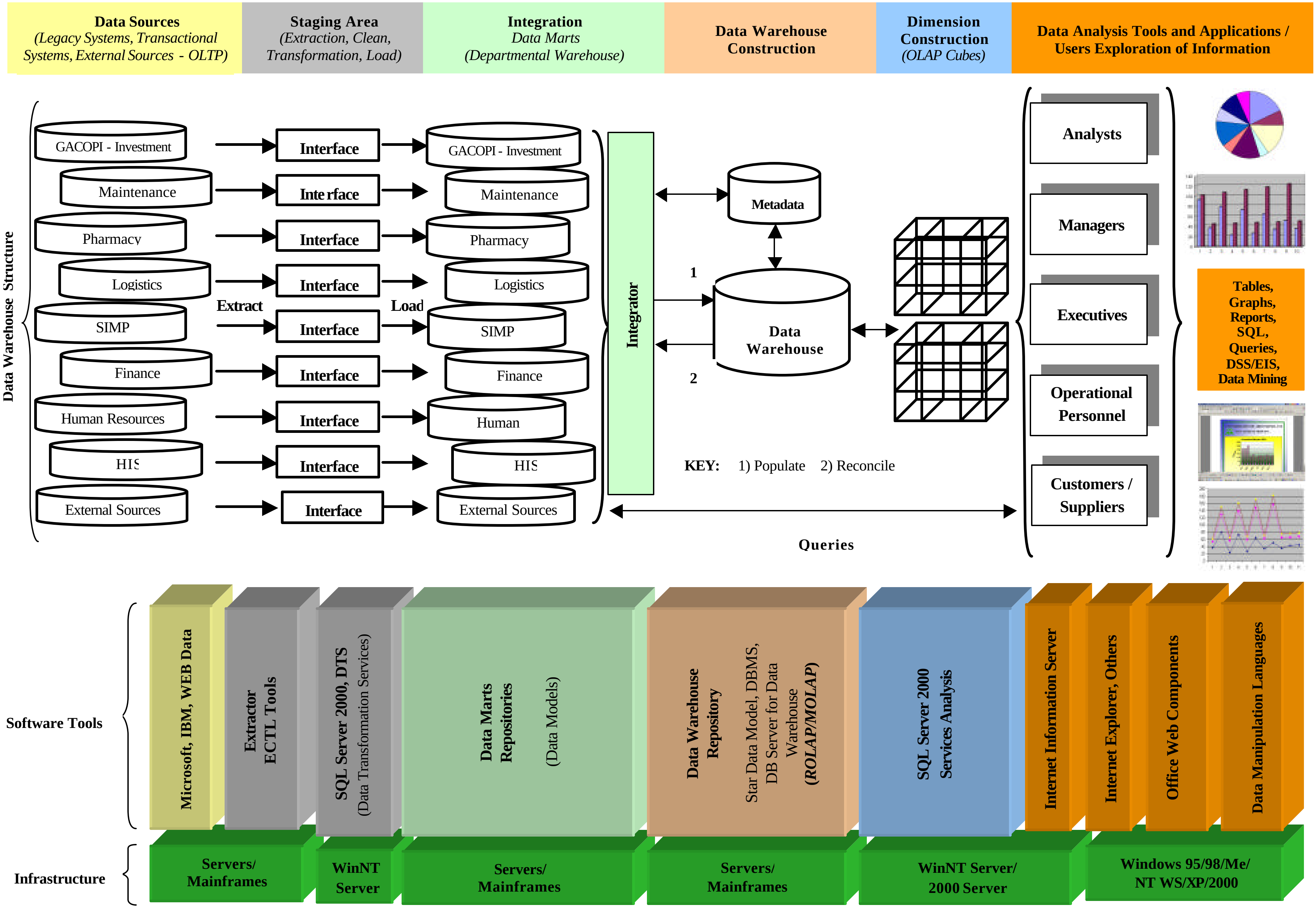
6.3 Summary

I chose the data warehousing approach in order to test if it is feasible to implement in the Ministry of Health. The selection of this approach takes into account the advantages and the disadvantages. Among the main advantages we have the simplicity, because the data warehouse allows existing legacy systems to continue in operation; better quality data; and it provides a unique organization data repository. The main disadvantages that the Ministry of Health can face are related to the time period that it will take to build the data warehouse, the budget involved in the construction process, and the training.

Another important aspect to look at is related to the definition of the essential data set. This aspect will raise implications for the design of the data warehouse data model. The essential data set will provide the data that the managers and decision-makers need. A combination of different data sets from different areas will provide an integrated data.

This chapter presented a description of different components of the proposed data warehouse architecture model, following the theory presented in chapter 3. Different phases of the data warehouse life cycle were covered. The projected data warehouse environment for the Ministry of Health allows that first, we define the data sources systems, then we extract, clean, transform and load data in the data marts, and finally we load data from the data marts in the data warehouse. The subsequent processes include access and analysis of data contained in the data warehouse. Finally, it seems that the adoption of Oracle tools will facilitate the data warehouse development.

Figure 6.2 - The Proposed Data Warehouse Architecture Model for the Ministry of Health.



Chapter 7

Discussion, Conclusions and Recommendations

7.1 Introduction

The main objective of the research was to present a data warehouse architecture model. The proposed model will contribute to the current integration process in course at the Ministry of Health. This chapter presents a discussion based on the objectives, the research questions, the literature review, the empirical findings presented in the different chapters of this thesis and focuses on the main conclusions, recommendations, contributions, limitations of the research study and suggestions for further research.

The strengths of the study are the proposed data warehouse architecture model that is simple and the study related to the information flow that provides a view of the current status of the information flow in the National Health Information System.

The weaknesses are related to the proposed model. The model lacks the full involvement of the Ministry of Health stakeholders due to their availability. The model was developed according to the available information deriving from the interviews, observations, questionnaires and documents analysis. Additional assumptions were based on the literature review.

7.2 Health Information Data Warehouse Model

7.2.1 Summary of the Model

Today, data warehousing has become an important strategy to integrate heterogeneous information sources in organizations. The proposed data warehouse architecture model presented in chapter 6, section 6.1.4.2, page 149, integrates different heterogeneous health data sources. Selected data sources include the following systems: GACOPI (Gabinete de Cooperação de Projectos de Investimento), Maintenance, Pharmacy, Logistics, Sistema Integrado de Planificação (SIMP), Finance, Human Resource, Health Information System and External Sources. Six different stages form the proposed model: Data Sources, Staging Area, Integration, Data Warehouse Construction, Dimension Construction and Data Analysis Tools and Applications. For each stage the suggested software tools are presented for implementation and infrastructure.

During the research, to design the data warehouse model I followed part of the data warehouse life cycle as the methodological approach (see Chapter 6).

7.2.2 Potential Benefits from the Model

From the proposed model, end-users can perform different tasks. The model covers only part of the selected legacy systems existing in the Ministry of Health, so the tasks will be around the selected systems. The proposed model will integrate different subject areas as presented in figure 6.1, chapter 6. Users may combine data from different areas to produce useful information.

The usability of the model can be demonstrated through the example of data integration from Pharmacy, Sistema de Informação de Saúde (SIS), External Sources (population data) and Finance areas. From the Pharmacy we can collect ***medicines*** (type and quantity), from the Finance we can learn about the existing ***budget*** (amount, donor and period), from Sistema de Informação de Saúde (SIS) we can have information about the

programs (Mother and Child Health and Immunization) that need the medicines and from the External Source, we need to know the ***population*** (children, women, men and age). This combination is useful for the management and decision-making process. Managers and decisions-makers need to see the performance of different services areas with integrated data and not isolated data, due to the fact that different services areas exchange data.

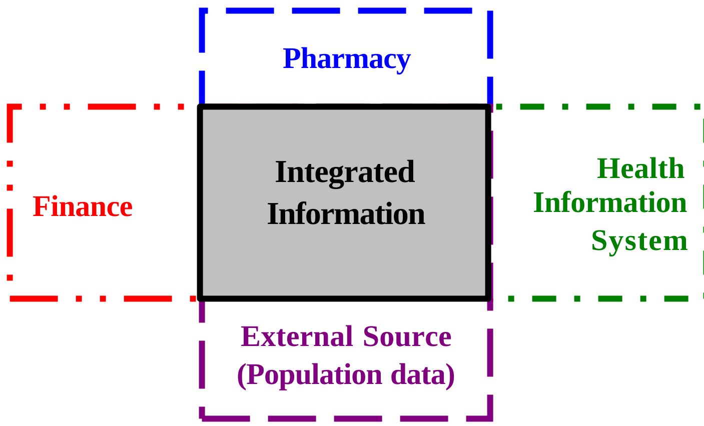


Figure 7.1: Information integrated from different areas.

Different authors argue that the data warehouse usage outcomes range from job productivity, decision efficiency, decision effectiveness, and knowledge discovery. Data warehouses have enhanced job productivity by enabling users to perform decision-making tasks faster and more comprehensively (Wixom and Watson, 2001), as well as reducing time and effort required to access data (Graham, 1996). Data warehousing usage can also create knowledge to support better quality decision-making (Brohman, Parent, Pearce and Wade, 2000; Cooper, Watson, Wixom and Goodhue, 2000). Different researchers show that data transformation and analysis tools and "intelligent" data mining tools help data miners to create knowledge from the data warehouse.

In short, the main benefits that the proposed data warehouse model can provide to the Ministry of Health are the following:

- *Single source of integrated and subject-oriented strategic data and information*. In the Ministry of Health the data warehouse will contain data and information to support the management decisions. The data warehouse will provide a unique, subject oriented, integrated, time variant and non-volatile source of data.

- *Complete view of existing data inside the Ministry of Health*. With the islands of information that the Ministry of Health has at the moment, it is impossible to have a complete view of the existing data. Using the data warehouse the Ministry of Health will integrate different systems. For the integration process all departments need to provide their data and at that time the Ministry of Health will have a complete view of the existing data.

- *Better data for the decision-making process*. Using the data warehouse the Ministry of Health will access data relations that are not easily accessible with a simple database system. Data warehouse usually focuses on the process of discovering valid, comprehensible, and potentially useful knowledge from large data sets with the goal to act on this knowledge in decision-making (Wang and Weigend, 2003). Figure 7.1 illustrates this benefit.

- *Data from different departments and areas in a unique data standard*. In the Ministry of Health each department are using different data standards (see chapter 3, section 3.3.2) even for the same data. The Ministry of Health need to establish standards for their data, different departments need "to talk the same language".

- *Flexible access to current data residing in the different data sources*. I have noticed that in the Ministry of Health, the problem of dispersed data is not different from other public organizations of the same type. In the Ministry of Health, useful data can be found in various applications, such as spreadsheets, but it is often an arduous and manual task to combine the information in these sources. Data warehousing systems are typically intended for helping decision-making process, based on combining large amounts of historical data from dispersed operational databases and the Ministry of Health should take advantage of that.

- *Ability for different departments and partners to share data and information*. The Ministry of Health strategy needs to consider the ability of different departments, levels of the Ministry of Health and partners to cooperate. The Ministry of Health needs to set out the policies and technical standards for achieving interoperability to enable a seamless flow of information and data across the public health sector. Data warehouses and marts give organizations the ability to consolidate information from multiple, heterogeneous production systems into a common, integrated database environment. Inmon (2003) argue, that information in this environment can then be organized, summarized, regularly refreshed, and used to great advantage by a wide variety of powerful analytical applications in different departments within and outside the organization.

7.2.3 Potential Limitations of the Model

While there are many positive points to adopting a data warehousing system, there are also drawbacks. The drawbacks can be the costs of a data warehousing system that is high. These costs are related to limited value of the data, irrelevant data collected, data collection costs, and costs of human capital. These costs are all potential drawbacks to the system. Data that is collected in a data warehouse is typically produced from the internal processes of the organization using historical data. This information has limited value because the purpose of this information does not, always contribute to forward-looking decision-making. Creating excess data for the data warehouse is another drawback to adopting this system. This typically occurs when the data warehouse system has been established. Organizations find that they can add unlimited data to a data warehouse. By adding unrelated data to the data warehouse, the costs of retrieving data and transforming it into a useable form are increased.

In the context of the Ministry of Health, I think that collection of irrelevant data may become a smaller problem in the future, because the Ministry has started the process of definition of essential data set at health facility, district, province and national levels, but if this process fails the whole process of data warehouse construction will be compromised. Another data warehouse drawback is that the costs to collect and deliver data in a timely fashion may be too expensive for the Ministry of Health. Cleansing the data is a vital part of this process.

Human capital is an expensive part of any database system. This is especially true with a data warehousing system. Many consultants and employees, who will use or maintain the system, do not have ample experience with the data warehousing life cycle and the associated processes. There are not many people that have extensive experience with data warehousing because it is a relatively new data storage system.

In the Ministry of Health the human capital is a scarce problem. Currently, the Health Information Department is running with two or three workers with background on computer science. Other staff of Health Information Department includes three consultants specialized in public health. To work on a data warehouse project the Ministry of Health needs to establish a team. There are two areas of discussion relating to the data warehouse project team: First is whether to use external consultants or hire permanent employees. The second is on what type of personnel is recommended for a data warehouse project. Using external consultants has the advantage that, they are usually more experienced in data warehouse implementation.

Different authors argue that, as a matter of fact, even today, people with extensive data warehousing backgrounds are difficult to find. So, when there is a need to ramp up a team quickly, the easiest route to go is to hire external consultants. The same authors argue that using permanent employees is less expensive and is a much more economical option. The following roles are typical for a data warehouse project (Lee, 2002):

> ***Project Manager***: This person will oversee the progress and be responsible for the success of the data warehousing project.
>
> ***Database Administrator***: This role is mainly for keeping the database running smoothly. Additional tasks for this role may be to plan and execute a backup/recovery plan, as well as performance tuning.

Technical Architect: This role is responsible for developing and implementing the overall technical architecture of the data warehouse, from the backend hardware/software to the client desktop configurations.

Extraction Transformation Load (ETL) Developer: This role deals with planning, developing, and deploying the extraction, transformation, and loading routine for the data warehouse.

Front End Developer: This person is responsible for developing the front-end, whether it is client-server or over the web.

Online Analytical Process (OLAP) Developer: This role is responsible for the development of OLAP cubes.

Data Modeler: This role, is responsible for taking the existing data structure in the enterprise and model it into a schema that is suitable for OLAP analysis.

Trainer: A significant role is that of the trainer. After the data warehouse is implemented, a person on the data warehouse team needs to work with the end users to get them familiar with how the front end is set up so that the end users can get the most benefit out of the data warehouse system.

The above list describes roles, and one person does not necessarily correspond to only one role. In fact, it is very common in a data warehousing team to find a person taking on multiple roles. In Mozambique it is difficult to find people with these skills. Available people are working in different organizations. In order to solve the problem of Information Technology (IT) human resources the Ministry of Health need to contract people and use the knowledge of the consultants.

Different case studies from the literature show that many organizations succeed with their data warehouse projects, whereas others fail. The exact failure rate is not known. Obstacles to success that have been observed (Gray and Israel, 1999) include: Data quality, transforming legacy data, business rule analysis and business data modeling. These obstacles can, and often do, lead to failure. Reasons for these failures observed (Gray and Israel, 1999) include:

- *Lack of skills in managing large-scale IT projects.* The data warehouse project requires a great persistence of managers. Currently the Ministry of Health has lack of IT staff, so they will need to invest in new IT staff and choose the coordinator of the project or they can use the consultants experience to coordinate the project.

- *Inadequate planning*. "Many information technology initiatives are not successful because inadequate planning and analysis is done before getting started" (Cohen, 1999). The Ministry of Health need to define and document the requirements for the data warehouse project and continuously review them to assure that they have not changed. Cohen argues that, "any project should take an iterative approach allowing for ongoing user usage and evaluation to refine the project and ensure that the outcome for the project will support the cost".

- *Political, social and cultural effects of the data warehouse*. Political, social and cultural requirements are those associated with the ethnic, geographic, religious and interpersonal relationships that affect those who participate in and/or derive value from your data warehousing project (Borland and Murphey, 2001). If the Ministry of Health fails to fully assess and address these considerations may cause considerable disruption and dissatisfaction with the data warehouse project.

- *System integration failures*. System integration failure occurs due to the failures in designs (incomplete specifications), poor communication among members (managers, analysts, designers, programmers) in a team. Other issues include the data sources that have the data set to be integrated. Each department at the Ministry of Health should provide the set of data to be integrated.

- *Lack of sufficient funding*. The Ministry of Health usually works with different partners in the implementation of the projects. Usually, these partners such as WHO and donors funding in collaboration with the Government of Mozambique, so I think, depending on the proposed budget, that the Ministry of Health will find funding agents for the data warehouse project.

- *Inability to provide desired user access*. Access to integrated data and information to support management process improvements is the root problem, according to different authors. Interoperability of the infrastructure to gain such access is the main technical challenge for the Ministry of Health.

As a conclusion, the costs associated with the data warehousing system are the major drawbacks to implementing the data warehouse. These drawbacks include limited value of the data, irrelevant data collected, data collection, and costs of human capital. To avoid or minimize the drawbacks the Ministry of Health need to:

1) Finalize the process of definition of the essential data set for the healthcare sector in Mozambique;
2) Focus on the data quality; and
3) Invest in IT human resources.

7.2.4 How to Implement the Model

The starting point of the implementation process is the definition of the data sources for the data warehouse project. At the same time, we need to select the data warehouse architecture approach. The Ministry of Health can adopt the bottom-up approach suggested by Kimball (see chapter 3, section 3.4.2, figure 3.3). This approach focuses on the construction of the departmental databases, the data marts. The next step will be the system and requirement analysis. Having the data sources, we will start with the extraction, transformation and cleansing processes. The data marts construction will be the subsequent step. Data marts accommodate the extracted data. Different departments will have its own data mart. At this stage, we need to break the project up into small achievable parts, so this will let each department present their data to be loaded in the data warehouse. Using this approach we can have different projects running in each department and at the end we will integrate different data marts (see, e.g. White, 1996). In the data warehouse construction stage we need to define the dimensional data model that should be a star model. The final step will be the definition of different types of views for data contained in the data warehouse, so that, users can explore the information using different analysis tools and applications.

In terms of tools, the Ministry of Health can use different software tools such as Oracle or Microsoft SQL Server. Another option is to test all tools used by Domenico (2001) to implement the data warehouse. It is important to note that the process of selecting tools is often dependent on the existing technical infrastructure of the organization. Currently, the Ministry of Health does not have the proposed tools.

7.2.5 What to do to extend the proposed model?

From my point of view, I think that the following steps need to be taken in the future in order to extend the proposed data warehouse architecture model:

- Deep analysis of data sources (legacy systems) selected for the data warehouse project;
- Definition of the data model to be implemented in each data mart;
- Design of the model of data integration;
- Design of data dimension needed for users to perform their analysis.

From the above, I have started with the first step together with four colleagues of the Masters in Information Systems at the University of Oslo and Eduardo Mondlane University. We conducted a study called "*Analysis of the Health Information Flows in the Ministry of Health, Mozambique*". The main results of the study are related to the information flow among different levels of the National Health Information System in Mozambique with focus to the data collection, the data process, the data analysis, the data presentation, the data use, and the data interpretation. Most of the findings are discussed in the chapter 5.

7.3 Answers to Research Questions

In this section, I present the answers to the research questions formulated in chapter 1, section 1.5 based on the study findings and in the literature. For each question, I present the summary of the literature related to the question, the study findings and finally a brief relationship between the study findings and the literature.

7.3.1 *How to collect or extract health data from heterogeneous sources?*

Moura and Victorino (2003) point out that the process of extracting data from heterogeneous sources is still very difficult, due to important aspects such as: need to process huge amounts of data of heterogeneous and complex types; data are located in distributed and heterogeneous repositories, hard to access (data are private or at a high cost), lacking a standard storage format; and inaccuracy of the information.

SILICO (2001) argues, that the environment composed by potential sources of data is a complex data environment, because data sources can be structured, semi-structured and unstructured and relational databases or flat files. So the process of data collection will also be complex and difficult. Different authors, e.g. Inmon (2000), suggest that each data source needs to uses different applications to collect and extract data. In this process, data are normally collected in one of two ways. Most commonly, information is downloaded into a separate database. Alternatively, an application is designed to interact with legacy systems (Levine, 2002).

From the research study at the Ministry of Health in Mozambique, I found that the collection process of heterogeneous health data could be achieved by selection of the data sources and a detailed analysis of the status of all data sources that will contribute to the data warehouse. The selected data sources should play an important role in the decision-making process. Potential data sources in the Ministry of Health include GACOPI-Investment, Maintenance, Pharmacy, Logistics, SIMP, Finance, Health Information System, Human Resources systems and External Sources. However, the data sources described in chapter 6, section 6.1.2 do not encompass the full extent of available data sources in the Ministry of Health.

After the data sources selection, we need to select the correct software to collect and extract data from different data sources. In the study, we choose the first approach suggested by Levine (2002), downloading information into a separate database. We used the *Database Filer 1.0* application. This application was designed by computer science students to extract data from the Computerized Sistema de Informação de Saúde (CSIS) in the Ministry of Health. This software has the ability to read data directly from the data sources and store data extracted in the software relational database management system. (see Annex I).

Levine (2000) argues, that it is essential to verify the quality of potential data sources. Data structure and content errors need to be identified and corrected as early as possible in the data conversion process, otherwise data quality problems will have a negative impact on downstream processes such as the transformation and load jobs used to extract source data for loading into the staging area.

A strategy to obtain better results in the extraction and collection processes, is first, analyze the data sources carefully, because the analysis will give the status of all systems. The second point is related to the software to be used. In the data warehousing market we can find different extraction tools like *Informatica*, *Datastage*, *Datajunction*, *DataMirror*, or *SAS Warehouse Administrator*, but many authors (Inmon, 2000 and Levine, 2002) argue that we need to construct our own extraction software in order to obtain good results, otherwise we can use the tools presented above and spend more resources (time and money) to achieve the extraction process objectives.

From the literature and research study findings, I can conclude that to collect or extract health data from heterogeneous sources, we need to first, identify the data sources, secondly, make a deep analysis of the identified data sources; third select or construct a software to perform the extraction process and then select a target database to store the extracted data. At this time it is important to focus on data quality.

7.3.2 *How to clean and transform heterogeneous health data?*

> "Data warehouses require a great deal of investment in data cleaning and transformation before the data can be loaded into the warehouse".
>
> (SILICO, 2001)

The data cleansing process consists in detecting and removing errors and inconsistencies in data to improve its quality. It is a process to check data for adherence to standards, internal consistency, referential integrity, valid domain, and to replace or repair incorrect data with correct data. Jermyn et al. (2003) argue, that inappropriate cleaning process can damage the data.

In the research study, the cleansing and transformation processes focus on the Computerized Sistema de Informação de Saúde. From the system were selected four forms, A04, B06, B07 and B07 (see chapter 6, section 6.1.3.2). The main data problems that occurred in datasets arise from the absence of data or missing data, presence of dummy values, violation of validation rules, business rules and calculation errors.

The cleansing and transformation is a time consuming process, because we had to go through each record to search for data to fill the fields with missing data, correct the dummy values and correct general errors. In this process, help provided by workers was useful. The process consisted in sitting together in front of the computer with health workers to correct data.

Jermyn et al. (2003) argue that good data preparation is a key prerequisite, the foundation to successful data analysis. Conventional wisdom suggests that data preparation takes about 60% to 80% of the time involved in a data analysis exercise.

> "For data warehouse clean data is useful data. Many transaction systems have some extent of data corruption either due to errors in the system or due to the data entry problems. Data cleaning being a continuous process through out the life of the data warehouse involves analyzing the data with respect to possible variances of

> the clean data and devising rules to clean the data according to the requirement of the data marts and data warehouse".
>
> (Rasheed, 2003).

This is usually the most critical aspect of the data warehouse process and the key factor in determining the data warehouse success or failure. It is paramount that clean data is provided for strategic decision-makers (Inmon, 2000).

From the literature and research study findings, I can conclude, that the cleansing and transformation process are of extreme importance. The process starts with the extracted data loaded in a database in the stage area and finishes with use of different approaches to correct the main data problems. Data with errors can lead to wrong decisions; so, much effort must be expended in the cleansing and transformation process.

> "Errors can occur for example because of excessive working hours or poor systems. When it occurs, it is necessary to distinguish between the normal level of human error and error caused by incompetence or other circumstances".
>
> (BMA, 1999)

The Ministry of Health can reduce or avoid data errors by:

- Defining different procedures related to data collection process; and,
- Using the technology.

Examples of Niassa and Gaza provinces show that considerable improvements occurred in relation to data errors. In these provinces, the Ministry of Health started a project with the aim to improve accuracy, timeliness, usefulness, and actual use of health information. Activities comprised introduction of computers at district level, installation of the software District Health Information System (DHIS), and training of health information managers, clerks and nurses. Computerization of the Health Information System (HIS) in combination with training has improved data quality. The software needs continuous support. Computerization reduces the time taken for data consolidation and compilation compared to processing by hand-held calculator. It improves data quality (summation errors are avoided).

7.3.3 *How to integrate heterogeneous health data?*

In order to integrate heterogeneous data, we have to follow the data warehouse life cycle. The starting point is the definition of the correspondence/mapping between data in different formats in different data sources. The next step is the development of an interface that will allow users to load data from data sources to data marts. At this stage, loaded data need to be exported to the data warehouse (Inmon, 2000).

The main finding of my research is a data warehouse architecture model for the Ministry of Health. An extensive description of the proposed model is presented in the chapter 6. Other findings are related to data collection, information processing, analysis, presentation, interpretation and use of health information (see chapter 5, section 5.3). The idea of building the data warehouse as a central data collection made available for decision support applications in an organization is widely accepted, and the Ministry of Health is not far from this point of view.

The proposed data warehouse architecture model previews data integration from different data marts. Data integration process consists in the creation of the integrator that will read data from different data marts.

7.3.4 *What type of methodological approach should be used to design a data warehouse model for the health sector?*

The data warehouse life cycle is the main methodological approach used in the design of the data warehouse model for the health sector. In this research, I used and covered parts of the data warehouse life cycle suggested by Flanagan and Safdie (1998), presented in chapter 3, section 3.5. Depending on the area where we want to implement the data warehouse we need to adapt the life cycle to the conditions of the environment where we want to implement. The stages of the life cycle covered are the following: document organization legacy systems that will take part in the data warehouse project, model the data warehouse and clean data. The steps followed are related to the extension of the data warehouse life cycle: definition of the data sources, extract, transform and clean data, create the data model and the physical design of the data warehouse and choose the database technology for the data warehouse.

7.3.5 *How can data quality influence the decision-making process?*

Different authors (SILICO, 2001; Inmon, 2000a, and Watson, 2001) argue that typically data quality problems occur in the source systems. From the study, I found that the data sources are the main source of poor data in the Ministry of Health (see chapter 5). The Computerized Sistema de Informação de Saúde at different levels of the National Health Information Systems stores poor data.

The data quality will influence the decision-making process in a way that, managers and decision makers access and use poor data in their decisions. That happens because they do not perform data analysis before the decisions is taken.

Different authors state that data stored in the data warehouse help to improve the effectiveness and efficiency of decision-making process. There are two prerequisites for maximizing data quality: 1) The organization must consider its data to be as important as any other asset (e.g., human resources). Data quality needs to be part of every health workers job and 2) Employees should be assigned responsibility for data. The Ministry of Health needs to establish and enforce policies concerning the acquisition, use and dissemination of the data.

To summarize, section 7.3 presents a discussion related to the research questions formulated in chapter 1, section 1.2. Based on the research findings, I can conclude that creating a data warehouse is an organizational effort. The data collection and extraction processes proved that one challenge for the Ministry of Health is on the selection of the correct software that will extract data and the major challenge will be the improvement of the data quality.

The data cleansing and transformation processes should be the focus of the whole process of data warehouse construction, because data are the main element of the data warehouse. Decision-makers need data, so data should have quality. The Ministry of Health can ensure data quality in the cleansing and transformation processes.

Finally, different authors argue that the data integration is a competitive differentiator that can enable managers to respond more quickly to different solicitations. In the case of the Ministry of Health, due to the heterogeneous data, complete integration architecture will necessarily include multiple technology approaches to match the integration issue.

7.4 Summary of Findings and Conclusions of the study

The major result that I presented in this study is a data warehouse architecture model (see chapter 6) that allows the integration of data and information at the Ministry of Health. The proposed model integrates different data sources from different departments and will constitute the central data repository for the Ministry of Health. During the research, special attention was paid to the information flow at and among district, province and national levels (see chapter 5) of the National Health Information System. From the research study, I can conclude that:

(1) The Ministry of Health keeps poor historical health data. Possibly that occurs due to:
   - No data quality control at any level of the National Health Information System;
   - Lack of feedback from the higher levels, managers and decision-makers;
   - Excessive amounts of data being collected (there is no essential data set for all health facilities, districts and provinces);

- Data aggregation that conduct to data details lost;
- Inexistence of tools and procedures for data collection process;

(2) There are some indications that poor data are also related to:
   - Incorrect management of the existing human resources at the different levels of the National Health Information System;
   - Inexistence of basic working conditions. Some health facilities and health directorates are running without the basic conditions of work – electricity, water and communications;
   - Misuse of the available system, District Health Information System (DHIS) in districts where the system is installed;

(3) There are some evidences that health workers do not have the culture of information use, so they do not use information produced for the planning, monitoring, evaluation of services and decision-making processes;

(4) The Ministry of Health is not monitoring the progress of the District Health Information System (DHIS) at different districts;

(5) There are several evidences that the Ministry of Health have "islands of information". The reasons for that are:
   - Each department is running their own systems and databases;
   - Inexistence of share mechanism. Different departments are not sharing the information produced in each department;
   - The Ministry of Health do not have a central data repository;

(6) Creating a Data Warehouse is an organizational effort. Interactions between different departments and units and the IT department must be established.

It seems that the conclusions from this study conducted in Gaza and presented above are the mirror of what happens in other provinces in Mozambique. One study conducted by

Health Information System Project (HISP) in Niassa proved that the situation is the same as in Gaza.

7.5 Research Contributions

This thesis has demonstrated how a simple data warehouse model can be used in the Ministry of Health, for integration purpose. It is to be hoped that this study can help to shed some light on process that is currently running in the Ministry of Health, the process that will culminate with the integration of different heterogeneous systems. The proposed model will integrate data from different subject areas to produce useful information.

7.5.1 Theoretical Contribution

One of the three main goals of the research described in this thesis was to propose a data warehouse architecture model to the Ministry of Health. The approach used for the construction of the data warehouse did not follow a traditional approach of systems development. The development of the data warehouse is data oriented and is not like other systems that are requirements oriented. Inmon (2000a) mentions that we need to populate the data warehouse in order to gather the data warehouse requirements. This occurs because the data is the main focus of the data warehouse.

Different authors (Inmon, 2000a; Watson, 2001; SILICO, 2001) argue that the data sources sometimes are the source of errors. In my study, I observed that the quality of data used is poor. I found that the data sources are the main source of poor data (see chapter 5), so, my findings do not totally corroborate with the findings of cited authors. So for making data warehouses for health sector, the quality of the data sources have to be checked carefully.

In order to avoid the poor data, systems using validation rules should indicate to the users the existence of an error in a specific field so that a suitably authorized person can later correct this error. Careful checking will also help to reduce the problem, but will not eliminate it. Another approach is to have the data entered independently by two people and to check for discrepancies. If errors are rare and randomly distributed, this is effective. It is also, however, expensive, which is the reason it is not done for every entry, only for specific entries. Watson (2001) argues that in order to avoid errors in a large scale; we need to select the correct extraction, transformation and load (ETL) software, purchasing or write the software.

Watson (2001) state that the cleansing and transformation process is, complex, time consuming and expensive. The statement is in line with my findings. In my study, the cleansing and transformation was a time consuming processes, because we had to go through each record to search for data to fill the fields with missing data, correct the dummy values and correct general errors. In this processes, help provided by health workers in the Ministry of Health, at the province and districts was useful. The process consisted in sitting together in front of the computer with health workers to correct data. This way of correcting data could be used when we have non-complex data and less volume of data.

The proposed data warehouse model presents six (6) different stages of the construction process and include the software tools and the infrastructure for each stage. In previous studies, different authors presented only two stages in the data warehouse architecture model: integration stage (data marts – departmental warehouse) and the data warehouse construction stage. The proposed model focuses on the architecture and does not incorporate the business perspective like other models presented by different authors.

To implement the data warehouse the Ministry of Health will faces different types of challenges, including organizational, technical and operational challenges. Organizational challenges include the existence of different projects running at the same time. These projects are related to the data integration at the Ministry of Health and are financed by different donors. Few Information Technology human resources and communication infrastructure at the Ministry of Health are other organizational challenges. The technical challenges include the selection of software that will allow heterogeneous systems to communicate. Integrating but maintaining "business as usual" will be the main operational challenge.

7.6 Conditions that the Ministry of Health need to establish

In order to implement the data warehouse project the Ministry of Health needs to create the following conditions:

- Perform deep analysis and document the existing legacy systems;
- Analysis of the potential data sources for the data warehouse project;
- Improve the data quality, because the data warehouse focuses on data to work properly and data is the main tool for the decision-makers during the decision-making process.

Other conditions include:

- Reformulation of the current procedures or mechanisms of data collection and data quality control;
- Establish the IT team.
- Involvement of the top managers in the project.

7.7 Recommendations for the Ministry of Health

The recommendations are related to the empirical findings and to the proposed data warehouse architecture model. The Ministry of Health will gain benefits in their decision-making process if they can improve the quality of their data.

To avoid or minimize several drawbacks during the data warehouse implementation the Ministry of Health needs to, finalize the process of definition of the essential data set for the healthcare sector in Mozambique, needs to establish mechanisms and procedures to avoid poor data quality, and needs to invest in IT human resources.

Based on the proposed data warehouse model, I suggest that the Ministry of Health start the construction of the data marts from those systems that are very important in the decision-making process. The best strategy to construct a data warehouse is to start with departmental data marts and then the central repository, the data warehouse. Each department in the Ministry of Health should construct the departmental data warehouse. Data marts will be constructed around a specific subject area. Due to the fact that subject areas can relate with each other, data warehouse will integrate them. From the research study it was also possible to notice that the health workers do not perform the data quality control, only at districts with the District Health Information Systems (DHIS). Based on this observation, I think that the data warehouse development in the Ministry of Health should involve and start with data from sources that has reasonable data quality control and different mechanism for data quality control need to be established.

Data warehouse is a technology that several organizations are using and it provides good solutions in information recovery. The proposed model will, in some way, facilitate the decision-making process, by providing consistent and coherent data to the decision makers. The implementation of the model will require a greater collaboration of different stakeholders in the Ministry of Health; particularly from the administrative personnel, managers, health workers, information and technology (IT) staff and consultants, so the Ministry of Health need to involve them in the data warehouse project.

In order to run the data warehouse project the Ministry of health need to investment in IT personnel. Build a team of internal consultants with appropriate subject-matter expertise and internal staff familiar with in-house systems will be one option. Another option will be outsourcing of the data warehouse development.

Other recommendations include the definition of the essential data set for all health facilities, districts and provinces and definition of strategies for feedback at all levels.

7.8 Further Research

As a result of the research, I have learned that there is more work that can be done with the proposed data warehouse architecture model, both in practice and in an academic setting. To build a more complete model we need to define a project and a team, where each one should have a role, e.g. project manager, database administrator or trainer (see chapter 7, section 7.2.3). It would also be interesting to look at the entire problem from a different perspective, for example, use a specific tool (e.g. Oracle or Microsoft SQL) and start with the development of the data marts and see the results then according to the results, construct the data warehouse. For that we need a serious involvement of different stakeholders in the Ministry of Health.

References

Anjos, S.J., 1999. Uma contribuição para a arquitectura de infromações estratégicas (AIE) para sectores de pesquisa em Universidades Brasileiras. Thesis (Masters). Universidade Federal de Santa Catarina, Florianópolis, Brazil.

Barker, C., 1983. The Mozambique pharmaceutical policy. Lancet, 1 October: 780-782.

Barker, K., 2000. Data Warehousing Overview. University of Calgary, Alberta, Canada.

Becker, H., 1998. *Tricks of the Trade: How to think about your research while you are doing it*. Chicago and London: The University of Chicago Press.

Bell, J., 1993. Doing your research project: A guide for first time researchers in education and social science. Buckingham: Open University Press.

Boehm, B. W., 1979. *Software Engineering: R &D trends and defense needs*. Cambridge, MA: MIT Press.

Borland, R. and Murphey C., 2001. Design Requirements for the Data Warehouse: Part VI [online]. Published 16 August 2001 on http://www.datawarehouse.com. Accessed from: http://www.datawarehouse.com/iknowledge/articles/article.cfm?ContentID=1556 [Accessed: 14 April 2003].

Braa J., Nampossa, J., Macome, E. and Mavimbe, J., 2000. A study of use and diffusion of information and communication technology at district and provincial levels in Mozambique. Proceedings of the IFIP. WG 9.4.2000.

Braa, J., Monteiro, E. and Sahay, S., 2002. Networks of action: Sustainable health information systems across developing countries. Submission for special issue on action research, Management Information Systems Quarterly (MISQ), forthcoming.

Bradburn, N. M., 1982. Question-wording effects in surveys. In N.M. Bradburn (Ed.): New directions for methodology of social and behavioral science: Question framing and response consistency, Jossey-Bass, San Francisco.

Brohman, M. K., Parent, M., Pearce, M. R. and Wade, M., 2000. The Business Intelligence Value Chain. *Proceedings from the Thirty-third Hawaii International Conference on System Sciences (HICSS).*

Brohman, M., Parent, M., Goodhue, D. and Pearce, M., 2002. Using task-technology fit to explain knowledge creation in data warehousing. Ontario, Canada: The University of Western.

Bruun-Lie, D. A., 1998. *Systems development on Internet*. Thesis (Cand Scient). Department of Informatics. University of Olso, Norway.

Budde, R., Kautz, K., Kuhlenkamp, K. and Züllighoven, H., 1992. Prototyping, In *Prototyping. An Approach to Evolutionary System Development*, p.33-48, Springer-Verlag, Berlin.

Bødker, S., 1987. Prototyping revisited: design with users in a cooperative setting. Proceedings of the 10th IRIS conference. p.71-92.

Cliff, J. and Noormahomed, A. R., 1988. Health as a target: South Africa's destabilization in Mozambique. Society of Scientific Medicine, 1988; 27: 717-22.

Coffey, A. and Atkinson, P., 1996. *Making Sense of Qualitative Data: Complementary Research Strategies.* London: SAGE Publications Ltd.

Cohen, M. 1999. Data warehousing for improved decisions [online]. Available from: http://www.physiciansnews.com/computers/899.html [Accessed 14 April 2003]

Cooper, B. L., Watson, H. J., Wixom, B. H. and Goodhue, D. L., 2000. Data Warehousing Supports Corporate Strategy at First American Corporation. *MIS Quarterly* (24:2), p.547-567.

Data Warehouse Team, 2001. *A brief history of Data Mining* [online].
Available from: http://www.datawarehouse.inf.br/. [Accessed: 24 November 2001].

Davis, A. M., 1993. Software Requirements: Objects, Functions, and States. Prentice Hall.

Domenico, J. A., 2001. *Definition of a Data Warehouse environment in a High Education Institution*. Thesis (MSc). Universidade Federal de Santa Catarina, Florianópolis, Brazil.

Flanagan, T. and Safdie, E., 1998. *The DNA of Data Warehousing*. The Applied Technologies Group, Inc.

Floyd, C., 1984. *A systematic look at prototyping*. In Approaches to Prototyping. Springer-Verlag, Berlin, p.1-18.

Gatziu, S., Jeusfeld, M., Staudt, M. and Vassiliou, Y., 1999. Design and Management of Data Warehouses. Report on the DMDW '99 Wokshop.

Government of Mozambique, 1998. *Understanding Poverty and Well-Being in Mozambique: The First National Assessment (1996-97).* Maputo. Mozambique.

Gomes, A. and Johnson K., 1994. Creation of a centralized health data bank for Mozambique: Feasibility study report. Maputo.

Gray, P. and Israel, C., 1999. *The Data Warehouse Industry*. California. Center for Research on Information Technology and Organizations – University of California.

Hackney, D. and Allen, P., 2001. *Seven Deadly Sins Data Warehousing*. 1st edition. Addison-Wesley.

Haley, B. J., 1998. *Implementing the decision support infrastructure: Key success factors in data warehousing.* Athens. University of Georgia.

Hayes, H. A., 2003. *IBM Software, Creating a flexible infrastructure for integrating information*. IBM Silicon Valley Laboratory. GC18-7560

Hearling, Kurt. 2001. *Data Mining, CRM, Decision Support and Database Marketing* [online]. Available from: http://www3.shore.net/~kht/text/dmwhite/dmwhite.htm. [Accessed 24 November 2001].

Horwitt, E., 1999. Common Information Model [online]. Available from: http://www.computerworld.com/networkingtopics/networking/story/0,10801,43442,00.html [Accessed 14 April 2003].

Inmon, W. H. and Hackthorn, R. D., 1994. *Using the Data Warehouse*. Wiley-QED Publication.

Inmon, W. H., 1996. *Building the Data Warehous*e, John Wiley & Sons.

Inmon, W. H., 1997. *How to construct a Data Warehouse*. Rio de Janeiro. Campus

Inmon, W. H., 2000. *Building the Data Warehouse: Getting started* [online]. Available from: http://www.sergiomaturana.com/SAG/ttbuild.pdf [Accessed 2 March 2003].

Inmon, W. H., 2000a. *Data Warehouse and software development* [online]. Available from: http://www.billinmon.com/library/whiteprs/earlywp/ttswdev.pdf [Accessed 2 March 2003].

Inmon, W. H., 2003. Data Marts and Data Warehouse: Information Architecture for the Millenium [online]. Available from: http://www.billinmon.com/library/whiteprs/infx_dm.pdf [Accessed 14 April 2003].

Jarvenpaa, S. L., 1991. *Panning for gold in IS research: 'Second-hand' data*. Klein and Hirschheim (Eds.), *Information System Research: Contemporary approaches and emergent traditions*. Elseveir, Amsterdam, p.63-80.

Jermyn, P., Dixon, M. and Read, B., 2003. *Preparing clean views of data for data mining* [online]. Available from: http://www.ercim.org/publication/ws-proceedings/12th-edrg/edrg12_jedire.pdf [Accessed 15 March 2003].

Järvinen, P., 2001. *On research methods*. Finland. Tampere.

Kerkri, E., Quantin, C., Allaert, F., Cottin, Y., Charve, P., Jouanot, F. and Yétongnon, K., 2001. An approach for Integrating Heterogeneous Information Sources in a Medical Data Warehouse. *Journal of Medical Systems*, Vol. 25, No.3.

Kimball, R., 1997. *The Data Warehouse Toolkit*. Paris. International Thomson. France.

Kimball, R., Reeves, L, Ross, M., Thornthwaite, W., 1998. *The Data Warehouse Lifecycle Toolkit: Expert Methods for Design, developing and Deploying Data Warehouses*. New York. John Wiley & Sons Inc.

Kvale, S., 1983. The quantitative research interview: A phenomena and a hermeneutical mode of understanding. *Journal of Phenomenological Psychology* 14, No. 2, 171-196.

Lee, C., 2002. Data Warehouse and Data Warehousing [online]. Available from: http://www.1keydata.com/datawarehousing/datawarehouse.html [Accessed 5 April 2003].

Levine, E., 2002. Building a Data Warehouse: How you collect, manage, and report data may be the difference between success and failure [online]. *American School Board Journal*: November 2002 Technology Focus 2. Available from: http://www.paperbasket.com/html/samples/asbj11_02.PDF [Accessed 13 March 2003].

Linden, A., 1999. *Data Mining and Advanced Analytics: Extended Coverage*. Gartner Group: Research Note and Key Issues.

List, B., Schiefer, J. and Bruckner, R., 2002. A Holistic Approach for Managing Requirements of Data Warehouse Systems. In: 8th Americas Conference on Information Systems. Available from: http://www.ifs.tuwien.ac.at/~bruckner/pubs/amcis2002_managing_requirements.pdf [Accessed 15th March 2003].

Lofland, J. & Lofland, L.H., 1995. *Analyzing Social Settings: A Guide to Qualitative Observation and Analysis*. Belmont, California: Wadsworth Publishing Company.

Lungo, J., Fumo, T., Mukama, F., Macueve, G. and Tore, C. 2002. Analysis of the Health Information Flows in the Ministry of Health, Mozambique (Unpublished). Ministry of Health, Mozambique.

Macome, E., 1992. *Health Information Systems in Mozambique*. Thesis (Masters). London. London School of Economics.

Marshall, C.; Brown, A. & Fernandes, A., 1988. Análise do Sistema de Informação de Saúde e proposta de mudança. Quelimane: MISAU/DPS.

Maslow, A. A., 1954. *Motivation and Personality*. New York: Harper and Row.

Miles, M. & Huberman, A., 1994. *Qualitative Data Analysis: An Expanded Sourcebook*. 2nd edition, London: SAGE Publications Ltd.

Ministério da Administração Estatal., 2002a. *Information about Manjacazi*. [online]. Available from: http://www.mae.gov.mz/gaza/imandlak.htm [Accessed February 2003].

Ministério da Administração Estatal., 2002b. *Information about Chibuto* [online]. Available from: http://www.mae.gov.mz/gaza/ichibuto.htm [Accessed February 2003].

Ministério da Saúde., 1979. Cuidados de Saúde Primários em Moçambique. Maputo.

Ministério da Saúde., 2003. Relatório Final do Projecto Plano Director de Informatização do Sector de Saúde. MISAU/EuroSys. Moçambique.

Ministry of Health., 2001. Strategic Plan for the Health Sector (PESS): 2001-2005-(2010). Approved in the XI ordinary session of the Council of Ministers, 24th April 2001.

Moura, A. and Victorino, 2003. M. *Using Mediator and Data Warehouse Technologies for Developing an Environmental Decision Support System*. [online]. Military Institute of Engineering. Rio de Janeiro, Brazil, Computer Science Department. Available from: http://www.cos.ufrj.br/wiiw/papers/19-Ana_Moura1(11).pdf [Accessed 25 May 2003].

Muquingue, H., Kaasboll, J. and Berg, O., 2001. The status of health management and information management in the Mozambican health districts: Overview and preliminary findings. Norway.

Murlrow, C., Langhome, P. and Grimshaw, J., 1997. *Integrating Heterogeneous Pieces of Evidence in Systematic Reviews.* [online]. Annals of Internal Medicine. Academia and Clinic. Available from: http://www.acponline.org/journals/annals/01dec97/evidence.htm [Accessed 15 March 2003].

Noormahomed, A. R. and Segall, M., 1992. The Public Health Sector in Mozambique: A post-war strategy for rehabilitation and sustained development. (Portuguese original, 1992; English version printed by WHO in 1993).

Oster, R., 1998. Data Warehouse Development. [online]. Available from: http://www.rtantiques.com/dataware.htm [Accessed 17 March 2003].

Patton, M. Q., 1990. *Qualitative Evaluation and Research Methods*. 2nd edition. London: Sage Publications Ltd.

Pavignani, E. 2002. The Reconstruction Process of the Health Sector in Mozambique - *A messy affair with a happy end?*. [online]. Available from: http://www.who.int/disasters/hbp/case_studies/reconstruction.htm. [Accessed October 2002].

Piotti, B. and Sitói, A., 1999. Avaliação do sistema de informação para a saúde de Moçambique (SIS): Volume 1 Anexos ao relatório preliminary: Avaliação conjunta pelo Ministério da Saúde e Organização Mundial da Saúde. Maputo, MISAU/DPC.

Rasheed, M., 2003. Data warehousing in a nutshell. [online]. Available from: http://www.dawn.com/weekly/science/science1.htm [Accessed 15 March 2003].

Ryan, J., 1999. *Building and Developing an Enterprise Data Warehouse.* The Technology Guide Series.

Sakaguchi, T. and Frolik Mark, N., 1996. *A Review of the Data Warehousing Literature*. [online]. Available from: http://www.nku.edu/~sakaguch/dw-web.htm [Accessed February 2003].

Sakaguchi, T. and Frolick, M. N., 1997. A Review of the Data Warehousing Literature. *Journal of Data Warehousing* (2:1), p.34-54.

Sanchez, A., 1998. *Data Warehousing with Informix: Best Practices*. New Jersey, USA: Prentice Hall – Informix Press.

SAS, 2003. SAS Data Warehousing: A complete perspective for managing enterprise data [online]. Available from: http://www.sas.com/technologies/data_warehouse/47395_0102.pdf [Accessed 14 April 2003].

Shanks, G. and Corbitt, B. 1999. Understanding Data Quality: Social and Cultural Aspects ACIS 1999, 789-797.

SILICO, 2001. *Data Integration in Biopharmaceutical and Development Process.* SILICO Insight paper.

Silver, M. S., 1991. Decisional guidance for computer-based decision support. *MIS Quarterly* (15:1). p.105-122.

Silverman, D., 2000. *Doing qualitative research: A pratical handbook.* London. SAGE Publications.

Sommerville, I. and Kotonya, G., 1998. *Requirements Engineering: Processes and Techniques*, Horizon Pubs & Distributors Inc.

Sommerville, I., 2001. *Software Engineering*. Sixth Edition. USA. Addison–Wesley.

Staccini, P., Joubert, M., Fieschi, M. and Fieschi, D., 1998. Towards semantic integration within an existing medical information system. *In* B. Cesnik et al., eds. NFO 98. pp. 935-939.

Steensboe, C., 2002. Information Integration in the Global Enterprise. *IDC Papers*.
Van den Hoven, J., 1998. Data warehousing: Bringing it all together. *Information Systems Management* (15:2), p.71-73.

The British Medical Association (BMA), 1999. Memorandum (ACI 92) [online]. Proceedings of the Health Committee. Available from: http://www.parliament.the-stationery-office.co.uk/pa/cm199899/cmselect/cmhealth/549/549ap15.htm [Accessed 14 April 2003].

Vanhanen, J., Risku, K. and Kilponen, P., 2000. Combining Data from Existing Company Data Sources: Architecture and Experiences [online]. Proceedings of the 33rd Hawaii International Conference on System Sciences – 2000. Helsinki University of Technology, Finland. Available from: www.computer.org/proceedings/hicss/0493/04937/04937010.pdf [Accessed 14 April 2003].

Vassiliadis, P., 2000. Data Warehouse Modeling and Quality Issues. National Technical. Greece. University of Athens.

Vavouras, A., 2000. Data Warehouse Refreshment using SIRIUS. *Proceedings Paper*.
Walt, G. and Melamed, A., 1983. *Mozambique Towards a people's health service.* London: Zed Books

Wang, H. and Weigend, A., 2003. Data Mining for Financial Decision Making [online]. Special Issue on Journal of Decision Support Systems, *Guest Editors:* Wang, H. and Weigend, A. Available from: http://www.weigend.com/dss.html [Accessed 14 April 2003].

White, C., 1996. The Data Mart: A New Approach to Data Warehousing [online]. DataBase Associates International. Available from: http://www.databaseassociates.com/pdf/ibi.pdf [Accessed 5 April 2003].

White, C. J., 2000. *An Analysis-Led Approach to Data Warehouse Design and Development*. DataBase Associates.

Wixom, B. H., Gray, P. and Watson, H. J., 2000. Introduction to the Minitrack on Data Warehousing. In: *Proceedings of the Thirty-Third Annual Hawaii International Conference on Systems Sciences (HICSS)*, p.174.

Annex A – *Questionnaire for the Masters Students, Brazil*

1. Data integration
2. What is a data warehouse?
3. Types of data warehouses?
4. Models of data warehouses: Dimensional and relational models
5. What is a data mart?
6. Data mart project
7. What is data mining?
8. What is metadata?
9. The data warehouse architecture
10. Online transaction process (OLTP) and Online analysis process (OLAP)
11. Project team for data warehouse implementation (Functions, roles, etc)
12. End-users involvement and implementation
13. Problems during data warehouse architecture design and implementation
14. The data warehouse prototyping
15. The data warehouse life cycle
16. The data warehouse costs
17. Business and Intelligence areas
18. Plan for the data warehouse project
19. Operational database systems (legacy systems)
20. Tools in data warehousing

Annex B – *List of Contacts in Magude District*

1. Mr. Agapito Nhamtumbo, Teacher, FRELIMO first secretary
2. Mr. José, Barber.
3. Dr. Alberto Matambo, Clinic Manager in Magude Health Center
4. Mrs. Olímpia Augusto, Pharmacy Agent
5. Ms. Sandra, Hairdresser
6. Mr. Chaúque, Telecommunication of Mozambique (TDM) employee
7. Mr. João Carlos, District Department of Agriculture and Development
8. Mrs. Milagrosa Francisco, FRELIMO
9. Mr. Lopes, Former Manager of District Department of Agriculture and Development
10. Mr. Alberto, Teacher
11. Mr. André, Priest (IURD – Universal Church of God Kingdom)
12. Mrs. Leia Nhancale, Magude District Directorate of Health Manager

Annex C – *Interviews with Health Staff*

Place	Name	Place, Data and Occupation
MoH (*5*) Observations – More than 30 hrs	Mr. Francisco Macuácuá	MoH – 2/05/2002
	Dr. Bruno Piotti	MoH – 30/04/2002 MoH – 28/01/2002
	Dr. Burt Bruins	MoH – 30/04/2002
	Dr. Capeone	-
	Dr. Amélia Ganâncio	-
Manjacaze (*6*) Observations **–** 3 weeks TRAINING	Mr. Simão Cau	Manjacaze Rural Hospital – in charge of NHI – 05/05/2002
	Mr. Dinis	In charge of human resources – Manjacaze Rural Hospital - 05/05/2002
	Mr. Nhande	Manjacaze Rural Hospital – Acceptance Sector – 05/05/2002
	Mr. David Nhantumbo	Medicine agent
	Mr. Jusumba	Manjacaze Rural Hospital – Nurse
	Mr. Felizardo	Tuberculosis program – NHI
Chibuto (*6*) Observations – 3 weeks TRAINING	Mrs. Dalila	Manager of Chibuto Rural Hospital
	Mrs. Elisabete	Chibuto Rural Hospital – Administrator
	Nurse	In charge of EPI program at NHI
	Nurse	Working at NHI
	-	Administrative
	Nurse	NHI Nurse
PDH Gaza (*4*) Observations – 3 weeks TRAINING **-Chicumbane**	Mrs. Beatriz	PDH Gaza – NHI
	Dr. Danilo Sakur	Manager of PDH in Gaza
	Nurse	In charge of NHI at PDH
	Nurse	WEB (PDH)
Magude (*3*) Observations – 1 week	Dr. Alberto Matambo	Magude Health Centre Clinic Director – 20/03/2002
	Mrs. Olímpia Augusto	Pharmacy Agent – 20/03/2002 – Magude
	Mrs. Leia Nhancale	Manager of DDH – 22/02/2002 – Magude
Total Interviewed (*24*)		

Annex D – *Interviews Questions*

D.1 Questions for Magude Field Work

1. **Divisão Administrativa**
   - Área
   - Localização
   - Limites (Norte, Sul, Este e Oeste)
   - Postos administrativos
2. **Geografia do distrito**
   - Planaltos, vales e rios
3. **População e Infra-estruturas**
   - População
   - Serviços disponíveis
   - Principais culturas
   - Sistema educacional
   - Perfil sócio-económico
   - Sistema de transporte
   - Vias de acesso
   - Sistemas de comunicação (Telefone, rádios comunitárias, etc.)
   - Factos históricos
   - Ambiente político
4. **Perfil da saúde no distrito**
   - Serviços das unidades sanitárias
5. **Infra-estruturas e pessoal de saúde**
   - Número de unidades hospitalares no distrito e sua distribuição
   - Número de pessoal (medicos, técnicos, enfermeiros, agents, serventes e outros) distribuídos pelas diferentes unidades sanitárias
   - Serviços de saúde existentes

- Principais problemas (Transporte, medicamentos, alimentação, pessoal, ambulância, edifícios)

6. **Saúde distrital**
   - Principais doenças
   - Maior causa de mortes
   - Problemas crónicos
7. **Programas chave**
   - DTS/HIV
   - Tuberculose
   - Saúde materno infantil
   - Doenças crónicas
8. **Fluxo de informação sobre saúde**
   - Tratamento dos dados
   - Forma em que os dados existem
   - Análise da informação
   - Meios para o armazenamento de dados
   - Meios usados na produção da informação
   - Relatórios? Seminários?
   - Quem produz a informação? Periodicidade?
   - Qualidade da informação produzida
   - Acesso a informação produzida
   - Retroalimentação
   - Feedback da informação
   - Contribuição da população
   - Maiores problemas no envio de informação para o nível seguinte
   - Informação do sector tradicional? privado? ONG's?

D.2 Questions for the Ministry of Health

1. **Organization data**
    - Structure of Ministry of Health
    - Computational resources
    - Operational legacy systems
    - Decision level
    - Information flow among different levels
    - Existing systems: Name, platform, database, language, outputs
    - Data standards
2. **Data acquisition**
    - Who acquires data?
    - Basis for data acquisition
    - Types of data
    - Who collects data?
    - Tools used to acquire data
    - Type of information that managers needs
    - Data models
3. **Data storage**
    - How data is stored
    - Which form
    - Use of computer? Other hardware?
    - Data volume
    - Data stored are divided into subjects?
    - Data stored are summarized or detailed?
    - Data stored are used to operate or to analyze the business
4. **Data analysis**
    - What kind of information can be produced?
    - What kind of analysis can be achieved?
    - Tools used in the analysis

- Who can do analysis?

5. **Other**

- Need any other type of analysis?
- Other data that would like to have
- Other systems that would like to have access to

D.3 Questions for the Gaza Field Work

1. **Análise situacional**

   a) Recolha de dados

   - Ferramentas usadas
   - Pacote mínimo de dados
   - Práticas na recolha de dados
   - Elaboração dos relatórios

   b) Processamento de dados

   - Qualidade e validação
   - Procedimentos mensais do fluxo de informação
   - Usos de diferentes ferramentas
   - Tipos de relatórios

   c) *Análise*

   - Ferramentas usadas para análise
   - Indicadores em uso
   - Mecanismos de retroalimentação
   - Uso dos princípios epidemiológicos
   - Envolvimento dos gestores

   d) *Disseminação*

   - Tipos de tabelas construídas
   - Tipos de gráficos construídos
   - Tipo de retroalimentação? Conteúdo?
   - Responsáveis pela retroalimentação

e) *Uso*

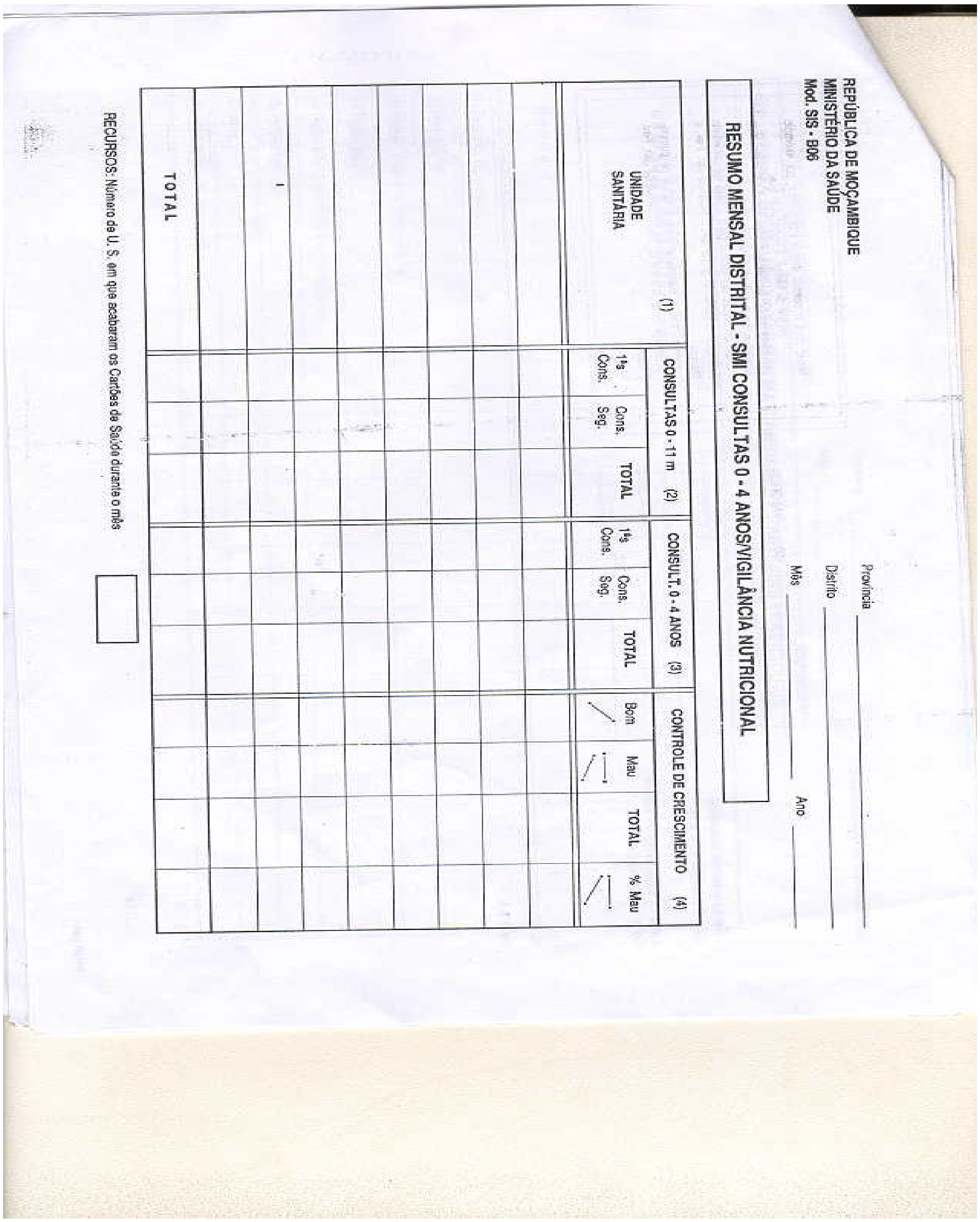
- Evidências do uso de informação na tomada de decisões
- Uso de informação nas reuniões
- Planificação usa informação? Evidências
- Uso de informação nas supervisões

2. Base de Dados Sistema de Informação de Saúde Distrital

- Completar a base de dados
- Avaliar a qualidade dos dados
- Analisar dados relevantes para o distrito
- Análise do ciclo vital
- Construção de gráficos

Annex E – *Sample Forms used in Sistema de Informação de Saúde*

Form B06



REPÚBLICA DE MOÇAMBIQUE
MINISTÉRIO DA SAÚDE
Mod. SIS - B06

Província ____
Distrito ____
Mês ____ Ano ____

RESUMO MENSAL DISTRITAL - SMI CONSULTAS 0 - 4 ANOS/VIGILÂNCIA NUTRICIONAL

UNIDADE SANITÁRIA (1)	CONSULTAS 0 - 11 m (2)			CONSULT. 0 - 4 ANOS (3)			CONTROLE DE CRESCIMENTO (4)			
	1ªs Cons.	Cons. Seg.	TOTAL	1ªs Cons.	Cons. Seg.	TOTAL	Bom	Mau	TOTAL	% Mau
TOTAL										

RECURSOS: Número de U. S. em que acabaram os Cartões de Saúde durante o mês

Form B08

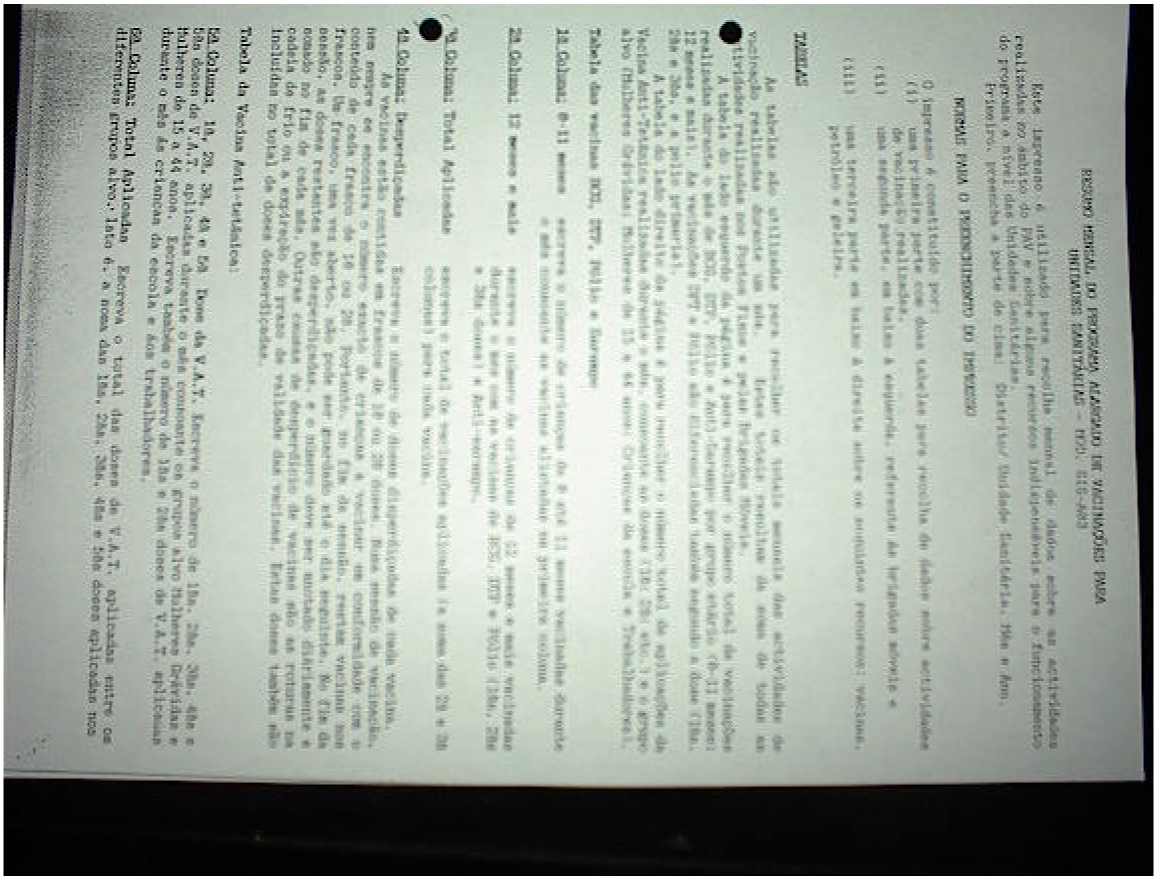
REPÚBLICA DE MOÇAMBIQUE
MINISTÉRIO DA SAÚDE
Mod. SIB - B08

Província ____
Distrito ____
Mês ____ Ano ____

RESUMO MENSAL DISTRITAL - SMI CONSULTAS PRÉ-NATAIS/POST-PARTO/PF

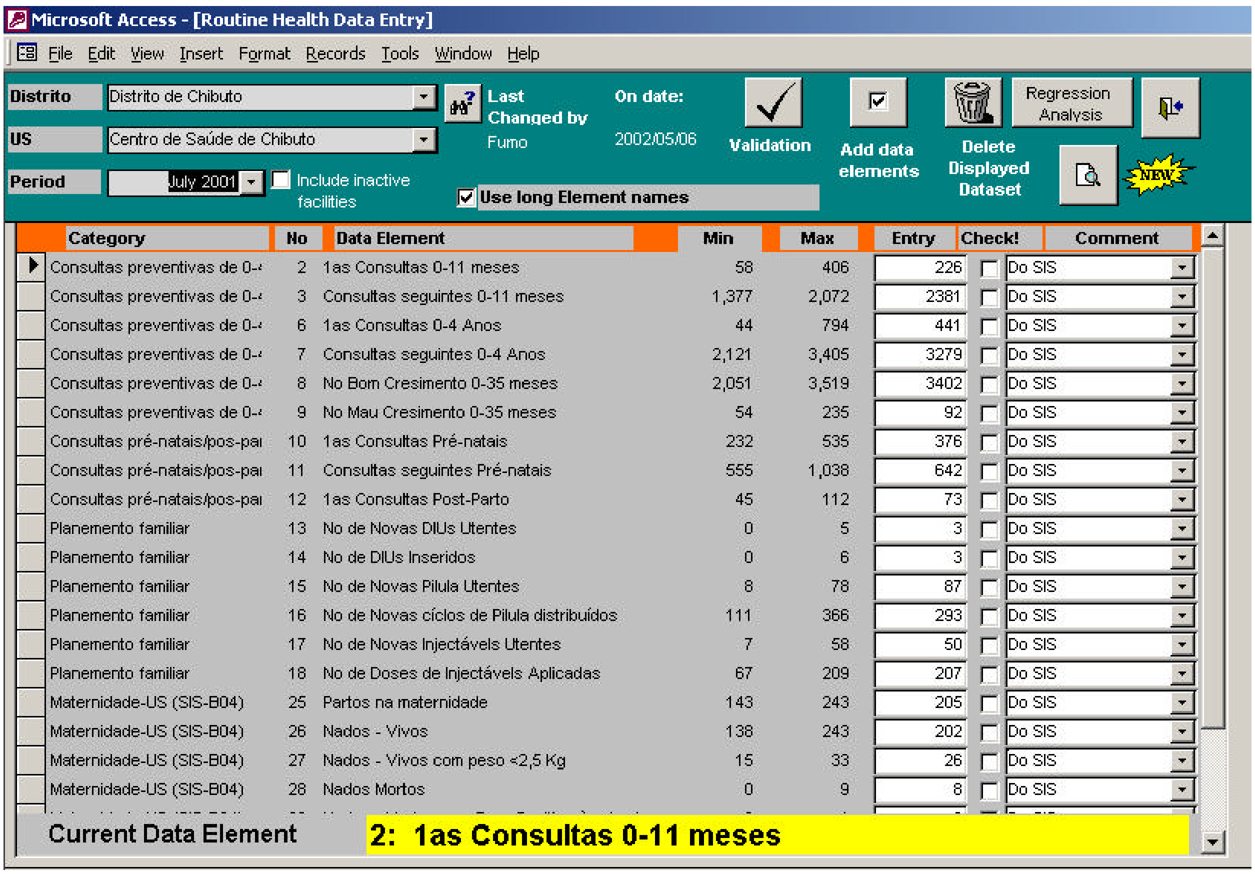
UNIDADE SANITÁRIA (1)	CONS. PRÉ-NATAIS (2)			1ª CONS. POST-PARTO (3)	PLANEAMENTO FAMILIAR					
					DIU		PÍLULA		INJECTÁVEL	
	1ª Cons.	Cons. Segs	TOTAL		Nº de Novas Utentes (4)	Nº de DIUs Inseridos (5)	Nº de Novas Utentes (6)	Nº Ciclos Distribuídos (7)	Nº de Novas Utentes (8)	Nº Doses Aplicadas (9)

Annex F – *Procedures in the Forms*



Annex G – *Activity and Semi-Permanent Data*

G.1 Activity Data from District Health Information System (DHIS)



Microsoft Access - [Routine Health Data Entry]

File Edit View Insert Format Records Tools Window Help

Distrito: Distrito de Chibuto

US: Centro de Saúde de Chibuto

Period: July 2001

Include inactive facilities

Last Changed by: Fumo

On date: 2002/05/06

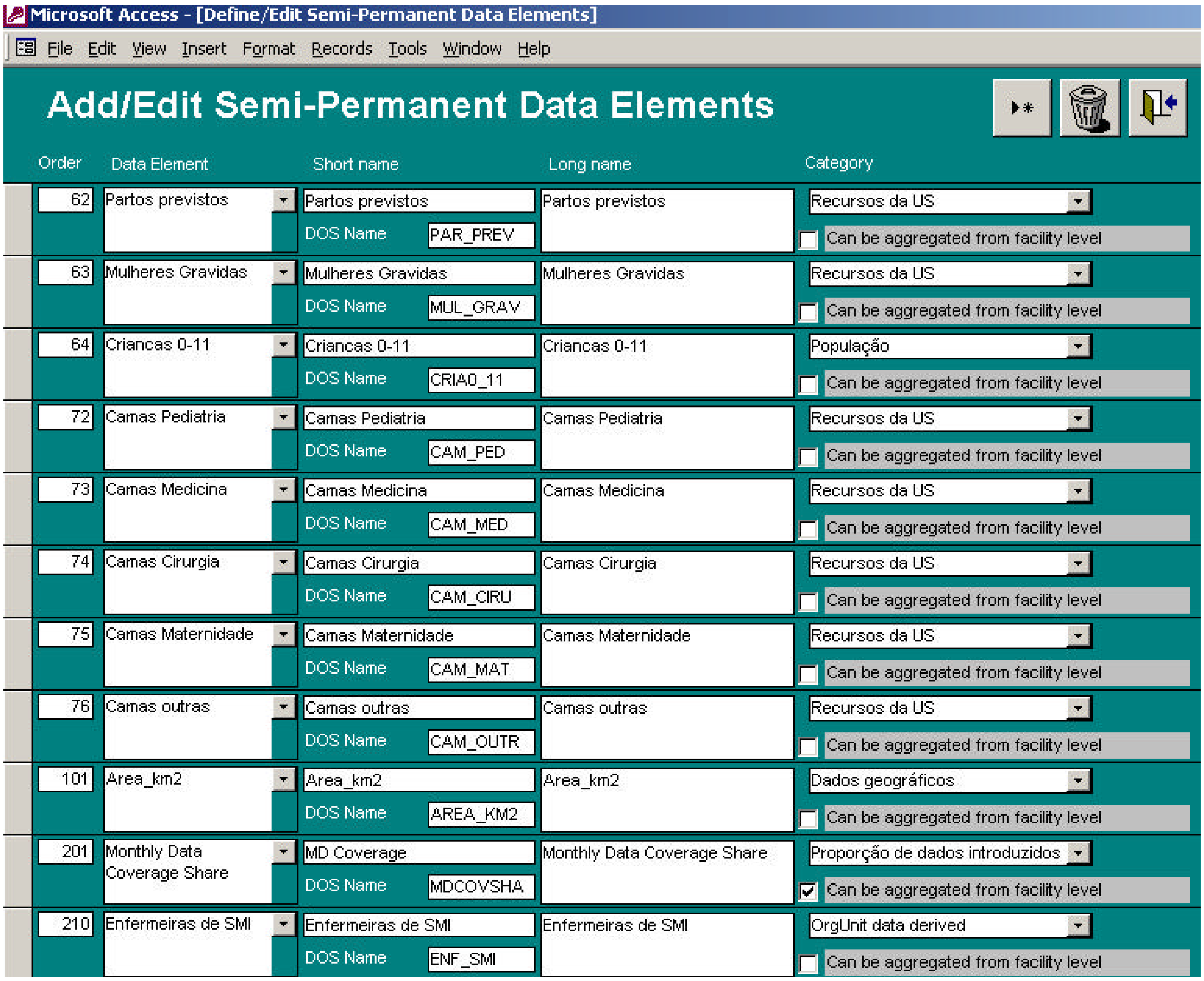
Validation | Add data elements | Delete Displayed Dataset | Regression Analysis | NEW

Use long Element names

Category	No	Data Element	Min	Max	Entry	Check!	Comment
Consultas preventivas de 0-4	2	1as Consultas 0-11 meses	58	406	226		Do SIS
Consultas preventivas de 0-4	3	Consultas seguintes 0-11 meses	1,377	2,072	2381		Do SIS
Consultas preventivas de 0-4	6	1as Consultas 0-4 Anos	44	794	441		Do SIS
Consultas preventivas de 0-4	7	Consultas seguintes 0-4 Anos	2,121	3,405	3279		Do SIS
Consultas preventivas de 0-4	8	No Bom Cresimento 0-35 meses	2,051	3,519	3402		Do SIS
Consultas preventivas de 0-4	9	No Mau Cresimento 0-35 meses	54	235	92		Do SIS
Consultas pré-natais/pos-par	10	1as Consultas Pré-natais	232	535	376		Do SIS
Consultas pré-natais/pos-par	11	Consultas seguintes Pré-natais	555	1,038	642		Do SIS
Consultas pré-natais/pos-par	12	1as Consultas Post-Parto	45	112	73		Do SIS
Planemento familiar	13	No de Novas DIUs Utentes	0	5	3		Do SIS
Planemento familiar	14	No de DIUs Inseridos	0	6	3		Do SIS
Planemento familiar	15	No de Novas Pilula Utentes	8	78	87		Do SIS
Planemento familiar	16	No de Novas cíclos de Pilula distribuídos	111	366	293		Do SIS
Planemento familiar	17	No de Novas Injectávels Utentes	7	58	50		Do SIS
Planemento familiar	18	No de Doses de Injectávels Aplicadas	67	209	207		Do SIS
Maternidade-US (SIS-B04)	25	Partos na maternidade	143	243	205		Do SIS
Maternidade-US (SIS-B04)	26	Nados - Vivos	138	243	202		Do SIS
Maternidade-US (SIS-B04)	27	Nados - Vivos com peso <2,5 Kg	15	33	26		Do SIS
Maternidade-US (SIS-B04)	28	Nados Mortos	0	9	8		Do SIS

Current Data Element: 2: 1as Consultas 0-11 meses

G.2 Semi-Permanent Data from District Health Information System (DHIS)


Microsoft Access - [Define/Edit Semi-Permanent Data Elements]

File Edit View Insert Format Records Tools Window Help

Add/Edit Semi-Permanent Data Elements

Order	Data Element	Short name	DOS Name	Long name	Category	Can be aggregated from facility level
62	Partos previstos	Partos previstos	PAR_PREV	Partos previstos	Recursos da US	☐
63	Mulheres Gravidas	Mulheres Gravidas	MUL_GRAV	Mulheres Gravidas	Recursos da US	☐
64	Criancas 0-11	Criancas 0-11	CRIA0_11	Criancas 0-11	População	☐
72	Camas Pediatria	Camas Pediatria	CAM_PED	Camas Pediatria	Recursos da US	☐
73	Camas Medicina	Camas Medicina	CAM_MED	Camas Medicina	Recursos da US	☐
74	Camas Cirurgia	Camas Cirurgia	CAM_CIRU	Camas Cirurgia	Recursos da US	☐
75	Camas Maternidade	Camas Maternidade	CAM_MAT	Camas Maternidade	Recursos da US	☐
76	Camas outras	Camas outras	CAM_OUTR	Camas outras	Recursos da US	☐
101	Area_km2	Area_km2	AREA_KM2	Area_km2	Dados geográficos	☐
201	Monthly Data Coverage Share	MD Coverage	MDCOVSHA	Monthly Data Coverage Share	Proporção de dados introduzidos	☑
210	Enfermeiras de SMI	Enfermeiras de SMI	ENF_SMI	Enfermeiras de SMI	OrgUnit data derived	☐

Annex H – *List of some of the Forms used in the Sistema de Informação de Saúde*

Item	Code	Designation	Fulfilling Level	Periodicity
Registers Books				
1	SIS-C01	Livro de registos de consultas externas	Health facility	Daily
2	SIS-D01	Livro de registo de internamentos	Health facility	Daily
3	SIS-B01	Livros de registos de maternidade	Health facility	Daily
4	SIS-E01	Livro de registos do laboratório	Health facility	Daily
…	…	…	…	…
Printed Forms				
5	SIS-A01	Registo diário do PAV (BCG, DPT)	Health facility	Daily
6	SIS-A02	Registo diário do PAV (VAT)	Health facility	Daily
7	SIS-A03	Resumo mensal do PAV para unidades sanitárias	Health facility	Monthly
8	SIS-A04	Resumo mensal do PAV para os distritos	District	Monthly
9	SIS-A05	Resumo annual distrital de recursos – PAV	District, Province	Annual
10	SIS-B02	Regsito diário de consultas pré-natais/pós-parto	Health facility	Daily
11	SIS-B03	Registo de consultas de crianças 0-4 anos	Health facility	Daily
12	SIS-B04	Resumo mensal de SMI/vigilância nutricional	Health facility	Monthly
13	SIS-B05	Controle de stocks de métodos anti-conceptivos	Health facility	Monthly
14	SIS-B06	Resumo mensal distrital (SMI, consultas, vigilância)	District	Monthly
15	SIS-B07	Resumo mensal distrital das maternidades e parteiras	District	Monthly
16	SIS-B08	Resumo mensal distrital (SMI/consultas pré-natais/pós-parto e planeamento familiar)	District	Monthly
17	SIS-C02	Ficha de contagem de novos casos de doenças de declaração obrigatória (Semanal)	Health facility	Weekly
18	SIS-C03	Boletim epidemiológico seminal – Unidades sanitárias e distritos	Health facility, district, province	Weekly
19	SIS-D03	Resumo mensal de internamentos para centros e postos de saúde	Health facility	Monthly
20	SIS-D04	Resumo de internamentos para hospitais distritais , rurais e gerais	Rural hospital /general, province	Monthly, semester, Annual
21	SIS-J02	Resumo anual da rede sanitaria e camas	Province	Annual
22	SIS-K01	Resumo anual distrital de recursos – pessoal de nível primário e secundário	District, Province	Annual
23	SIS-K02	Resumo anual provincial de recursos – pessoal de nível primário (rural and urban), secondary, provincial directorate and training institutions	Province	Annual
…	…	…	…	…

Annex I –*Description of the Software Application used to Extract Data from Computerized Sistema de Informação de Saúde*

RESTRIEVING DATA FROM OLD HEALTH INFORMATION DATABASES: A CASE OF MOZAMBIQUE HEALTH INFORMATION SYSTEM.

By Juma Lungo, May 2002

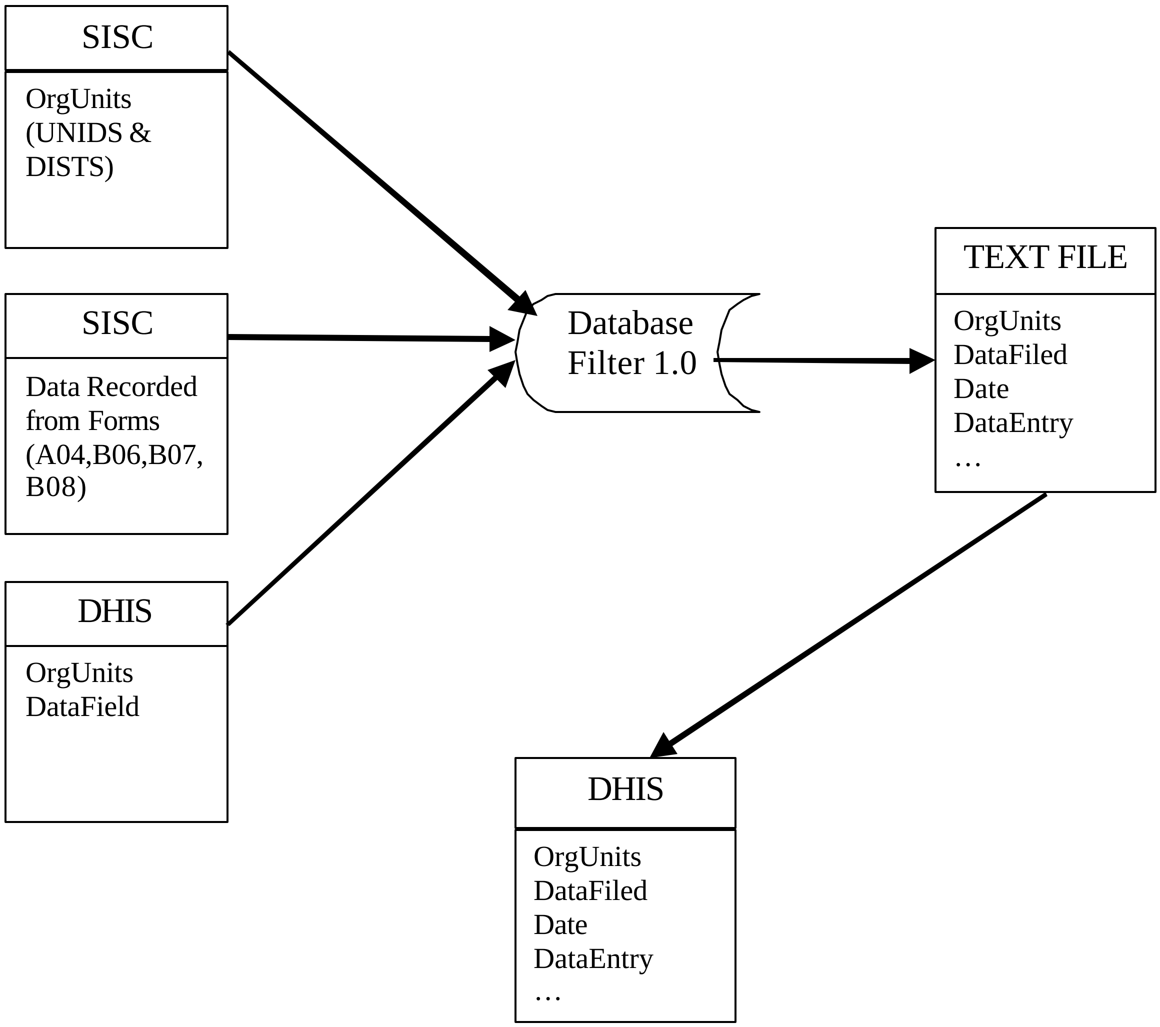
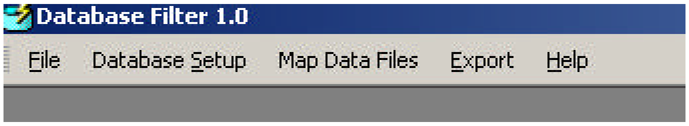


Figure 1: Model of the Logic on how the program works. SISC is the MISAU database, DHIS is the new database and *Database Filter 1.0* is the application that extracts data.

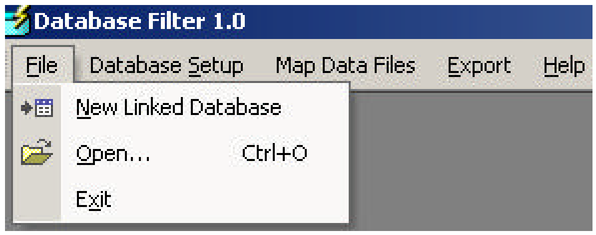
IMPLEMETATION OF THE SOFTWARE (Database Filter Application)

The application was designed and implemented after the analysis of the problem. It uses linked tables to connect to SIS dBase database and DHIS Microsoft Access database. The following figure is the menu list of the application developed.



File Menu

This menu has three commands: New Linked Database, Open and exist.

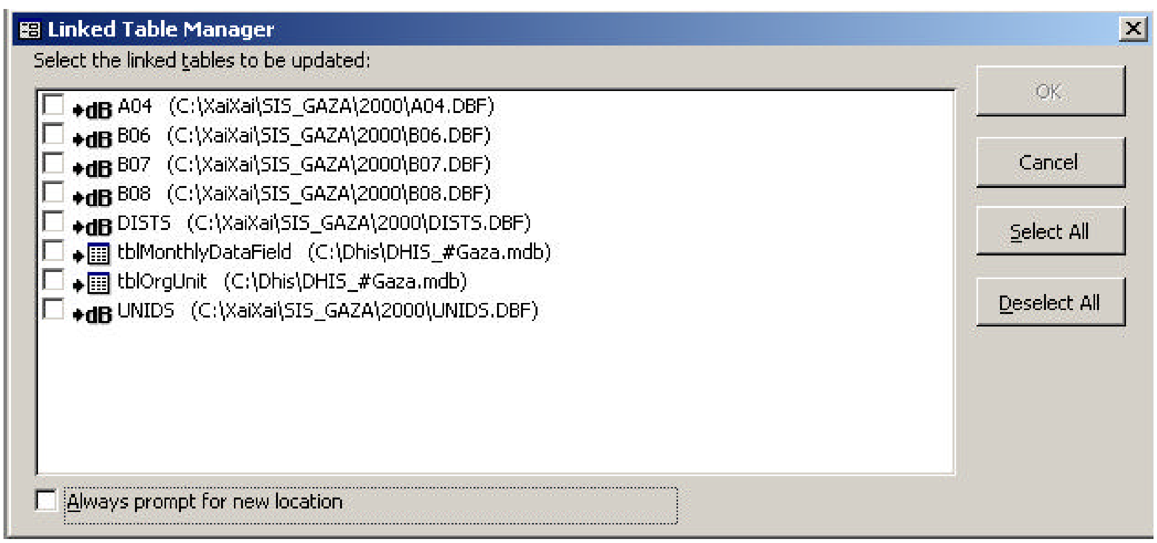


New Linked Database: Allows you to link more tables. For the time being, the application only support table A04, B06, B07, B08, DISTS, and UNIDS of SIS Database and tblOrgUnit and tblMonthlyDataField of the DHIS.

The **Open** command opens Access Database and the **Exit** commands ends the session of your Database Filter application.

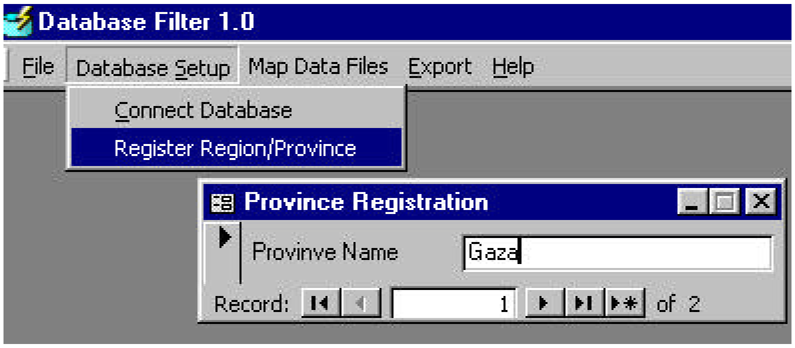
Database Setup Menu

This menu has only one command – **Connect Database**



Connect Database command gives a possibility of connecting to different databases. On clicking this command the following screen will appear. If the old Database (SIS) is not located in the path shown (C:\XaiXai\SIS_GAZA\2000\), select the check boxes of the dBase tables and check the “Always prompt for new location” check box. Click OK and follow the instructions to locate your Legacy Database.

Register Region/Province



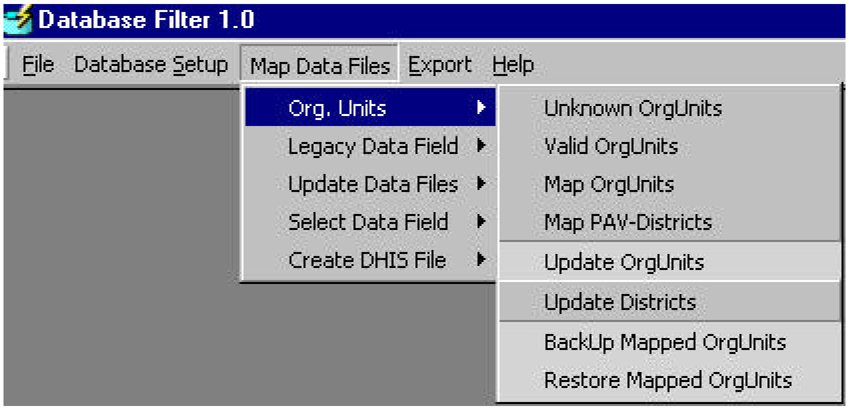
The small screen "Province Registration" is used to register provinces. This was purposely designed in order to save and retrieve mapping of various provinces.

Map Data Files Menu

This menu consists of five commands.

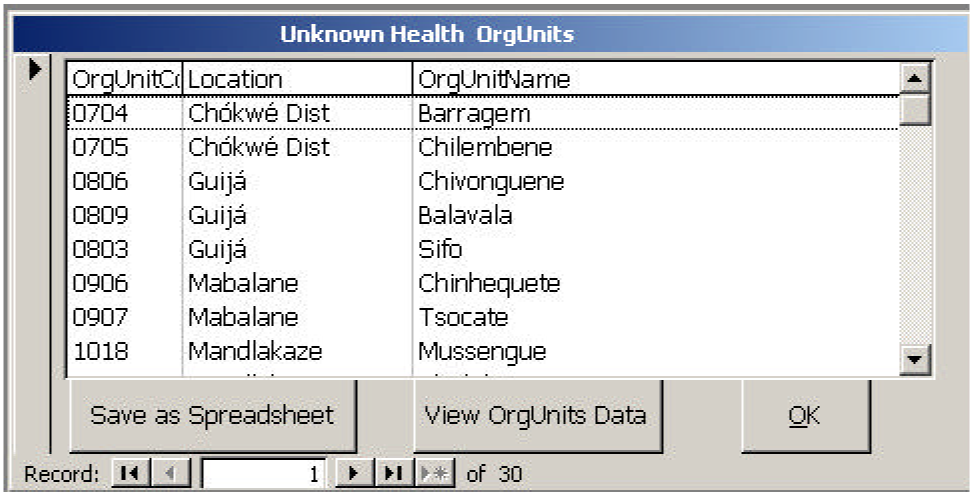
1. **Org. Units**

   It has sub commands to manipulate Organization Units Mapping.

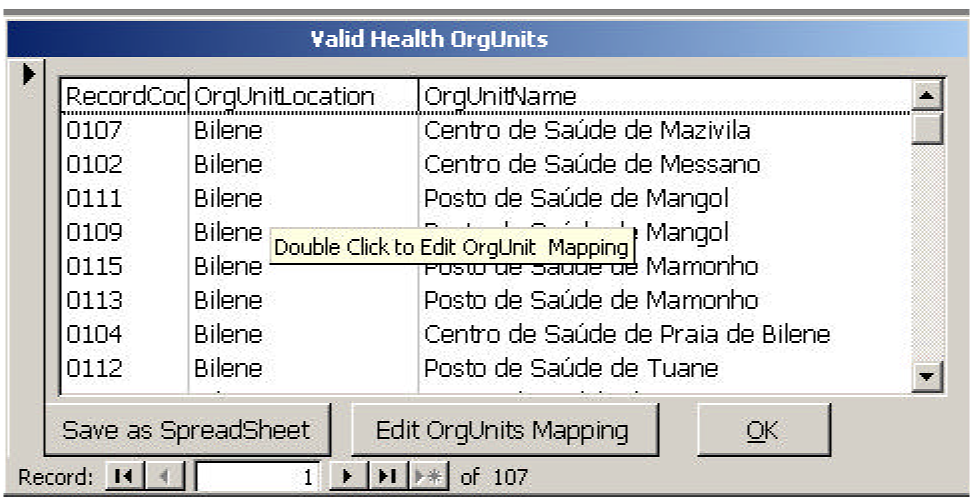


Unknown OrgUnits: Displays the names of Organization Units that exist in the Legacy Database (SIS) but has no corresponding OrgUnits in DHIS Database. By unknown health units, we refer to those un-mapped health units to the existing health units in DHIS software. Even if the health unit has a correspondent health unit name in DHIS database, if was not mapped, it will appear in the Unknown OrgUnits List.

Added Values: *with the commands on the screen showing the unknown list, you can save the list in Excel file format, you can view/check if these ghost units contain data and by* ***double-clicking a health unit name****,* ***the mapping window appear where you can map the health unit.*** *The "****OK****" command closes the screen.*



Valid OrgUnits: Displays the names of Organization Units exist in both the old Database and DHIS Database. All mapped health units will appear in this list.



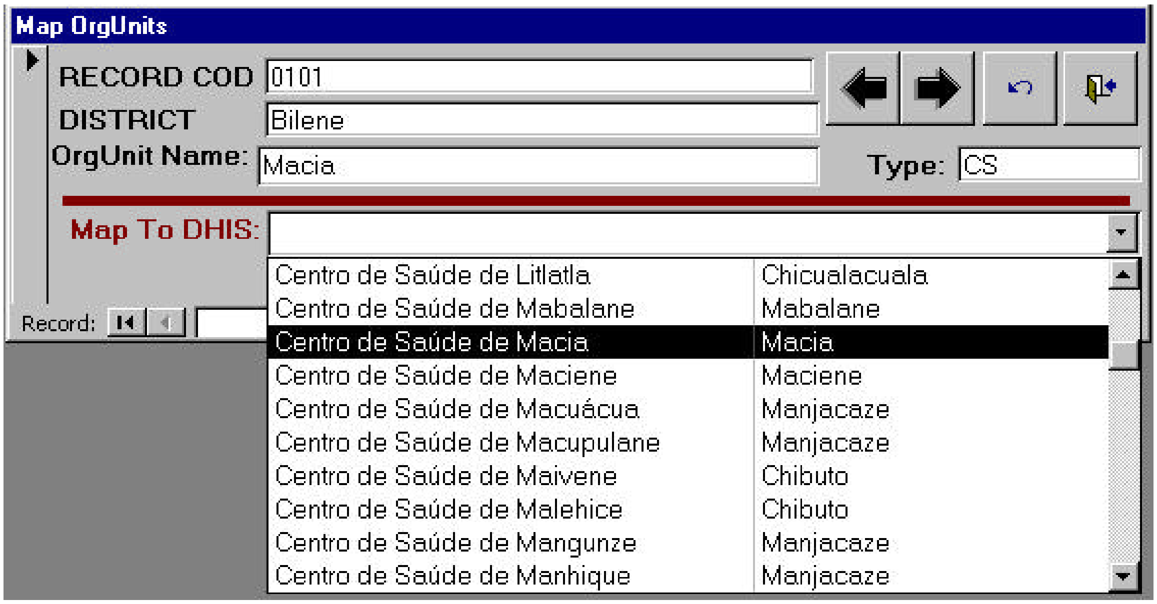
Added Values: *with the commands on the screen showing the unknown list, you can save the list in Excel file format. On clicking "**Edit Orgnits Mapping"** command or **double click a health unit name**, **the mapping window appear where you can map the health unit.** The "**OK**" command closes the screen.*

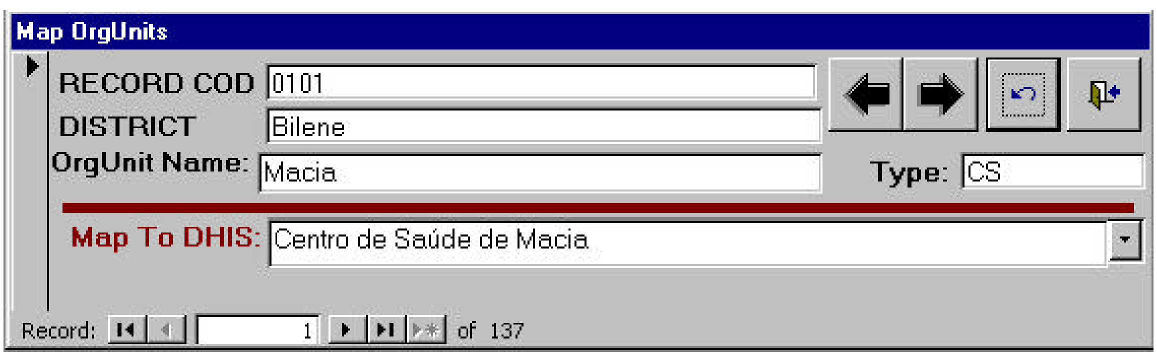
Map OrgUnits: Displays a screen for mapping the organization Units.

In the SIS Database, an orgUnit is identified by three data fields: ***District, OrgUnit Name*** and ***OrgUnit Type***. Below **the red line** is a combo box which lookup the DHIS

OrgUnits registered. Use the navigation arrows to navigate from one record to another. To map an orgUnit, click the combo box and select on the list the corresponding orgUnit name.

Added Values: *the commands on the screen are Previous Record (the back arrow), Next record (forward arrow), undo command and close form command. On mouse over the button, the name of the button will be displayed in context text. To select a matching record (health in unit) in DHIS click the record. The" event on click" will save your mapping.*





Mapped record.

Update OrgUnits: This commands copies all SIS OrgUnits into a Temporary table where then can be accessed and mapped to DHIS orgUnits. Use this command only once when you want to map orgUnits of new Province. For example, now we have mapped all GAZA orgUnits, if we want to mapp orgUnits from Niassa, we have to link the database of Niassa and then use this command to save the records in a temporary file for mapping purposes.

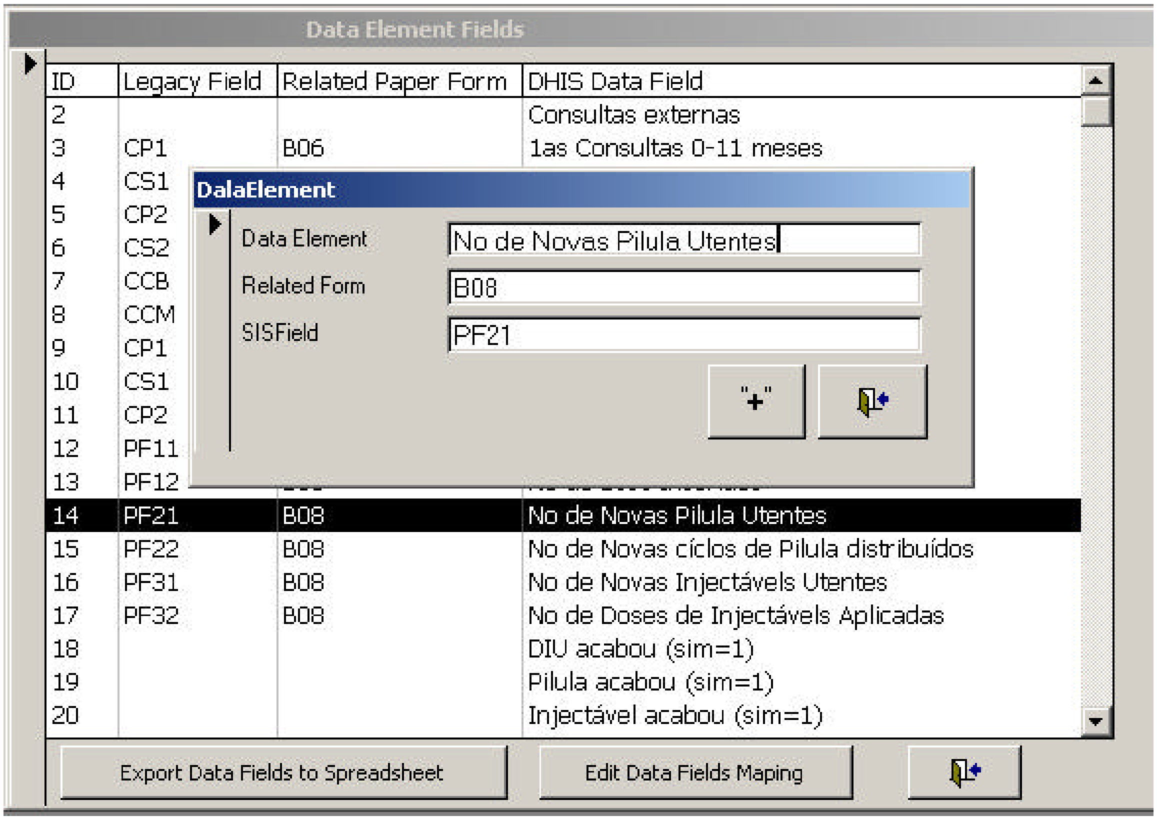
Backup Mapped OrgUnits: Save a Backup copy of the mapped orgUnits

Select the Province which you are working with to save the mapping.

Restore Mapped OrgUnits: this restores all Mapped OrgUnits from the backup file. So if it happens someone runs the Update OrgUnit commands, all mapping links will get lost! To restore the mapping links, use this command.

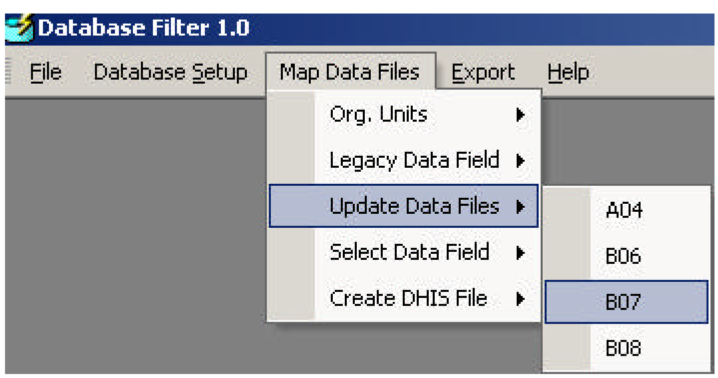
2. **Legacy Data Field**
   This command has only one sub command – **View Data Field ID**
   The form displayed is for mapping the data fields from SIS database to DHIS data element fields. The mapping is used to determine the correct data element IDs of the DHIS software when creating DHIS import Data File.

   To add new data field mapping, double click the data element to and a popup window will appear. Enter the correct entries of every field and then click the plus"**+**" Button. The "**Export Data Field to Spreadsheet**" saves the list in Excel file.



***Added Values:** with the commands on the screen showing the mapped and unmapped data fields, you can save the list in Excel file format. On clicking "**Edit Data Fields mapping"** command or **double click a row**, **the mapping window appears where you can map the data fields.** The "**close**" command closes the screen.*

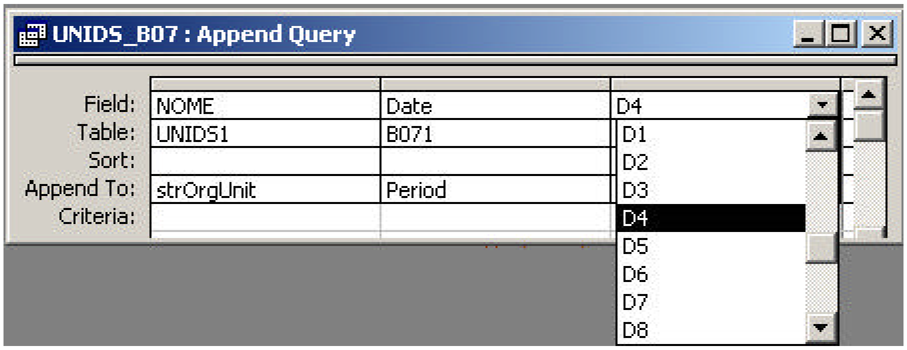
Update Data Files: the command has four subcommands: **A04, B06, B07** and **B08.** Use this command to update the data after connecting to new database. E.g. if you were connected to SIS dBase 1999, when you finish you connect the dBase 2000. In this case you must update all temporary files.



3. **Select Data Field:** the command displays the list of all data fields of the Database Table in the legacy database (Sistema de Informação de Saúde). Select one field to

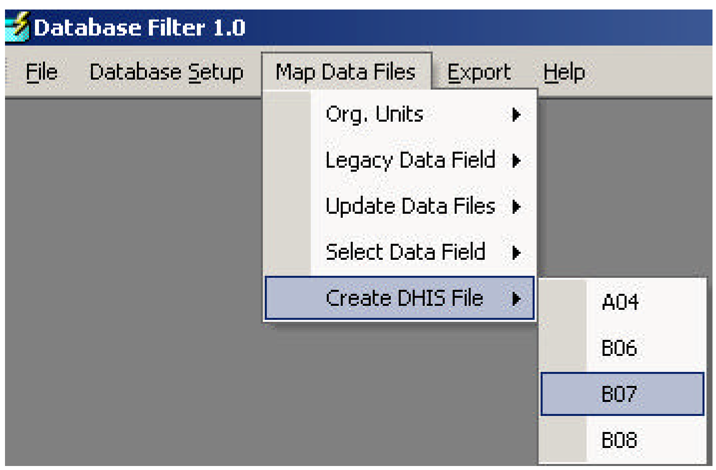
extract data at a time. The example below is from FORM B07. If one wants to extract data from B07 then he/she must select one data field to extract its data item. Select a data field from the dblEntry list.

Warning: Do not disturb the strOrgUnit and Period column!



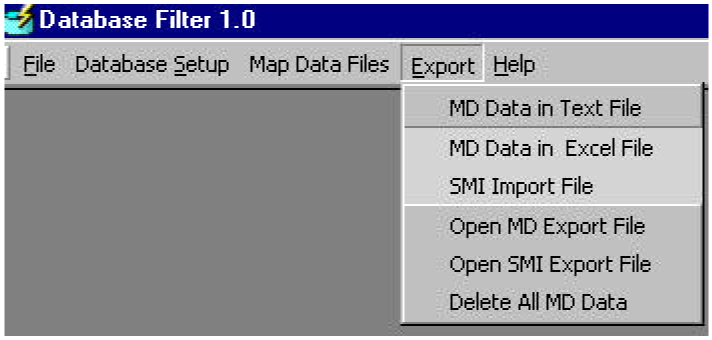
The same applies for all data files (A04, B06, B07, and B08).

4. **Create DHIS File**: Use this command to create the DHIS Data file from any one from the four data files. To use this command you must have chosen the data field as instructed in the **Select Data Field** above.



Export Menu

The Export Menu has commands that allow you to **view (open)** the created DHIS Datafile, Exporting the DHIS datafile in **Spreadshee** and in **Text** file.

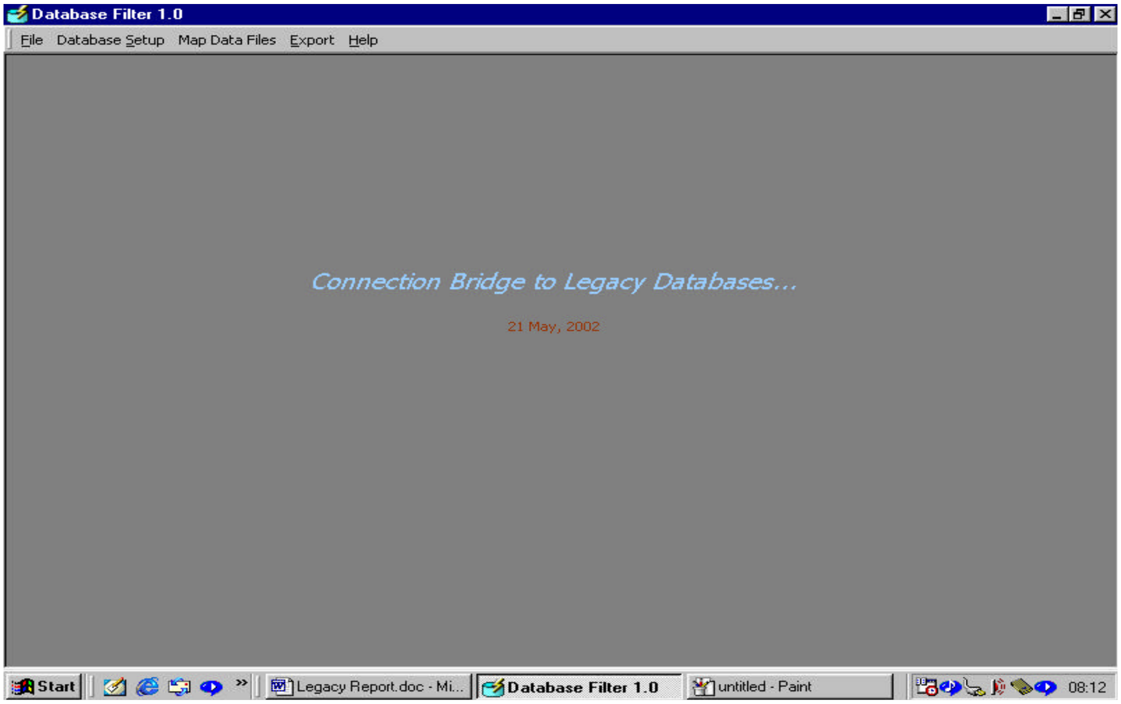


For the time being, only the MD (Monthly Data) commands are working. Thus you can use the command "**MD Data in Text File**", "**MD Data in Excel File**" and "**Open MD Export File**". The SMI command is under construction!

Added Values: *The application, stores all the data in a temporary table. Before exporting to DHIS text file or Spreadsheet file, you can view the data. This is important because DHIS date format depends on the regional settings of your computer. If you set e.g. Norwegian, DHIS database will not accept your data because of the date format. Thus why after generating few records, you are strongly advised to check/view you data.*

Help Menu

Added Value: *The application database Filter displays the current date on the middle of your screen all the time. This was made purposely to remind you your current date format and hence to ensure that you set your computer in a setting that will produce date format that will be accepted by DHIS tool.*


Database Filter 1.0
File Database Setup Map Data Files Export Help
Connection Bridge to Legacy Databases...
21 May, 2002
Start
Legacy Report.doc - Mi...
Database Filter 1.0
untitled - Paint
08:12