Assessment for a Relational Database Management System (RDBMS) for the Ethiopian Central Statistical Agency (CSA)

Survey Information System Analysis: Feasibility study for the development and implementation of RDBMS

Final report

Funded by the European Commission
CRIS 120-352
GCP/ETH/071/EC

Birru Dori
National Database Programmer and DBA Expert
September, 2008
Addis Ababa, Ethiopia.

Table of contents

Acknowledgements

I would like to express my heartfelt appreciation and thank to the FAO Chief Technical Advisor, Mr. Raphy Favre for his assistance in facilitating all my activities. I also thank Mr. Thomas Gabrielle and Mr. Yakob Mudesir for their technical advices and guidance in undertaking this study.

Finally, I would like to thank CSA staff members; Alemayehu G/Tsadik, Eleni Kebede, Tabit Yasin , Amare Legese, Girma Tadesse, Mageru Haile, Alemayehu Gebre, Tekelab, Shimelis Mulugeta, and Biratu in their effective responses to consultations I have made with them towards producing a good and fruitful result of this study.

Birru Dori.

Definitions and acronyms

1. **Dataset**: is any organized collection of data or information that has a common theme. A dataset might be a list of objects, a digital map, records of geological borehole samples, a collection of photographs at a certain location or of a certain subject, etc. One may also see dataset as individual-level results of a survey, conceptualized as a table or "matrix" where the rows are individual respondents and the answers given by each respondent or values derived from those answers.
2. **Database**: is a structured collection of related records or data that is stored in a computer so that a program can consult it to answer queries. The records retrieved in answer to queries become information that can be used to make decisions. It can also be seen as a logical collection of interrelated information, managed and stored as a unit, usually on some form of mass-storage system such as magnetic tape or disk. A logical collection of interrelated information managed and stored as a unit, usually on some form of mass-storage system such as magnetic tape or disk.
3. **RDBMS**: Relational database management system (RDBMS) is a database management system with the ability to access data organized in tabular files that can be related to each other by a common field (item). An RDBMS has the capability to recombine the data items from different files, providing powerful tools for data usage. RDBMS is the most popular architecture for storing data in a database
4. **Data Dictionary:** A catalog of all data held in a database, or a list of items giving data names and structures. Also referred to as DD/D for data dictionary/directory. Commercial RDBMS have online data dictionaries stored in special tables called system tables.
5. **Standard**: Something established as a basis of comparison in measuring or judging capacity, quantity, content, value, quality, etc.; a specified set of safety or performance qualities which a device or process must possess. .... It can be considered as a basis for comparison; a reference point against which other things can be evaluated or seen.
6. **Flowchart** is a representation, primarily through the use of symbols, of the sequence of activities in a system (process, operation, function, or activity).
7. **SMS:** Subject Matter Specialist
8. **DP:** Data Processing
9. **DBA:** Data Base Administrator
10. **ICT:** Information communications Technologies
11. **Belg:** {in Amharic language } short rainy season in Ethiopia
12. **EA:** Enumerators Area
13. **Field editor:** Supervisor
14. **Free and Open Source Software (FOSS)** programs are programs whose licenses permit users the freedom to run the program for any

purpose, to study and modify the program, and freely redistribute copies of the original or modified program

15. **Free Software** is different from Freeware, Shareware, Adware, Spyware or Crippleware, which are all types of proprietary software made available at no price, providing various degrees of freedom of use, but in most cases not other freedoms as described by FSF.
16. **Open Source Software (OSS)** is software where the source code (the language in which the program is written) is freely distributed with the right to modify the code, and on the condition that redistribution is not restricted, and indeed is obtainable for no more than the reasonable cost of production.
17. **FSF** stands for Free Software foundation.
18. **Proprietary software** is software, which has been designed and coded by or for a specific person, organization or group of organizations, which hold ownership or intellectual property rights over the software. An individual or a company (usually the one that developed it) owns proprietary software.
19. **ISIS:** Integrated Survey Information System
20. **ISCI:** Industry Standard Commercial Identifier
21. **DDI:** Data Documentation Initiative

Part One: An overview of Survey Information System problems and solutions

1. Introduction

1.1 Rationales and Justifications

This feasibility study for the implementation of a Relational Database Management System (RDBMS) is undertaken under the framework of an EC funded "Support to Food Security Information System Project – GCP/ETH/071/EC". The project is a contribution to the overall efforts to improve the food security situation of the country. It aims at addressing a number of bottlenecks in the systems of information generation and analysis and related decision making processes with respect to food security. There are: a) the lack of reliability of certain key food security data/information; b) lateness of some of the information generated; c) the lack of standardization of the data and information produced with often conflicting methodologies; d) the often limited control and ownership over the process of information generation of mandated national and regional institutions; e) the fact that information generated is not always properly analyzed under a food security perspective; and f) a tenuous link between food security analysis and the related decision making processes. In practice, all these problems affect the relevance and timeliness of responses to food insecurity undertaken by national and international stakeholders.

The GCP/ETH/071/EC project comes under an EC-MoFED Financing Agreement ET/FOOD/2005/17756 in which CSA had anticipated to build a food security database and website for food security related data. It is estimated that 70 to 80% of CSA datasets are food security related. There is no central database system at CSA per se. Data are entered into CSPro and extracted to other software such as SPSS and Excel for further analysis. CSA, with the support of the project, has identified the need to establish a Relational Database Management System (RDMS) which would enhance the data management efficiency, prevent losses of information by making available only processed information in pre-designed format and improve timeliness of information sharing. A centralized relational database management system will streamline existing procedures and methods, promote standards, and provide the ability of statistical and spatial analysis across time and subject matters for both CSA and other users. It will represent a powerful tool for food security analysis.

The project Steering Committee requested the project to conduct this feasibility study. The study assessed CSA's current data management infrastructure and provide recommendations for developing a centralized RDBMS for data entry, processing, storage, analysis, and dissemination. It highlights all associated infrastructures and financial requirements for an internally-sustained RDBMS. The proposed implementation approach can be 'phased' inasmuch as each data set must first be organized into a single schema which can then be compared and eventually included in the relational database.

1.2 CSA background

The Central Statistical Agency is the statistical arm of the Government of the Federal Democratic Republic of Ethiopia. Since its establishment in 1960 it has been and is involved in socio-economic and demographic data collection, processing, evaluation and dissemination that have been used for the country's socio-economic development and planning, monitoring and policy formulation. This main function of the Agency is performed through running National Integrated Household and Enterprise Survey Program (NIHESP), undertaking ad-hoc surveys, conducting census, and compilation of secondary data from administrative records.

Considering the limited resources available in Ethiopia, the NIHESP enabled CSA to run a number of annual national socio-economic and demographic surveys using the Agency's available infrastructure, field staff (enumerators, supervisors, drivers…etc), logistic support (field equipment and vehicles), data processing facilities, …etc.

Under the umbrella of the National Integrated Household and Enterprise Survey Program, the Agency plans and executes a number of national socio-economic and demographic surveys on annual basis.

The Agency has carried out several socio-economic and demographic surveys that include agriculture, price, household income, consumption and expenditure, welfare monitoring, large and medium scale manufacturing and electricity industries, small scale manufacturing industries, cottage industries, construction, mining and quarrying, transport and communications, informal sector, distributive trade and services, manpower, demography, family and fertility, health and nutrition, child labour, etc.

The Agency runs the National Integrated Survey Program on annual basis and this operation involves quite substantial number of professional staff (Statisticians, Economists, Demographers, Mathematicians, Computer programmers, etc). There are also semi professionals that include statistical technicians, data editors and coders, data entry operators, field supervisors, enumerators and other supporting staff. The Agency also occasionally undertakes an ad-hoc survey that requires specialized personnel like only female enumerators, supervisors, field editors, etc. In such cases the office hires the field staff on temporary basis for the survey period and lays them off as soon as the field work of the survey in question is completed.

The Agency has a total of 3,400 employees out of which about 1,400 of them are permanent and the remaining 2,000 are working on contract basis. Among the employees working in the Agency, about 10 percent of the employees constitute professionals from various disciplines, more than 50 percent are sub-

professionals (mainly statistical technicians, data editors and coders, data entry clerks, field supervisors and enumerators) and the remaining are supporting staffs (administrative, personnel, finance, mechanics, drivers, etc).

1.3 Objectives of the Study

The main objectives of this study are:

- To identify the various data sets/surveys that CSA collect regularly, ad hoc, and in the future
- To identify the existing data management systems for each data set/survey and the methods of entering, storing, processing, retrieving, analyzing, and disseminating data
- To identify existing and planned human infrastructure that are involved with data collection, processing, management, analysis, and dissemination
- To identify the existing and planned networking and communication infrastructure at federal and branch office level (review Master Network Plan) and its relation to the RDMS
- To assess/analyze these systems and recommend various options for building a relational database management system
- To document and present these options, providing pros/cons, hardware, software, human requirements, including cost estimates on these aspects for installation and future maintenance

1.4 Available Surveys and datasets

There are about 40 datasets which are electronically available under 10 major surveys excluding the national population census and price datasets.

Table 1: Available electronic datasets, maximum number of cases and frequency of surveys

Survey	Max. records/cases per dataset	Frequency of a survey	Available datasets	Related with food security
Annual Agricultural Sample Survey	596,326	Yearly	11	Yes
Agricultural Sample Enumeration	3,592,193	10 Years	1	Yes
Child Labour Survey	189,936	Once	1	No
Ethiopia Welfare Monitoring Survey	1,052,674	Five years	5	Yes
National Labour Force Survey		Five years	2	Yes
Large and Medium Manufacturing Industries Survey	6,223 (in part 3)	Yearly	12	No

Survey	Max. records/cases per dataset	Frequency of a survey	Available datasets	Related with food security
Livestock and Livestock Characteristics Survey	356,595	yearly	5	Yes
Urban annual Employment Unemployment Survey	60,282	Yearly	4	Yes
Demographic and Health Survey	120,660	Five years	2	No
Household Income Consumption and Expenditure Survey	3,009,993	four-five years	3	Yes
Price Survey		Monthly	132	Yes
Population and Housing Census	N/A	10 years	2	No

Cost of surveys

Out of all surveys the largest is Agricultural survey. This survey takes a large amount of money (refer to table 2).

The total operation cost of all survey is about 39 million ETB for fiscal year 2008/09.

According to the information obtained from the planning department of CSA, it is obvious that investments on the survey are the major cost of the Agency. The surveys cost 38,890,900 ETB for 2008/9 budget year while recurrent budget of the Agency is 26,998,600 ETB. The detailed investment information at a survey level is given in table 2.

Table 2: Sample operational cost of all surveys

Sr.No	Survey	Operational cost in ETB
1.	Annual Agricultural Sample Survey	23,252,000.00 (2005/6)
2.	Agricultural Sample Enumeration	131,965,000.00(2002)
3.	Child Labour Survey	3,819,000.00(2001)
4.	Ethiopia Welfare Monitoring Survey	5,910,000.00(2004)
5.	Large and Medium Manufacturing Industries Survey	1,361,000.00(2007)
6.	Livestock and Livestock Characteristics Survey	1,826,162.00(2007)
7.	Urban annual Employment Unemployment Survey	1,934,000.00(2006)
8.	Demographic and Health Survey	6,056,000.00(2005)

Sr.No	Survey	Operational cost in ETB
9.	Household Income Consumption and Expenditure Survey	19,206,000.00(2004-5)
10.	Price Survey	2,889,000.00(2007)

2. Problem-Solution

CSA have identified the following major challenges to their current information management systems. These are as follows:

1. Timeliness: Timeliness in data availability and slow data processing mechanisms
2. Security, Safety, and Standards: Security and Safety of data leading to ensured institutional memory. Standards which allow internal and external comparisons
3. Analysis and presentations: Lack of systems which enable quick and automated data analysis

2.1 Timeliness

It is obvious that data processing system takes substantial time (3/4 of the survey time) so that survey data is not available to users till the whole data processing is completed.

2.2 Security, Safety and standards

The issue of data security is important, but unfortunately the issue of backup and recovery are getting overlooked in CSA. In a lot of cases there is little guidance or support for backup and recovery of critical data - the data CSA provides to its targeted clients.

Available microdata files are in ASCII format, along with its data dictionary, and the data file in SPSS format. The backups are kept only in the Agency.

- One copy is kept in Data Processing Department
- Same copy of the data and system is kept by ICT for Central Data Bank for achieving and dissemination purposes.

The way the data is kept is not well described. This shows us that CSA is missing a backup strategy, electronic data preservations and data disaster recovery policies and guidelines.

Some of the staff members are keeping backups on their PCs which is not appropriate for any kind of backup. The use of FTP server was also not manifested including its access.

We learnt that CSA is generally following International standards such as ISIC. However the Agency has no internal standards so as to have uniform work/codes which can help it to exchange data and undertake analysis accordingly.

Locating datasets of a given survey in a central place helps the Agency to keep all electronically available datasets for future use both for technical and historical purposes. Keeping the datasets in one place helps CSA to prepare datasets for both internal and external users. However, currently this is not in a place.

2.3 Analysis and presentations

It was indicated that CSA needs to generate time series data to trace changes over time and space and to get answers of the points on the finger tips. However it has been identified that in CSA:

- Getting data over range of time and space on a finger tips is not available under the current situations
- There is no Interactive Analysis on the electronically available datasets of all surveys
- Analytical analysis can not be performed frequently, i.e. the analysis is one time to generate a targeted report

It was revealed that the use of GIS is under its inception in CSA. It has been applied only for population census of 2007.

2.4 Solution

Among all possible options, a RDBMS is the best solution for CSA as it can solve the aforementioned problems. RDBMS can help

- Audiences to get time series data over applicable space/s. The space can be the country at large, regional states, Zones within a given region.
- Audiences to exhaustively using datasets which the government or other body has invested on
- CSA to save money and enhance quality by re-using existing datasets
- To bring in positive change in planning, policy making and research results
- To fulfill the statistical data requirements essential for planning, policy formulation, monitoring and evaluation, socio-economic policy analysis, food security and research activities in general.
- A lot in setting up systems and mechanisms to ensure a sustainable flow of statistical data in Ethiopia and thereby wherever possible bridge over the existing statistical data gap.
- Get microdata for spatial analysis

Out of widely used RDBMS, MS-SQL server has been recommend because of:

- availability of experts in Ethiopia,
- partner ecosystem,

- total cost of ownership,
- its performance,
- scalability, and
- Its support for spatial functionality.

A corresponding application development programme is recommended to be web-enabled that can be ASP.NET based on anyone of the .NET programming. Here we do not expect additional software to be installed on a client machine in order to use the ISIS. Any browsers should be used to access ISIS.

2.4.1 System Development approach

Amongst the different development approaches, the most appropriate for CSA is to make use of the nationally available experts by integrating its internal staff (staff members of CSA) so as to have technology transfer. This will make the system more sustainable.

3. Recommended System Specifications

3.1 Human resources requirements

Major skills required for the proposed system are: project manager, system analyst, information architect, database administrator and database programmers.

In addition the existing system data encoders, data cleaners, and data entry personnel are required by the system both at HQ and branch offices.

3.2. Hardware requirements

Since CSA has committed to implement the master network plan, no additional hardware is required at the initial stage of the system development and implementation phase.

Establishing network connectivity of branch offices to head quarter is critical for the full-fledged implementation of RDBMS because its implementation can bring in the envisaged benefits or return.

3.3 Software requirements

The following are major software required to implement the proposed system. These are MS-SQL server 2008 enterprise edition, system development tool (Visual studio or Java).

3.4 Process

Responsibilities shall be officially given to an appropriate section of the Division to take care of the implementation process and decision should be made on which category of experts or combinations to use for the system development and maintenance. Securing fund shall also be undertaken.

The recommended system should be able to import microdata from CSPro and/or SPSS in a flexible manner. It should also be able to export microdata to SPSS, STAT and tables to MS-Excel and MS-Word for final touch of the analysis and report generations.

As an option if the Agency insists to have the existing system, CSPRO can be used as usual and the clean & final datasets can be migrated to the production server. This approach can be used only at the initial stage of ISIS. Eventually it should give a way to the use of template for the respective database. Then after, CSPro will give its way to the new system.

System maintenance: For the system maintenance there should be an agreement with system developer/s to provide support services for sometime, a year or two.

3.5 Costs associated with RDBMS

Table 5: Initial cost estimates of the proposed RDBMS

Items	Estimated cost	Remark
Database and associated software	3500 USD[1]	Minimum number license considered is five
Additional hardware at HQ and branch offices	00 USD	Implementation of the master plan assumed
Databases development	50,000USD[2]	For national company or expert
Migration existing datasets	10,000USD[3]	
Training on MS-SQL server for DBA	1,000 USD	For professionals who are working on the system
Training on how to use the developed system	200USD	Two days training for different group of users

[1] One time cost unless upgrading of the software is required.
[2] It has been estimated that on an average each of the survey needs about 5,000USD for national company or expert.
[3] Migration of datasets for each survey has been estimated to be 1,000USD on an average.

Total estimated cost =	**64,000USD**	

3.6 System migration

Based on the need of CSA and benefits which can be obtained from the conversion the surveys for which CSA needs to develop RDBMS is suggested to be in the following priority order: These are:

Category I: monthly datasets

- Price surveys

Category II: yearly datasets

- Annual Agricultural Sample Survey
- Livestock and Livestock Characteristics Survey
- Large and Medium Manufacturing Industries Survey
- Urban annual Employment and Unemployment Survey

Category III: Less frequent datasets

- National Labour Force Survey
- Ethiopia Welfare Monitoring Survey
- Demographic and Health Survey
- Household Income Consumption and Expenditure Survey

The development and conversion timing of system depends on the type of employed experts and/or availability of fund and commitment of the senior managements.

Part Two: Detailed Survey Information System Study

Table of contents **Page**

Part Two: Detailed Survey Information System Study

1. Introduction

1.1 Overview of Activities

The consultancy project was divided into four major tasks. These are to asses and Document existing Data Sets and Data Management Systems; to assess and document existing Analytical Systems; to assess and document existing Reporting and dissemination Systems; and to present Recommendations for a RDBMS which will be used in CSA.

1.2 Brief introduction to the CSA

The Central Statistical Agency is the statistical arm of the Government of the Federal Democratic Republic of Ethiopia. Since its establishment in 1960 it has been and is involved in socio-economic and demographic data collection, processing, evaluation and dissemination that have been used for the country's socio-economic development and planning, monitoring and policy formulation. This main function of the Agency is performed through running National Integrated Household and Enterprise Survey Program (NIHESP), undertaking ad-hoc surveys, conducting census, and compilation of secondary data from administrative records.

Considering the limited resources available in Ethiopia, the NIHESP enabled CSA to run a number of annual national socio-economic and demographic surveys using the Agency's available infrastructure, field staff (enumerators, supervisors, drivers…etc), logistic support (field equipment and vehicles), data processing facilities, …etc.

Under the umbrella of the National Integrated Household and Enterprise Survey Program, the Agency plans and executes a number of national socio-economic and demographic surveys on annual basis.

The Agency has carried out several socio-economic and demographic surveys that include agriculture, price, household income, consumption and expenditure, welfare monitoring, large and medium scale manufacturing and electricity industries, small scale manufacturing industries, cottage industries, construction, mining and quarrying, transport and communications, informal sector, distributive trade and services, manpower, demography, family and fertility, health and nutrition, child labour, etc.

The Agency runs the National Integrated Survey Program on annual basis and this operation involves quite substantial number of professional staff (Statisticians, Economists, Demographers, Mathematicians, Computer

programmers, etc). There are also semi professionals that include statistical technicians, data editors and coders, data entry operators, field supervisors, enumerators and other supporting staff. The Agency also occasionally undertakes an ad-hoc survey that requires specialized personnel like only female enumerators, supervisors, field editors, etc. In such cases the office hires the field staff on temporary basis for the survey period and lays them off as soon as the field work of the survey in question is completed.

The Agency has a total of 3,400 employees out of which about 1,400 of them are permanent and the remaining 2,000 are working on contract basis. Among the employees working in the Agency, about 10 percent of the employees constitute professionals from various disciplines, more than 50 percent are sub-professionals (mainly statistical technicians, data editors and coders, data entry clerks, field supervisors and enumerators) and the remaining are supporting staffs (administrative, personnel, finance, mechanics, drivers, etc).

1.3 Organizational structure of Central Statistical Agency

CSA has 25 branch offices in the country and head quarter in Addis Ababa. These Branch Statistical Offices are responsible for coordinating the data collection activities in rural and urban sample sites (Enumeration Areas). Moreover, for the execution of the first ever National Agricultural Sample Census Enumeration

A Director General and three Deputy General Directors currently head the Central Statistical Agency. The deputy general Director for Economic Statistics leads three technical departments namely:

- Natural Resources and Agricultural Statistics
- Industry, Trade and Services Statistics and
- Household Budget, Welfare Monitoring & Price Statistics

On the other hand, the Deputy General Director for social and Demographic Statistics leads:

- Population and housing statistics,
- Vital statistics and
- Man power and social statistics departments.

The five service rendering departments under the Deputy General Director of Operation, Methodology and Data Processing are:

- The Statistical methodology
- The Regional Offices and Field Operations
- Data processing
- Information Communication Technology (ICT) Development departments and

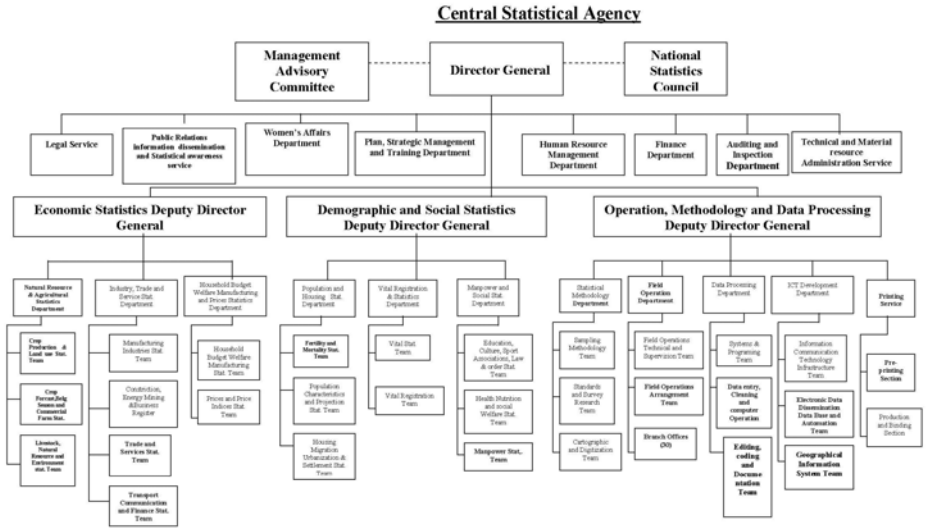
- The printing service

In addition to the above mentioned departments headed by the Deputy Director Generals, there are nine support rendering units that are directly accountable to the Director General. these are:

- Legal Service
- Public Relations & Data Dissemination Service
- Women Affairs Department
- Plan and Program Department
- Human Resource Management Department
- Finance Department
- Audit and Inspection Department
- Technical and material Resource Administration Service

The departments are further subdivided into three or four expert teams (see organizational chart of the agency).

Figure 1: Organogram of CSA



1.4. Methodology

The method of primary data collection was in-depth and used semi-structured interviews. Heads or representatives of some of departments of CSA and heads of the departments under the Operation, Methodology and Data Processing division were interviewed individually. The responsible persons interviewed were drawn from different departments within the Agency. The interview explored the views, experiences and concerns of the respondents in relation to the use of information systems and their roles and responsibilities.

In addition to these interviews, observation carried out for short period of time. Documentary materials such as relevant reports, formats, organizational charts and other publications were also studied.

In undertaking the study structured system analysis and design methods has been employed.

2. Completed, latest and planned Surveys and their audiences

2.1 Completed and latest surveys

As one of CSA's initiative to do better dissemination by making survey and censes data readily available in a user friendly manner to data users, a number of surveys have been compiled and archived to be available to the public. The following list of surveys provides access to the individual survey metadata and documentation as well as variable level information through the website and CD-ROMs. These surveys are:

1. Agricultural Sample Enumeration
2. Annual Agricultural Sample Survey
3. Child Labour Survey
4. Ethiopia Welfare Monitoring Survey
5. Labour Force Survey
6. Large and Medium Manufacturing Industries Survey
7. Livestock and Livestock Characteristics Survey
8. Urban annual Employment Unemployment Survey
9. Demographic and Health Survey
10. Household Income Consumption and Expenditure Survey
11. Population and Housing Census
12. Price Survey

Sample can vary from survey to survey with some additions or deletions of indicators. Frequency of survey varies from one to another as indicated in table 1.

It has also been identified that a given survey follows its own sampling procedure based on its own sampling design.

Each of the survey has its own procedures for data collection. In fact field work organization, training of field staff, data editing, cleaning, tabulations follow similar process for all surveys except the industry survey.

There are about 40 datasets which are currently electronically available under 11 major surveys excluding the national population census and price datasets.

Table 1: Available electronic datasets, maximum number of cases and frequency of surveys

Sr.No	Survey	Number of max. records/cases	Frequency of a survey	Available datasets
1	Annual Agricultural Sample Survey	596,326	Yearly	11
2	Agricultural Sample Enumeration	3,592,193	10 Years	1
3	Child Labour Survey	189,936	Once	1
4	Ethiopia Welfare Monitoring Survey	1,052,674	Five years	5
5	National Labour Force Survey		Five years	2
6	Large and Medium Manufacturing Industries Survey	6,223 (in part 3)	Yearly	12
7	Livestock and Livestock Characteristics Survey	356,595	yearly	5
8	Urban annual Employment Unemployment Survey	60,282	Yearly	4
9	Demographic and Health Survey	120,660	Five years	2
10	Household Income Consumption and Expenditure Survey	3,009,993	four-five years	3
11	Price Survey		Monthly	132
12	Population and Housing Census	N/A	10 years	2

Cost of surveys

Out of all surveys the largest is Agricultural survey. This survey also takes a large amount of money (refer to table 2).

The total operation cost of all survey is about 39 million ETB for fiscal year 2008/09.

It is obvious that investments on the survey are the major cost of the Agency. The surveys cost 38,890,900 ETB for 2008/9 budget year while recurrent budget of the Agency is 26,998,600 ETB. The detailed investment information at a survey level is given in table 2.

Table 2: Sample operational cost of all surveys

Sr.No	Survey	Operational cost in ETB
11.	Annual Agricultural Sample Survey	23,252,000.00 (2005/6)
12.	Agricultural Sample Enumeration	131,965,000.00(2002)
13.	Child Labour Survey	3,819,000.00(2001)
14.	Ethiopia Welfare Monitoring Survey	5,910,000.00(2004)
15.	Large and Medium Manufacturing Industries Survey	1,361,000.00(2007)
16.	Livestock and Livestock Characteristics Survey	1,826,162.00(2007)
17.	Urban annual Employment Unemployment Survey	1,934,000.00(2006)
18.	Demographic and Health Survey	6,056,000.00(2005)
19.	Household Income Consumption and Expenditure Survey	19,206,000.00(2004-5)
20.	Price Survey	2,889,000.00(2007)

2.2. Planned surveys

The surveys conducted in 2007/2008 are under process completion. These include:

1. Natural Resources and Agricultural Statistics

- Crop Production Forecast Survey
- Crop Production Survey (meher season)
    - Survey on Area and Production of Crops
    - Land Utilization Survey (Statistical tables to be produced soon)
    - Survey on Farm Management practices (Statistical tables to be produced soon)

- Livestock, poultry and Beehives Survey
- Crop production Survey (Belg season)
    - Survey on Area and production of Crops
    - Survey on Farm Management Practices
- Survey of State and Private Commercial farms

2. Industry, Trade, Transport and Communication Statistics

- Survey of Medium & Large Scale Manufacturing Industries
- Compilation of Transport & Communication

3. Consumer Price and Producer Price survey

Datasets of the above surveys are expected to be more or less similar to their pervious once.

CSA has planned the following surveys to be undertaken for the Ethiopian fiscal 2001 to 2003. The plan is tentatively set and subject to change whenever required.

2.2.1 Planned for 2008/9 (2001E.C)

Natural Resource and Agriculture Statistics

Crop production Forecast
Crop Production Survey /for Long Rainy Season/
Land Utilization Survey
Agricultural Inputs and Practices Survey /Farm management/
Livestock, Poultry and Beehives Survey
Crop Production Survey /for Short Rainy Season/
Survey of Large and Medium Scale Commercial Farms

Industry, Trade and Services Statistics

Survey of Medium and Large Scale Manufacturing Industries
Survey of Producers' price of Manufactured Products
Survey of Small Scale Manufacturing Industries
Survey of Contract Construction
Survey of Distributive Trade and Services
Compilation of Foreign Trade Statistics
Compilation of Transport & Communications Statistics

Household Budget and Price Statistics

Survey of Producers' price of Agricultural Commodities
Survey of Retail prices of Goods and Services
Construction of Consumer Price Index for National, Regional and Addis Ababa
Construction of Producer's Price Index for Agricultural Commodities

Demographic and Social statistics

Vital Events Statistics
Current Employment Survey

Pilot Survey of Orphanage and Street Children
Preparation of Analytical Report for 2007 Population Census

2.2.2 Planned for 2009/10 (2002E.C)

Natural Resource and Agriculture Statistics

Crop production Forecast
Crop Production Survey /for Long Rainy Season/
Land Utilization Survey
Agricultural Inputs and Practices Survey /Farm management/
Livestock, Poultry and Beehives Survey
Crop Production Survey /for Short Rainy Season/
Survey of Large and Medium Scale Commercial Farms
Rural Socio-economic Survey
Compilation of Wild Life Statistics

Industry, Trade and Services Statistics

Survey of Medium and Large Scale Manufacturing Industries
Survey of Producers' price of Manufactured Products
Survey of Small Scale Manufacturing Industries
Survey of Contract Construction
Survey of Mining & Quarrying Statistics
Census of Economic Establishments/Enterprises
Compilation of Foreign Trade Statistics
Compilation of Transport & Communications Statistics
Compilation of Energy production Supply and Consumption

Household Budget and Price Statistics

Survey of Producers' price of Agricultural Commodities
Survey of Retail prices of Goods and Services
Construction of Consumer Price Index for National, Regional and Addis Ababa
Household Income, Consumption and Expenditure Survey
Construction of Producer's Price Index for Agricultural Commodities

Demographic and Social statistics

Vital Events Statistics
Current Employment Survey
Demographic and Health Survey (DHS)
Preparation of Analytical Report for 2007 Population Census

2.2.3 Planned for 2010/11 (2003E.C)

Natural Resource and Agriculture Statistics

Crop production Forecast
Crop Production Survey /for Long Rainy Season/
Land Utilization Survey
Agricultural Inputs and Practices Survey /Farm management/
Livestock, Poultry and Beehives Survey
Crop Production Survey /for Short Rainy Season/
Survey of Large and Medium Scale Commercial Farms

Rural Socio-economic Survey
Compilation of Wild Life Statistics
Comprehensive Land Use Survey
Pilot Survey of Fishery Statistics

Industry, Trade and Services Statistics

Survey of Medium and Large Scale Manufacturing Industries
Survey of Producers' price of Manufactured Products
Survey of Small Scale Manufacturing Industries
Survey of Contract Construction
Survey of Distributive Trade and Services
Survey of Mining & Quarrying Statistics
Compilation of Foreign Trade Statistics
Survey of Informal Sector
Compilation of Transport & Communications Statistics
Compilation of Water Supply & Consumption Statistics
Compilation of Energy production Supply and Consumption

Household Budget and Price Statistics

Survey of Producers' price of Agricultural Commodities
Survey of Retail prices of Goods and Services
Construction of Consumer Price Index for National, Regional and Addis Ababa
Household Income, Consumption and Expenditure Survey
Welfare Monitoring Survey
Construction of Producer's Price Index for Agricultural Commodities

Demographic and Social statistics

Vital Events Statistics
Current Employment Survey
Survey of Orphanage and Street Children
Time Use Survey
Health and Nutrition Survey
Compilation of Employment Statistics
Construction of Wages and Salaries Index

2.3. Audiences of surveys/datasets

The survey data have a lot of users at different reporting levels. The major audiences are policy makers of the federal government and regional states and planners at different level. Generally the following are also target audiences:

- Researchers
- Students, particularly of higher learning institutes
- Journalists
- Sponsors of different Research on related subjects
- National and international survey sponsors
- Advocates / General Public
- UN and International organizations

3. Common Entities of all datasets and Variations within a survey

3.1 Common Entities of all datasets

In section 2 of the report1, all available datasets for the 12 surveys have been documented and assessed. In addition some of the attributes have been identified to be common for all datasets/surveys and their reports. The identified common entities are described in the following set of tables.

Table 4: Identification particulars

#	Attributes
1.	Region
2.	Zone
3.	wereda
4.	Farmer's association
5.	Enumeration area

Table 5: Survey identifications

#	attributes
	Country
2.	Title
3.	Abbreviation
4.	Survey Type
5.	ID Number
6.	Version Description
7.	Production Date
8.	Abstract
9.	Kind of Data
10.	Unit of Analysis
11.	Scope
12.	Description of Scope
13.	Geographic Coverage
14.	Universe
15.	Primary Investigator(s)
16.	Funding
17.	Metadata produced by
18.	Date of metadata publication
19.	DDI Document Version
20.	DDI Document ID

Table 6: Sampling Procedure

#	attributes
1.	Sampling frame
2.	Sample design
3.	Sample size and selection scheme
4.	Response rates

Table 7: Questionnaire

#	attributes
1	Overview
2	Form/s
3	Data_collector

Table 8: Forms for dataset

#	attributes
1	Form Identification
2	Form Description

Table 9: Data collection methods

#	attributes
1	Field_Work
2	Editing_Coding_Verfication
3	Training_Of_field _Staff
4	Entry_cleaning_Tabulation

Table 10: Data processing

#	attributes
1	Data Editing, Coding and Verification
2	Data Entry, Cleaning and Tabulation

Table 11: Available and associated technical documents

#	attributes
1	Title
2	Author
3	Date_of_Publications
4	Language
5	Country

6	Publisher
7	Document
8	Document_ID

Table 12: Applied Access policy

#	attributes
1	Access_Authority
2	Access_Condition
3	Citation
4	FullText

3.2. Variations within a survey

All responsible SMSs for each of the survey confirmed that there are some differences among datasets of a given survey. The identified differences among the surveys of different years are:

- Coverage area (EAs)
- Adding more indicators
- Adjusting some indicators
- Deleting some indicators

Most of the time, they are adding some questions to a given survey.

The RDBMS designer should consider these variations in the database design (make it flexible) and setting migration process from current system to the new one.

In fact, because of the feedback obtained from main users of data and generated report on Household Income Consumption and Expenditure Survey, the House hold Department has decided to come up with different questionnaire and approach. It is believed that the data gap can be narrowed through the newly designed survey. The RDBMS designer should wait and see what will come as a result of the suggested change before commencing the database.

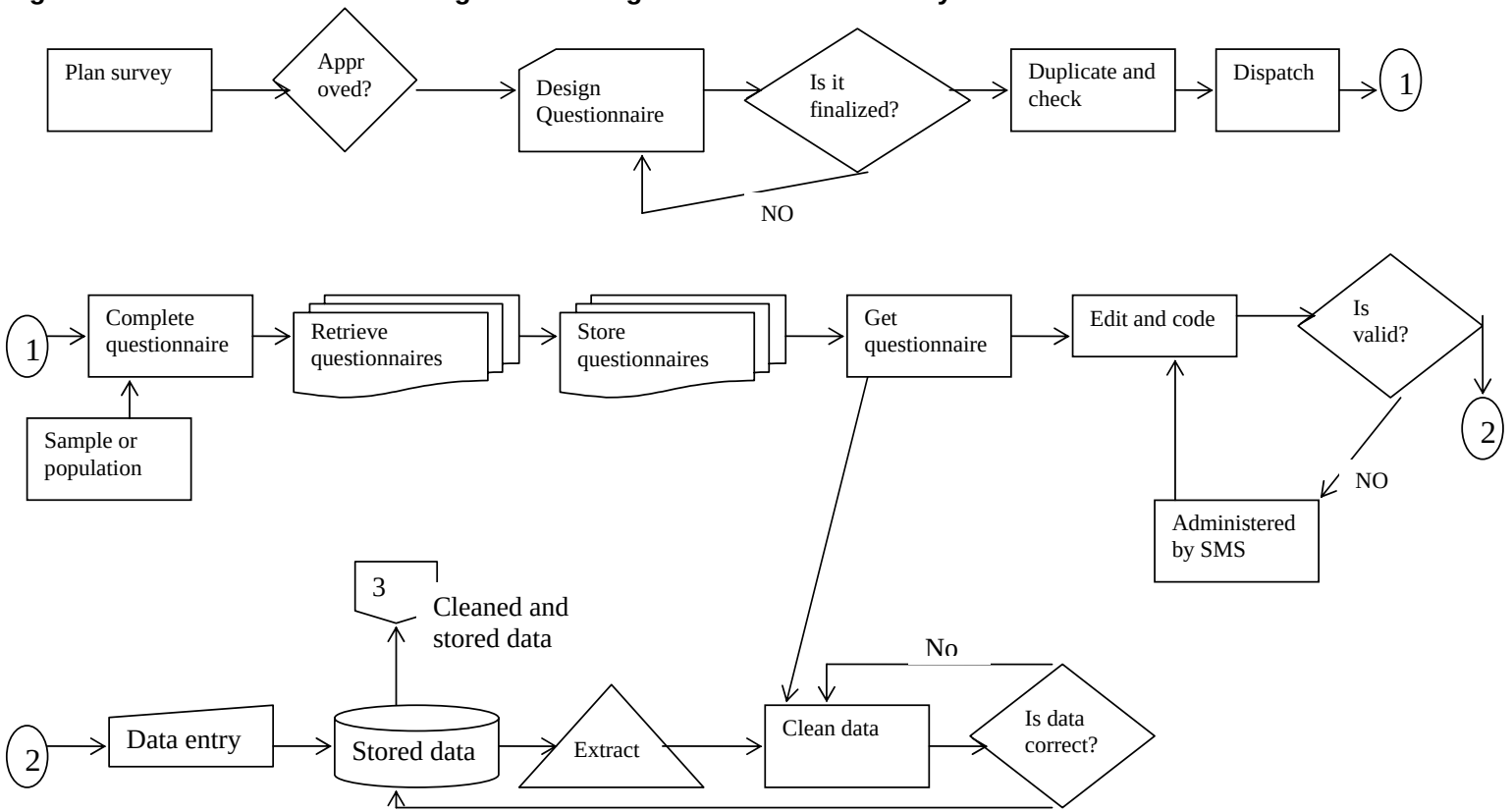
4. Dataflow and document flow of surveys/datasets

Table 15: Major activities being performed

Major activity in sequential order
1. Survey planned
2. Survey approved
3. Questionnaire

preparation/design
4. Questionnaire dispatch
5. Questionnaire filling
6. Questionnaire retrieval/collection
7. Questionnaire storage
8. Editing and coding
9. Data entry
10. Data cleaning
11. Reformatting data
12. Report generation(tabulation)
13. Data preparation/conversion
14. System documentation
15. System backup
16. Information dissemination

Figure 3: Flow chart for Data management for a given dataset of a survey



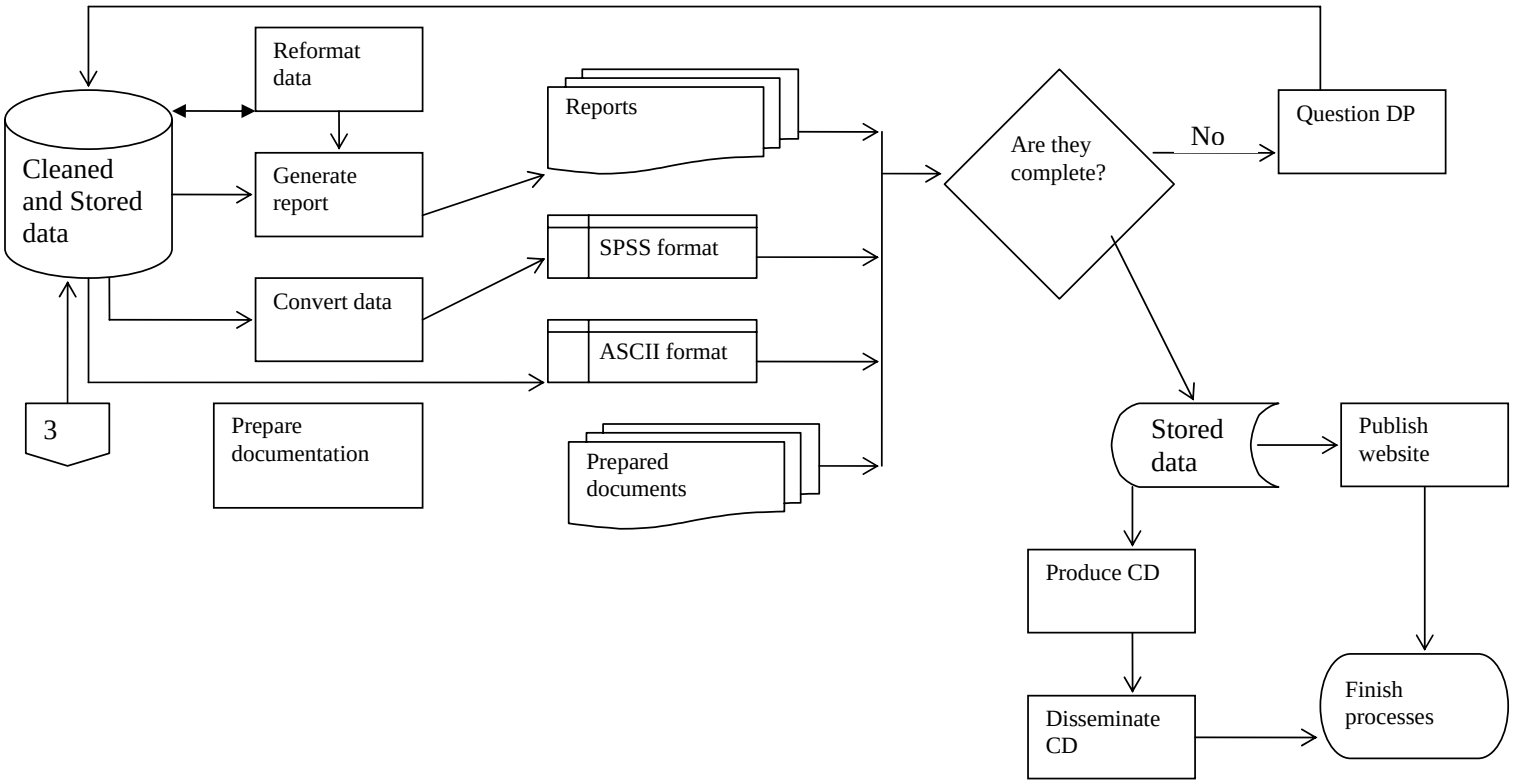
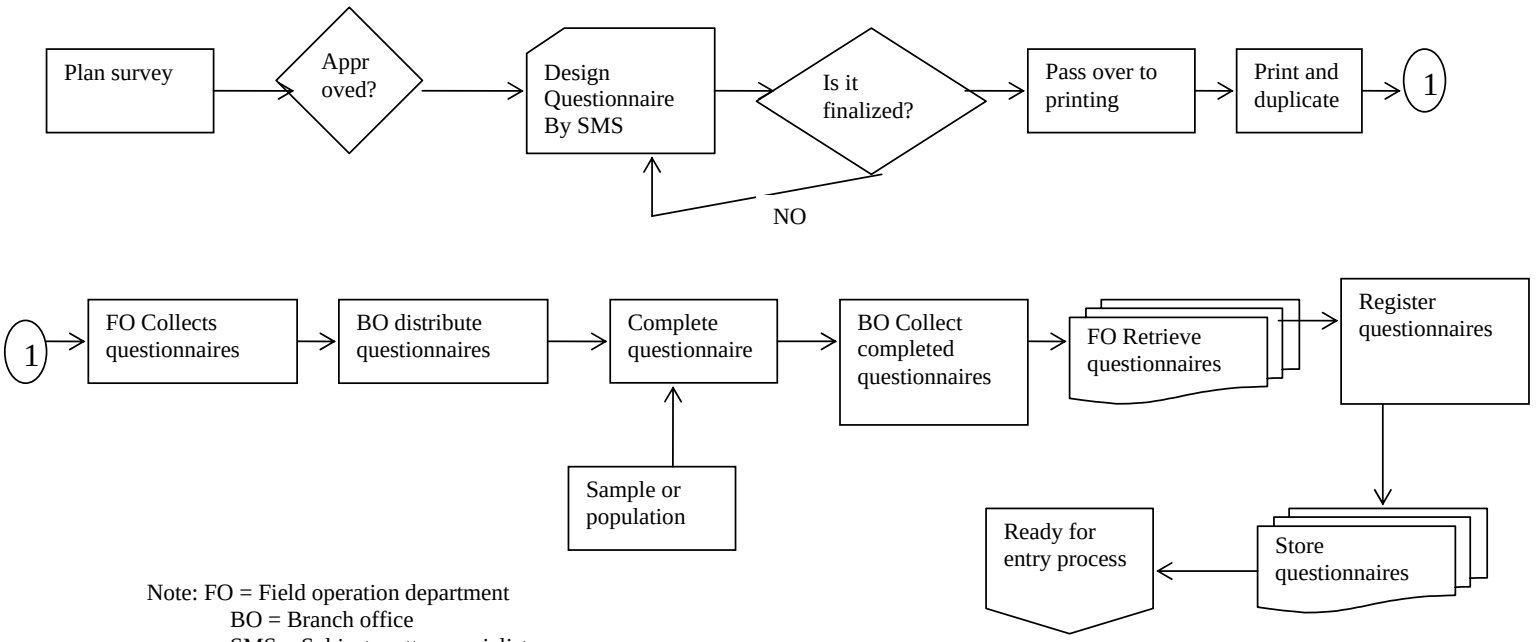

Reformat data
Cleaned and Stored data
Generate report
Reports
Are they complete?
No
Question DP
SPSS format
Convert data
ASCII format
3
Prepare documentation
Prepared documents
Stored data
Publish website
Produce CD
Disseminate CD
Finish processes

Figure 4: Questionnaire/s flow in CSA



Note: FO = Field operation department
BO = Branch office
SMS = Subject matter specialist

Flow of questionnaires

- Questionnaires will be ready by Subject matter specialists and passed over to printing for duplications
- Field Operation Department will collect the ready questionnaires
- Duplicated and ready questionnaires are distributed to branch offices through the Field Operation Department
- Filled-in questionnaires are retrieved from Statistical Branch Offices.
- Documents are received, systematically registered and organized at the documentation unit of the Data processing Department at the HQ.
- Dispatched to the editing and coding section for manual editing and coding purposes.
- Documents dispatched are returned to the documentation unit after manual editing, data entry and data cleaning activities are completed.
- After data cleaning is completed the survey questionnaires are being kept for three to four years based on availability of storage space
- Old questionnaires will be weed out by keeping few samples

5. Platform and infrastructure

So far telecommunication system has not been utilized to collect data from field levels. But there is a plan to use telecommunication infrastructure in the near future to connect branch offices to the head quarter through Wide Area Network (WAN). At the moment 10 branch offices have been selected ready to get connected to the head office through 64KB leased line.

5.1 Software systems

IMPS, CSPro, SPSS, SAS are being used in the organization. IMPS, CSPro are being used for data entry, editing (validation), and cleaning whereas SPSS and SAS are being used for processing purposes. All the software is being used by Data Processing department and subject material specialists. The major software packages used to process survey and census data are: Integrated Microcomputer Processing System (IMPS), Census and Survey Processing System (CSPro), and SPSS. ICT development department is also using IHSN/Nesstar and SPSS for dissemination and archiving purposes.

CSPro is the most important software which is being used in CSA.

MS-SQL version 2000 server DBMS is being used to handle price data and Large & Medium Manufacturing Industries datasets.

Currently, the Agency is using Windows platform both at server and workstation levels. Windows 2003 server operating system and Windows XP is being used for workstations.

The Agency is using Microsoft Office for different purposes. From DBMss Access and MS-SQL DBMS are in use. From the spreadsheet only Excel is in use.

Table 14: Software and personnel involved in data management

Software	Departments/Branch offices who are using the respective system
CSPro	Data Processing Department
IMPS	Data Processing Department
SPSS	All departments (SMS)+ ICT +Data Processing
SAS	Data Processing Department
MS-SQL server	ICT
MS-Excel	Data Processing Department + ICT
MS-Access	
Other MS-Offices	All staff membes of CSA who has PC
IHSN	ICT
PDF maker (specify)	ICT
EUROTRACER	Industry, Trade and Services Department

Note that CSPro and IMPS are free software which were obtained from US bureau of Cencus. CSPro is a software tool for entering, editing, tabulating, managing and disseminating data from surveys and censuses. IHSN is also free software from the World Bank.

5.2 Hardware

CSA has eight servers. These are Active Directory, file, web, backup/storage, ISA, anti-virus and mail servers. All of them are single CPU servers and some of them are older than five years. All the servers are being administered by the ICT Development Department.

It has been estimated that there are more than 200 workstations in the agency and out of which about 120 workstations are in Data Processing and ICT Development Departments. It has been identified that each branch offices do have at least two PCs each.

In fact, all the ICT development and Data processing staff members are using standalone computers to under take their day today activities.

There are two LANs in the institution and connected through fiber optics. The LANs are running on 5e UTP cable. Different switches and hubs are being used.

Currently there is no Wide Area Network which interconnects branch offices to the Head quarter.

It has been suggested and agreed that the ICT facilities have been well-used.

Cost of hardware and software

The Agency has purchased computers at different years through different sources of fund and got a number of PCs through donations. Therefore, it is not easy to get the exact cost hardware and software. However, it has been estimated as shown in table 3.

Table 3: Estimated cost of currently operational hardware and software related to data processing and ICT

Category	**Items**	**Qty/license number**	**Estimated cost[4] ETB**	**Remark**
Hardware	computers	250	2,500,000	
	Servers	10	400,000	
	Plotters		80,000	
	Scanners		20,000	
	DRS Scanners	11	2,200,000	
	Digitizers		100,000	
Software	SPSS	1	30,000	
	MS-SQL server	1	1,850	
	STATA	1	70,000	
	Nesstar	1	free	
	IHSN	1	free	
	CSPro	1	free	
	IMPS	1	free	
	Server Operating systems		n/a	Provided with the hardware
	Harvard graphics	1	free	
	EURO Tracer	1	free	

[4] Please note that this does not necessarily imply actual investment of CSA, rather it estimates what are currently available

Category	Items	Qty/license number	Estimated cost[4] ETB	Remark
	PDF maker	1	free	
	Backup tool	1	free	
Total			**5,401,850**	

However, it has been revealed that cost of software is insignificant because most of the software in use are free softwares.

5.3 Level of Internet utilization

CSA is using 256KB broadband/leased line to access Internet services. Staff members who are positioned team leaders and above do have 24 hours access to Internet while other staff members do have half a day access. Thus, the numbers of staff members who do have full access to Internet are about 50. It has been revealed that all PCs which are connected to the network do have Internet connectivity whereas some of the PCs are not connected to LAN because of their contribution and role. The ICT Development Department is claiming that the Internet has been used in its full capacity.

Filtering of information is not being done at the CSA level. So far the ICT Development Department has not received complaints about use of Internet from CSA. The security system of internet connection is very good, because they are using both hardware and software firewalls.

Cost of telecommunications and Internet

The Agency is using telephone, fax and internet in the head office. Since there costs are covering all uses, it is not easy to find out how much has been used in direct relation to survey activities. This is also true to get how much has been used for telephone and fax at the 25 branch offices. However, telephone and fax budget for the 25 branch offices is 285, 000 ETB.

The cost of the 256KB broad band (ADSL) for BFWA - Data Only; which is used at the head office is costing 6,567 ETB per month which is 78,804 ETB per annum (source ETC website).

5.4 Review of proposed Network infrastructure

This section of the report covers how the newly proposed network infrastructure can support the intended RDBMS for CSA. The proposed master plan for network (LAN and WAN) covers all technical details and managerial components of the network. This part of report is focusing on data entry, updating, storage

and retrieval. The report is not reviewing the technical design, standards and materials.

Rather it is emphasizing on components which are directly affecting the RDBMS from data entry to information dissemination. The Master Plan report proposes 10 PCs to be used at the branch offices level. They allocated five for data entry and five for other purposes. This figures put the number of data entry personnel to 125 in the 25 branch offices. This number is by far greater than the existing data entry staff member which is 70 at the head office. In addition the estimate shows us that the amount of data coming from each branch office is equal. The base for having 5 PCs in each branch office (125 data entry personnel) is that they need to reduce timing of data entry which has positive impact on timing of report generation.

However, cost implications on staff cost, staff relocation, office space, running cost, etc for the Agency has not been mentioned in the report. Since the data entry will be done from the branch offices, data coding and editing activities are also accomplished at branch office levels.

The minimum bandwidth has been calculated to be 23kbps and suggest for branch offices to be 64kbps frame relay. Here it should be clear enough that 64kbps is a connectivity which is supposed to be used by individuals from home in the near future, because individuals from developed countries are subscribing to 1Mbps lines. Therefore scalability and expansion of the system should be there.

The use of PDA for remote data entry from the field is covered. Use of dial-up system is proposed to be secured particularly by using callback in addition to others.

The uses of LAN at the branch offices are not clearly stated. Is the server used only for a backup purpose or has it any other role in using central system? Here the server is expected to have multi-purpose including hosting databases.

Connectivity of router for the frame really should be able to support dialup feature incase if the leased line is down as we can see currently just to minimize downtime of the network.

Network operating system or server operating system is not clearly set by indicating pros and cons of using open source and commercial systems. This could help CSA to select appropriate one for its future plan. The management of branch office LANs are not clearly described in the sense of how the system administrators and network administrators from the head office are interfering or assisting branch office systems remotely.

In the case of RDBMS implementation and in regard to the network operating system, it depends on which DBMS CSA is using in the near future.

Synchronization of database at branch office with the Central database at the HQ should be allowed by the system.

Whatever the system we are proposing Database Administrator should be able to have full access to the branch office version of the database and be able to assist data entry personnel of a given branch offices at any time.

5.5. Review of existing relational databases

Currently CSA is on a process to develop and implement relational databases for two surveys. These are Price and Large & Medium Manufacturing Industries Surveys.

The price Database is developed on MS-SQL server 2000 by using ASP script. The database had three tables at the beginning which were related to one another. After getting some feedback the number of tables has increased to 7. The presentation component is on a process of implementation. Migrations of datasets of three years were done for the price database.

Industry database contains about 28 related tables. The lists of tables with their respective fields are given in the following table.

Table Name	**Number of Fields**
General_Address	11
Head_Office	10
Industrial_Category	2
Establishment_Information	36
Capital	11
General_Occupation	2
Engaged_Person	14
Temporary_Employee	4
Salary	5
Employees_Benefit	4
Salary_Group	27
Product_List	5
Production_Data	7
Sales	7
Other_Product	5
Other_Income	10

Stock	12
Raw_Material_List	5
Raw_Material_Used	10
Non_Principal_Material	7
Industrial_Cost_List	4
Industrial_Cost	4
Other_Industrial_Cost	8
Non_Industrial_Expense	12
Tax_Paid	5
Fixed_Asset_List	2
Fixed_Asset	7
Financial_Source	9
Total number of fields =	**245**

The attempts are showing a good start. In fact well organized system documentation is required so that any professional can understand how the system has been designed and developed. Having schema definition is a very good habit for database designer and database programmers.

6. Institutional setup for survey information management

The institutional setup for data management is being lead by the Deputy Director General. Five service rendering departments under the Deputy General Director of Operation, Methodology and Data Processing are:

- The Statistical methodology
- The Regional Offices and Field Operations
- Data processing
- Information Communication Technology (ICT) Development departments and
- The printing service

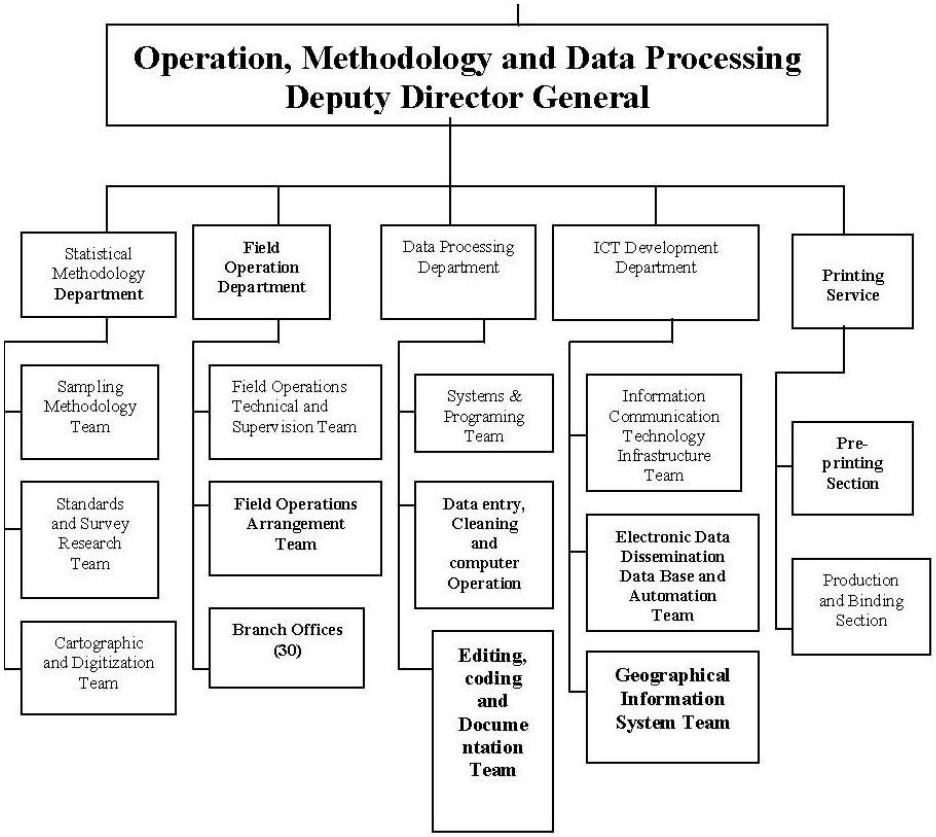


Figure 5: Organogram of Operation, Methodology and Data Processing Division

The Operation, Methodology and Data Processing division of CSA is giving services to the two line divisions. The major services are:

- Survey design services
- Data processing services
- Field operation services
- Information archiving and dissemination services
- Printing services

Information analysis is being undertaken by SMS not by this division. Horizontal communication is among the routine activities. DDG is involving in the activities of departments whenever there is a communications problem and there is a need to have an intervention.

Figure 6: Flow of activities with in the division is in the following order

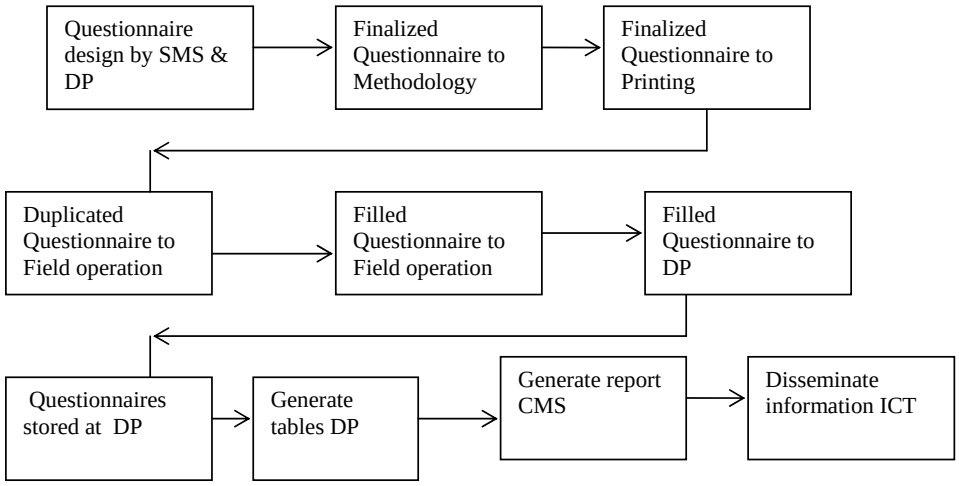


Figure: Flow of activities with in the division

6.1 Data processing Department

On an average, there are around 15 (both regular and ad-hoc) or more surveys conducted and processed at the organization every year. All the annual and periodic socio-economic surveys and censuses conducted by the subject-matter departments of the CSA; namely: Natural Resources and Agricultural Statistics, Industry, Trade, Finance, Transport and Communication Statistics, Household Budget and prices Statistics and Social Statistics. They are all processed at the Data Processing Department.

Major objectives of the data processing department are to:

- Analyze design and implement various surveys and census data processing systems.
- Develop, maintain and apply computer programs used in capturing, cleaning, organizing and generating surveys and census result.
- Produce accurate and timely electronic data that meets the users' needs.
- Organize and access raw data and summarized statistical reports to government and non-government agencies, researchers, and various local and international organizations.

The accomplishment of the above mentioned objectives basically follow four phases. These are:

1. Data editing and coding
2. Data entry

3. Data cleaning
4. Data analysis

In the data editing and coding phase, any inconsistencies are checked and corrected on the questionnaires obtained from the field in addition to giving various codes. On a regular basis, most of the time the process of editing and coding is first and then verifying is following. In the data entry phase, the edited and coded questionnaires are captured into the computer. Similar to editing and coding activities, the data entry work is verified thoroughly and completely for most of the surveys undertaken by the Agency. Before the data is used for analysis, the data will pass through a number of edit programs to validate the data and it is done by the data cleaning staffs of the department. In this cleaning phase, any invalid values are checked and corrected on the data. The final step after obtaining the clean data is attaching weights, analyzing the results, and producing various tables including the provision of variances for reliability measures of the data. There are about 90 computers for data entry and cleaning purposes and about 14 for data analysis in the department in order to perform the above mentioned activities.

In the Data Processing Department there are 11 systems and programming specialists; 70 data entry personnel; 15 cleaning personnel. In general, there are 96 professional and semi-professional staff members in the Department. The system and program team members' qualifications are ranging from BSC to MSC in Computer Science, Mathematics, and Statistics field of study. The other staff members' qualification is at diploma or equivalent level except the team leaders in the Department.

6.2 ICT Development Department

The ICT Development Department is responsible to keep metadata standardization for all surveys. It creates a linkage between metadata and micro data. Generally stating it assures quality of datasets output and their presentations.

What is a role of ICT Department in backup and storage? The department is responsible to take a backup and store datasets and reports for future use. One complete copy of the data and system is being kept by the ICT department for Central Data Bank for achieving and dissemination purposes.

Data automation and Disseminations of the ICT Development Department to build CD-ROMs and publish out put on the website by using IHSN toolkit. In addition it is responsible to take backups. This Department is not generating raw data except having few details. All the output designs and presentation on the website and CD-ROMs are done by Data automation and Disseminations section.

The Department receives cleaned and ready data in SPSS format and ASCII formats survey data from Data Processing Department and at the end PDF files are generated from MS-word reports by using IHSN toolkit.

The Department has a team which is dealing with GIS called Geographic Information System team which is responsible to prepare appropriate spatial data for the Agency so as to have all survey data to be accessible through GIS.

The ICT infrastructure team of the department is providing infrastructure to the Agency including all necessary support services.

Currently, the ICT development Department has 23 staff members whose qualification ranges from BSC to MSC. Their specialties include website developers, database specialists, ICT infrastructure administrators, GIS specialists and technicians.

6.3 Regional offices and Field Operations Department

All the 25 branch offices do have appropriate road infrastructure, telecommunication infrastructure and hydro electric power. Most widely used communication tool for routine activities is being made through fax. The staffs at the branch office level do not get involved in making analysis, storing, and dissemination of information.

Table 15: Branch offices and their human resources

#	**Branch office**	**#EA**	**Coordinators**	**Supervisors**	**Editors**
1.	Shire	56	1	6	0
2.	Mekele	106	3	11	0
3.	Asaiyta	45	1	17	1
4.	Gonder	88	2	26	4
5.	Dessie	156	2	31	6
6.	Debreberhan	50	1	21	3
7.	Bahir-dare	170	2	35	3
8.	Asebe Teferi	56	1	21	3
9.	**Harer**	49	1	20	3
10.	**Dire Dawa**	44	1	17	3
11.	Adama	109	2	31	4
12.	Goba	57	1	18	0
13.	Negelle	74	2	18	0
14.	Ambo	114	3	32	4
15.	**Jijiga**	29	1	11	2
16.	Assosa	55	1	20	1
17.	Nekemet	153	3	41	2

18.	Hossaena	155	2	24	5
19.	Awassa	138	3	29	3
20.	Sodo	91	2	14	0
21.	Arbamench	131	3	20	1
22.	Jimma	102	2	33	6
23.	Mizane	112	3	30	1
24.	**Gambella**	60	1	16	1
25.	**Addis Ababa**		2	14	3
	Total =	**2200**	**46**	**556**	**59**

There are 2200 EAs in Ethiopia which have been established by CSA for its data collections for various surveys. In undertaking the data collection it uses its 46 coordinators and 556 supervisors based on the sample size of a survey.

According to the new organizational structure there are 59 data editors in the branch offices.

Qualifications of Enumerators, field supervisors, coordinators are university diploma or equivalent in relevant field of studies. Their distribution is based on area they are assigned for. In fact there is a plan to have PDA to collect data from fields.

It has been noted that price survey is using markets as an EA.

Table 16: Retail prices enumeration areas

Region	**Branch office**	**Zone**	**Wereda**	**City**
Tigray	2	4	8	8
Afar	1	2	4	4
Amhara	4	10	20	20
Beshangul	1	2	2	2
Oromia	7	13	22	22
Southern People nationalities and nations	5	15	25	25
Somalia	1	1	4	4
Dire dawa	1	1	3	3
Harari	1	1	1	2
Addis Ababa	1		12	12
Gambela	1	3	3	3
Total =	**25**		**119**	**119**

The retail prices EA are 119 all over the country. They are different from other surveys enumeration areas. The nature of data collection and nature of data is different from other surveys.

Collected price data was sent to the HQ by using disks but because of high staff turnover most of them have stopped sending data through disks. Almost all of the branch offices are sending data in a hard copy.

Main responsibilities of the branch offices are:

- Collection of data
- Supervision of data collections
- Controlling quality of data and data collection processes
- Distributing report as assigned or instructed by the HQ

The personnel setups as related to survey undertakings are revealed in the following flow chart. In this chart role of SMS is not shown. This is because they are getting involved in the survey inception to its data dissemination at various degrees.

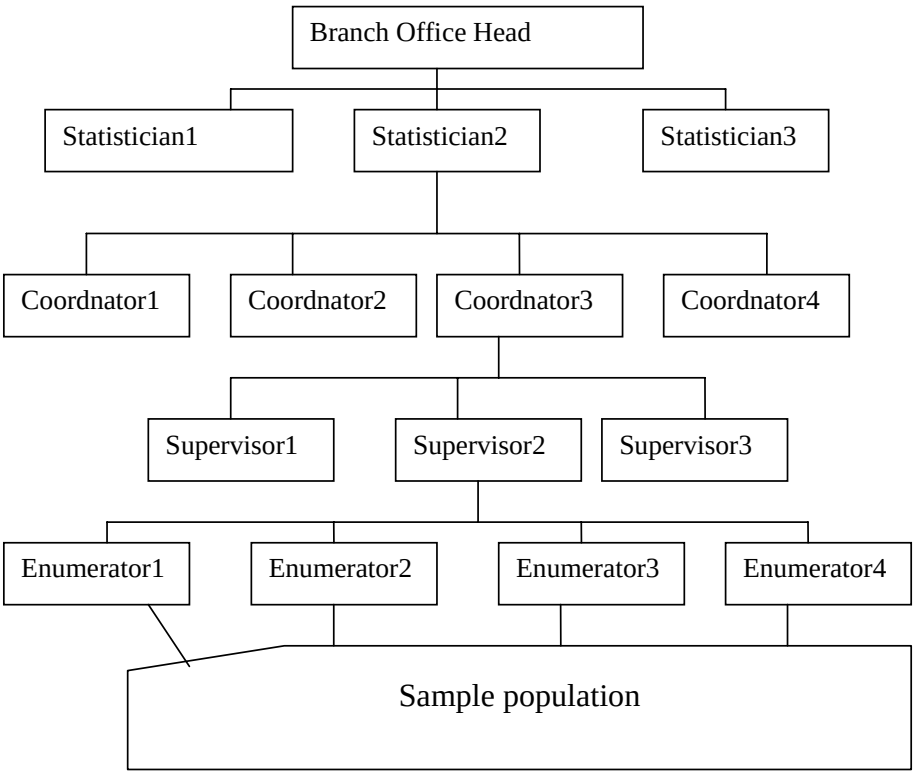


Figure 7: Institutional setup at the branch office level

Branch office is headed by a person who is BSC/BA holder in the field of geography, statistics, or economics. This person is responsible for all activities of the branch office.

On the next level of the setup there are statisticians. The number of statisticians in each of the branch office ranges from three to four on an average based on size of enumeration areas. If the number of EA is thirty, two statisticians are allocated, and if the number of EA is greater than 30, three to four statisticians are allocated based on the agro ecological system of a given area. For example, Bahir-dar has 170 EA, and there are four statisticians.

The statisticians are usually getting training on a given survey at the head office by the subject mater specialists. Then after, the statisticians are responsible to provide training to enumerators, supervisors, and coordinators. The training is same for all category of staff members so that they can speak same language. There is written exam at the end of the trainings to make sure that all participants can undertake their activities accordingly.

Secondly, the statisticians are undertaking spot checking.

Thirdly, they are controlling coordinators and supervisors for the purpose of quality control. If time allows them, they are swapping questionnaires do cross checking.

The third level in the branch office is Coordinators. Each of the coordinator is responsible to coordinate seven to eight supervisors.

At the fourth level of the branch office we can find supervisors. Each supervisor is responsible to supervise 3 to 6 enumerators. Supervisors are also getting involved in data collection related activities such as paying salary to enumerators among other activities.

The enumerators are temporarily employees of CSA.

The size of enumeration area is determined by house holds and is set CSA.

Reporting is being done as shown in the above figure. They are reporting, to the head of the Regional Offices and Field Operation Department and vise versa for instructions or assignments.

Usually the time of survey undertakings are determined or planned in collaboration with the subject matter specialists and data processing departments. Most of the time, the planned time is being kept as planned. But if the survey time is at pick time for house holds, delay can occur. The planned

time of series of each of survey of a given subject is uniform. The timing of one survey can differ from the other based on complexity and coverage.

The Department is using training manual and a working manual as a standard for data collection and keeping its quality. In accomplishing data collection the branch offices are using a number of forms and formats for taking tools such as compasses, meter, calculator, etc.

It has also been identified that there is horizontal communication and linkage among branch offices. For example, Gondor branch office is sending its data through Bahir-dar and Shire branch office is doing same through Mekele branch office.

6.4 Subject Matter Specialists (SMS)

In addition to the aforementioned departments Subject matter departments are also main stakeholders in the survey information management.

All the annual, periodic socio-economic surveys and censuses have been conducted by the subject-matter departments; namely: Natural Resources and Agricultural Statistics, Industry, Trade, Finance, Transport and Communication Statistics, Household Budget and prices Statistics and Social Statistics.

The SMSs are originators of all tabulations plan, edit specifications and produce statistical and analytical reports. Data analysis process has been initiated at the level of questionnaire design. SMS and Data processing are designing the questionnaires all together so as to make life easier for future analysis. In addition SMSs are developing all plans for tabulations. Then, programmers of Data Processing Department are making use of it and generate requested tabular data.

Generally stating SMS are the major once to generate any kind of reports or output. However Data processing Department is the one who is supporting output generations out of collected survey and census data.

Grass root for all output is users' requirement for datasets. Thus, they are major determinant of the analysis formats and output.

Whoever is initiating a given survey the respective department is designing a survey questions and getting feedback from user through a User Producer form.

The subject matter specialists are generally involved in a survey from inception to dissemination at different degree. For example, they fully involved in setting objectives, questionnaire preparations, data collection, manual editing, prepare tabulation plan for programmers, writing statistical reports, etc. The departments are generating analytical report rarely based on requests. These requests are coming from external agencies or institutions.

The skills of enumerators are set by SMS by considering the available skill and number. For example, for the large and medium scale survey all the data editing and coding were done by the SMS themselves.

The SMS are the once who are determining the frequency of a given survey. It has also been revealed that the frequency of given survey is based on the nature of the survey. Since there are some indicators which are expected to be answered by using yearly data (e.g. financial statements), they have to collect such data yearly. For instance area production and yield survey, land utilization survey, farm management and practice survey are expected to collect data only at the time production. In addition availability of resources also determines its frequency. Respondents fatigue is also another determining factor on determining the survey frequency.

7. Data collection

There is almost no difference in undertaking the following steps (table 13) in each of the 11 surveys described in chapter 3 of the report. However, the data editing and coding were done by subject matter specialist for the industry survey for its own reasons.

In fact, the current census has also different approach as compared to pervious censuses. The census of 1999 E.C (2007) data capturing and entering has been done by CSA by using a software which has been supplied by the UK based company called DRS. The system uses direct paper scanning and entering into the computer system. It does not need keying data from a keyboard. The correction and validation functions were accomplished by using the software. Once the Data processing Department finished all these activities, they exported the data to CSPro for further analysis.

Table 13: Major activities involved in survey administrations with their respective actors

Major activity	**Actor/s**
1. Plan Questionnaire	CSA, NGO, Public institutions
2. Questionnaire preparation/design	Subject matter specialist, Data processing department
3. Questionnaire dispatch	Field operation department, Subject matter specialist

4. Questionnaire filling	Enumerators, technical assistance from Subject matter specialist
5. Questionnaire retrieval/collection	Field operation department
6. Questionnaire storage	Data processing department through the Editing, coding and documentation team
7. Editing and coding	Data processing department through the Editing, coding and documentation team and Subject matter specialist
8. Data entry	Keying by Data Processing Department through data entry, cleaning and computer operation and checked by Subject matter specialist
9. Data cleaning	Data Processing Department through data entry, cleaning and computer operation and Subject matter specialist
10. Reformatting data	Data Processing Department through Systems and programming team
11. Report generation(tabulation)	Data Processing Department through Systems and programming team
12. Data preparation/conversion	Data Processing Department through Systems and programming team
13. System documentation	Data Processing Department through Systems and programming team
14. System backup	Prepared by Data Processing and sent to ICT department
15. Preparing documents in appropriate format	ICT
16. Information dissemination	ICT

In terms of the above major steps there is no difference within a given survey from year to year or from one dataset to another.

The timing/duration of the above major activity is known before hand. All surveys have plans. The duration or timing of the above activities have been done by all stakeholders in the survey undertaking. The stakeholders are subject matter specialist, Data processing department, and field operation department. The duration of a survey varies from one survey to another survey and it is based on its complexity and coverage.

Data collection is performed by following the steps mentioned below.

- Data collection manuals are prepared
- Appropriate trainings are provided to enumerators, supervisors, coordinators

- Ready and blank questionnaires are distributed to branch offices to be filled at sampled or required places
- Filled-in questionnaires are retrieved from Statistical Branch Offices. Papers are transported from branch offices to the Head quarter for the data entry purposes
- Documents are received, systematically registered and organized at the documentation unit of the Data processing Department.

8. Data processing and analysis

Before getting to details of this section we need to have clear meaning of two terms/ phrases statistical reports and analytical reports. Statistical reports are reports which are providing summarized data in tabular formats. Where as analytical reports are providing more analysis over the tabular data such as why it happened and etc. It is also possible to say statistical reports are first level or tier of analytical reports.

CSA has no analytical needs for itself, but it has a plan to generate analytical reports in the near future, if it has been requested for. CSA needs to analyze some data on a time series basis and provide information to the users. For example, users are expecting to get price data of certain commodities over range of times. Production data over certain time is also expected to be obtained at finger tips.

Livestock number and breeds with their distributions are also some important information which users are expecting to get over certain range of years.

Demographic and health situation over range of time and over certain geographical area is also one of the agenda to be addressed by CSA datasets. Specifically it needs to assess the working condition of children, identify who is most use child labour, assess welfare of working children, etc.

Land use information over time under certain region, volume of crop production, the corresponding input used, etc. are some of the queries which are expected to be answered.

In order to get existing poverty situation, CSA needs to identify poor and vulnerable groups over different regions and over different time.

Change of Industrial production, use of raw materials, identifications of problems in the industry sector over time and area are also some of the points to be answered on the finger tips.

Getting information on household consumption varies over time and space is also another important analytical data which will be used by the government for policy analysis and evaluation.

Furthermore, it needs to show distribution of unemployed or employed and to identify workforce over time and space.

Generally stating, CSA needs to generate time series data to trace changes over time and space. This is believed that the government can make use of it so as to make an intervention be it at policy level or other applicable means.

The SMSs are fully responsible to prepare and provide tabulation plan to the Data Processing Department based on set indicators in a given survey. The number of set of tables depends on the amount of used indicators and complexity of the survey in hand.

Then, the Data processing Department is developing computer program which can help them to generate the set tables. Then after all required tables are generated accordingly and given to respective SMS.

It has also been identified that Branch Offices are not involved directly in data analysis activities except data collection and report distributions.

a. Editing and Coding

- Instruction manuals used in editing and coding are prepared by subject-matter specialists.
- Editors are given training on editing and coding before the actual work begins.
- Filled-in forms are checked for consistency and completeness; inconsistent data are corrected.
- Verification is done to check the editors' work and ensure data quality (100% verification).
- Different forms are used to control the flow of questionnaires during the editing process (handover, returning, misplacement, loss…)

b. Data Entry

- Data entry applications needed to capture the data for the surveys are currently prepared using CSPro software
- It combines the features of IMPS and ISSA (Integrated Systems for Survey Analysis) software packages.
- CSEntry, a module of CSPro, is used to run the data entry application.
- Features of CSEntry
  - Questionnaire-oriented screens
  - Automatic range checks
  - Interactive skips
  - Modification
  - Verification

- Performance statistics
- ASCII format output data files

- Keying in data is carried out by data entry operators.
- Verification (re-entry) is done to check the entry work and improve the quality of data.
- Verification is carried out on 100% basis.
- At the end of data entry of an Enumeration Area (EA), structure edits are run to check the completeness of questionnaires for each case.

c. Data Cleaning

- The cleaning and computer editing for a survey data is done using CSPro Batch Edit Application
- A batch Edit Application contains edits (logic) that are used to:-
  - Validate individual data items
  - Test consistency between data items
  - Check questionnaire structure (completeness)
  - Generate edit reports used in cleaning the data
- CONCOR, a subsystem of IMPS, is sometimes used in computer editing
- Computer edit programs are prepared in accordance with the edit specifications provided by the subject matter specialists.
- Missing data is checked.
- Errors and inconsistencies in the data are checked and corrected.

Corrections are done through references made to the original documents, using printouts from the edit program.

d. Reformatting the Data

- The cleaned data is reformatted to facilitate easier report generation or tabulation.
- During reformatting
  - Data files may be re-organized
  - New variables can be created whenever required for analysis

e. Tabulation

- Tables are prepared by the subject-matter specialists and handed to computer programmers
- Computer programs used it to produce statistical tables.
- Tabulation module of CSPro is used to generate required tables
- Statistical tables generated are exported to MS-Word, MS-Excel, or MS-Word format for publication purposes.

f. Report generation

The generated tables will be checked for quality by SMS and Methodology Department. If it passed the quality control, then the respective SMS will write reports. The report can be analytical or statistical in nature.

8.1 Major features and purposes of datasets analysis systems

CSA is using a number of softwares for dataset analysis. The Major once are: Integrated Microcomputer Processing System (IMPS), Census and Survey Processing System (CSPro), MS-Excel, and SPSS. All line departments of CSA are using SPSS and MS-Excel on their own office machines for data analysis purpose.

8.1.1 CSPro

CSPro (Census and Survey Processing System) is a public-domain software package for entering, editing, tabulating and mapping census and survey data.

CSPro was designed and implemented through a joint effort among the developers of IMPS and ISSA: the United States Census Bureau.

It is the combined IMPS and ISSA Census and Survey Processing Systems which has been developed and used by US census bureau.

Census and Survey Processing System (CSPro) Software is to enter and verify on personal computers. CSPro is the most important software which is being used in CSA.

CSPro is free software which was obtained from US Bureau of Cencus. CSPro is a software tool for entering, editing, tabulating, managing and disseminating data from surveys and censuses.

This software has a number of modules which can be used for various activities. CSPro software has the following major modules:

- Data Entry Application
- Batch Edit Application
- Cross Tabulation Application
- Tools

CSEntry, a module of CSPro, is used to run the data entry application. Main features of CSEntry

- Questionnaire-oriented screens
- Automatic range checks
- Interactive skips
- Modification
- Verification
- Performance statistics
- ASCII format output data files

A batch Edit Application contains edits (logic) that is used to:-

- Validate individual data items
- Test consistency between data items
- Check questionnaire structure (completeness)
- Generate edit reports used in cleaning the data

Tabulation module of CSPro is used to generate required tables by Subject Matter Specialists (SMS). Statistical tables generated are converted to MS-Word, or Excel format for publication purposes and to SPSS format for further analysis.

8.1.2 SPSS (Statistical Package for Social Science)

SPSS has the features that allow efficient delivery of the following statistical data preparation, presentation, analysis and decision making.

1. Data entry, editing, importing and exporting, recoding and rearranging interfaces
2. features for Descriptive Statistics
    a. Frequency distribution
    b. Measures of central tendency (mean median, mode, etc...)
    c. Measures of dispersion (range, variance, standard deviation, etc...)
3. Graphical presentations
4. Reporting and printing features
5. Features for Inferential Statistics
    a. Comparing averages: Two-Sample and One-Sample Tests.
    b. Chi-Square test
    c. Measuring statistical association(Correlation test)
    d. Analysis of Variance-ANOVA (One-way, Two-way, etc…)
    e. Covariance Analysis
    f. Regression Analysis
    g. Discriminant Analysis
    h. Logistic Regression and
    i. Others

SPSS evolved in sequences of versions, and recently SPSS 16.0 is released. This package is used in market effectiveness, fraud detection, risk management, market research, academic activities, scientific research, business intelligence, data and text mining, predictive analysis, survey research, and others.

SPSS present and share dynamic, interactive results when you use the correspondence analysis and categorical regression analysis procedures, you can display your results in tables, graphs or report cubes that feature

unique pivoting technology. This statistical technology empowers you to discover new insights into your data — with a few mouse clicks. You can swap rows, columns and layers of report cubes or quickly change information and statistics in graphs for new levels of presentations.

SPSS can get data from MS-Excel, CSPro, etc. for further analysis. It has full support from its developers or vendors.

This software is one of the best tools to undertake statistical analytical activities. Users are expected to understand how it is working and how to use its advanced features which can bring in a lot of powerful features.

8.1.3 MS-Excel

MS-Excel (full name Microsoft Office Excel) is a spreadsheet software. The primary advantage of a computerized spreadsheet is its ability to redo the calculations should the data it stores be changed. Calculations can be made automatically as formulas have been preset into the spreadsheet.

Microsoft Excel is a proprietary spreadsheet application written and distributed by Microsoft for Microsoft Windows and Mac OS X. It features calculation, graphing tools, pivot tables and, except for Excel 2008 for Mac OS X, a macro programming language called VBA (Visual Basic for Applications). It is overwhelmingly the dominant spreadsheet application available for these platforms, and is bundled as part of Microsoft Office.

This software is the richest in exchanging data with a number of other systems such as Databases, spreadsheets, statistical softwares, word processing, information presentation tools, etc.

SMSs are using this software mostly to generate graphs from the tables which were generated by Data Processing Department. In addition they are using it to do some calculations on the tables in hand.

8.1.4 Integrated Microcomputer Processing System (IMPS)

The Integrated Microcomputer Processing System complemented by CSPro software was used for data entry, consistency checks and tabulation of survey results

The Programmers prepared the data entry programs using CENTRY, which is a data entry module of IMPS.

CSPro combines the features of IMPS and ISSA (Integrated Systems for Survey Analysis) software packages.

CONCOR, a subsystem of IMPS, is sometimes used in computer editing.

Please note that except few of its modules, IMPS is obsolete software which is being replaced by CSPro software. But some of its modules are still being used for data analysis purpose especially the CENTS, the tabulation component of IMPS.

Main features IMPS

The IMPS performs the major tasks in survey and census data processing: data entry, data editing, tabulation, data dissemination, statistical analysis and data capture control which can be used as a complete processing system or as stand-alone modules.

IMPS 3.1: includes the following software modules:

- **Data Dictionary** - This module allows for the definition of the characteristics of the data file(s) to be processed. This definition, which is stored in computer, is used by the data entry, editing, and tabulation components of IMPS to access data files.
- **Data Entry** - CENTRY is a screen-oriented, menu-driven package for developing data entry applications. Specifically it can be used for entering, editing, verifying and modifying data, and for collecting statistics on data entry operator performance. CENTRY executor can run on a 2-diskette drive system with no hard disk.
- **Editing** - CONCOR is a package for the rapid identification and correction of invalid and inconsistent data using any editing techniques which are considered appropriate. CONCOR can be used independently or in conjunction with CENTRY. It requires CA-Realia COBOL.
- **Tabulation** - QUICKTAB is a menu-driven package for the rapid production of frequency distributions and cross- tabulations.
- **CENTS** is a package for tabulating, summarizing, and displaying statistical tables for publication. It requires CA-Realia COBOL.
- **Data Dissemination** - Table Retrieval System (TRS) is a menu-driven package to easily select, retrieve, display, and print statistical tables. It is a tool for electronic data dissemination.
- **Variance Calculation** - CENVAR is a variance calculation package which produces reliability measures for estimates from stratified multistage sample surveys or simpler survey designs.
- **Operational Control** - CENTRACK is a management and control package to help census managers monitor, control, and track the various operations necessary between receipt of questionnaires from the field and data entry.

8.1.5 STATA

Among other features the latest Sata has the following major features. Stata is a complete, integrated statistical package that provides everything you need for data analysis, data management, and graphics. Stata 10 adds many new features such as multilevel mixed models, exact logistic regression, multiple correspondence analysis, a graph editor, and time-and-date variables.

Stata puts hundreds of statistical tools at the fingertips, from advanced techniques, such as survival models with frailty, dynamic panel data (DPD) regressions, generalized estimating equations (GEE), multilevel mixed models, models with sample selection, ARCH, and estimation with complex survey samples; to standard methods, such as linear and generalized linear models (GLM), regressions with count or binary outcomes, ANOVA/MANOVA, ARIMA, cluster analysis, standardization of rates, case–control analysis, and basic tabulations and summary statistics.

Stata's data-management commands gives complete control of all types of data: you can combine and reshape datasets, manage variables, and collect statistics across groups or replicates. You can work with byte, integer, long, float, double, and string variables. Stata also has advanced tools for managing specialized data such as survival/duration data, time-series data, panel/longitudinal data, categorical data, and survey data.
Publication-quality graphics: Stata makes it easy to generate publication-quality, distinctly styled graphs, including regression fit graphs, distributional plots, time-series graphs, and survival plots. With the integrated Graph Editor you click to change anything about your graph or to add titles, notes, lines, arrows, and text.

The software is one of the licensed systems in CSA. Some members of Natural resources and Agricultural statistics Department had taken training courses on the software. However, still it is not as popular as other analytical systems and not being used as required.

8.1.6 Other analytical systems

Among the softwares used in CSA in different departments one is called EUROTRACE. The Industry, Trade and Services Department is sharing trade data which they are getting from Customs Agency from ASICUDA. They export this data into software called EUROTRACE DBMS which they obtain through COMESA including its technical support. EUROTRACER software is being funded by EU. The data is given to anybody who is requesting for. Currently, they are organizing data on a yearly basis.

The Manpower and Social Statistics Department is using software called SPECTRUM/DM project for health and population related analysis and sharing data with others.

SAS software which provides extensive statistical capabilities is also being used in CSA.

8.2 Output of the data analysis system

The existing data analyses output is mainly statistical reports which do have charts, tables, etc, in their bodies. Maps and GIS are also on a process to be incorporated in most of the outputs. Website and CDs are major media for information disseminations.

SMSs are responsible to generate statistical reports for all surveys. The SMS departments are generating analytical report rarely based on requests. These requests are coming from external agencies or institutions. So far, analytical reports were generated for Ethiopia Welfare Monitoring; Household Income Consumption and Expenditure Surveys; Demographic Health survey of 2005; and labour force survey.

In addition in depth analysis were done for Population & housing and Agricultural censuses.

There is a plan to improve their data analyses procedures and outputs based on users' feedback. They are ready to add new indicator, to remove unwanted ones, to add frequency, and even to add new survey.

9. Data storage and sharing

Data is stored on servers and tapes for backup purpose and future use. Electronic datasets are stored for some years in flat files formats in both ASCII and SPSS formats. Copies of some datasets are being stored in individuals PC.

It was identified that electronic information or data sharing is being performed by the use of secured shared folder over the current LAN of the institute. If the network is not operating because of different reasons, they are using floppy disks and flash disks to share electronic files. The shared files can be datasets, tables and reports.

It was also identified that the Industry, Trade and Services Department is storing the database in the department which is obtained from Ethiopian Customs Authority so as to share the information with other users of the data.

The catalog entry of the national archive data provides users with the facility to search at the variable level by entering a key word as it may appear in the variable and a list of variables containing the word is provided across all surveys.

This creates additional access point to the existing information by reorganizing some of the associated documents.

Ethioinfo and price database are additional archiving and sharing tools of datasets for target audiences. Ethioinfo is used to share indicators related to Human Development, and to facilitate data sharing and indicator harmonization at global, regional and country level by making statistics available to a wide audience. It allowed presentation of data through Tables, Graphs and Maps.

Price database has been initiated to provide price information over time series and over space. This system is under development during this report writing.

9.1 Backup System

The issue of data security is important, but unfortunately the issue of backup and recovery are getting overlooked in a number of projects. In a lot of cases there is little guidance or support for backup and recovery of critical data - the data CSA provides to its targeted clients.

Available microdata files are in ASCII format, along with its data dictionary, and the data file in SPSS format. Where are the backups located?

- One copy is being kept in Data Processing Department
- Another complete copy of the data and system is being handled by the ICT department for Central Data Bank for achieving and dissemination purposes.
- There is a backup on high density magnetic tapes; copies are available in the DDG's office in a different campus. This is encouraging.

But, is it still sufficient to have them at where they are?

10. Reporting and dissemination

10.1 Existing reporting and dissemination systems

Existing reporting and dissemination software systems in CSA include MS-Word which is almost used by all staff members for various purposes; IHSN which exclusively used by ICT Development Department; MS-Excel which also used by most of the staff members for various purposes; Ethio-info; SPSS and Adobe PDFMaker are mostly used by the ICT Department for information publishing on the website and CD-ROMs.

Information sharing and dissemination is based on data access policy which was approved by the Ethiopian Ministers of Councils.

It was identified that the Agency is widely using facsimile technology in communicating with its branch offices. And it was identified that Email technology

is not widely used in the Agency for official communications where as they do have Micro-soft Exchange mail server.

During the data collection phase of this project, we came to learn that FTP is not being used because there is no WAN implemented at the time of writing this report. In fact most of the staff members are sharing reports and data by using secured shared folders.

The following section will cover major features of the software in use and for what purpose CSA is using them.

10.1.1 MS-Word

A word processor such as MS-Word is a computer program or software that enables users to create, edit, print and save documents (or textual files) for future retrieval and revision. Users enter text into the computer by using keyboard, which is displayed on the monitor. A key advantage of word processing software is that users can make changes such as spelling, margins, additions, deletions, movement of text, etc.

CSA is using Microsoft Word that runs on the ***Windows*** operating system. For instance, MS-Word lets users to combine text that has been formatted in a variety of styles with graphics and can include tables and data from other software such as spreadsheets (e.g. MS-Excel), SPSS, databases and graphics programs. The users are importing pictures (graphs) anywhere in a page in different sizes.

Generally stating, CSA is using this software so as to prepare and produce reports on all surveys. Then, the reports could be printed and same copy will be forwarded for electronic report disseminations.

It has also been observed that everybody is comfortable in using MS-Word in CSA.

10.1.2 IHSN Microdata Management Toolkit

The IHSN Microdata Management Toolkit developed by the World Bank Data Group for the International Household Survey Network aims to promote the adoption of international standards and best practices for microdata documentation, dissemination and preservation.

CSA is using all three modules of the Toolkit. They are using the Metadata Editor to document data in accordance with international metadata standards (DDI and Dublin Core). The Data Documentation Initiative (DDI) is an international project

to create a standard for information describing social science data. The DDI specification, written in XML, provides a format for content, exchange, and preservation of information. Version 3 of the DDI standard is expected to be released in 2008.

DDI 3.0 is used to Support Preservation, Management, Access and Dissemination Systems for Social Science Data.

The DDI-tree contains five main branches, or sections:

1. **The Document description**, which consists of bibliographic information describing the metadata document and the sources that have been used to create it
2. **The study description**, which contains information about the data collection.
3. **The Data files description**, which describes each single file of a data collection (formats, dimensions, processing information, missing data information etc.)
4. **The variable description,** which describe each single variable in a datafile (format, variable and value labels, definitions, question texts, imputations etc.).
5. **Other Study-Related materials**, which can include references to reports and publications, other machine-readable documentation

CSA is taking the advantage of DDI to classify, describe, and organize datasets of its surveys.

The Dublin Core metadata standard is a widely recognized meta-language to describe information resources. It contains fifteen elements such as coverage, creator, date, description, format, etc.

The Explorer, free reader for files generated by the Metadata Editor is also in use. The module allows users to view the metadata and to export the data into various common formats (Stata, SPSS, etc).

CSA through its ICT department is widely and frequently using its CD-ROM Builder to generate user-friendly outputs (CD-ROM, website) for dissemination and archiving.

CSA is usually using this toolkit to publish CD-ROMs and website on each of the datasets.

Some of the users are claiming that they are using Nesstar and some others are claiming that they are using IHSN Microdata Management Toolkit. So, is there any difference between the two systems?

Nesstar is a commercial software system for data publishing and online analysis. The software consists of tools which enables data providers to disseminate their data on the Web. Nesstar handles survey data and multidimensional tables as well as text resources. Users can search, browse and analyse the data online.

Nesstar Publisher's feature is being integrated with IHSN. This component consists of data and metadata conversion and editing tools, enabling the user to prepare these materials for publication to a Nesstar Server.

The Nesstar Server is built as an extension to a normal web server. As well as providing all the usual facilities for publishing web content, this server provides the ability to publish statistical information that can be searched, browsed, analysed and downloaded by users. This is done either by using a standard web browser or using Nesstar WebView.

The Nesstar represents a system of software architecture that makes it easy to create, locate, access and operate remotely on metadata and corresponding data. At the same time it does this while maintaining a high level of compatibility with the WWW.

Nesstar has the following analytical tools which are not available in IHSN:

- Cross tabulations
- Correlations
- Regressions
- Compute and recode
- Graphical representations of data in customizable forms
- Application of variable weights

In fact, IHSN toolkit has taken a lot of features from Nesstar. In fact, IHSN is using a lot of components from Nesstar as they are.

In reporting and dissemination of reports and datasets archiving, CSA is using IHSN. The staff members of Dissemination team are using IHSN to produce a single output particularly for CD-ROM version, then they are publishing same version to the website without optimizing it for web.

The ICT department is burning CD-ROMs for each of the surveys based on the number of target audiences of each of the surveys. Whenever more or extra copies are requested, they are producing requested CD-ROMs copies and sent them to public relations for distributions.

10.1.3 MS-Excel

The primary advantage of a computerized spreadsheet is its ability to redo the calculations should the data it stores be changed. Calculations can be made automatically as formulas have been preset into the spreadsheet.

This software is the richest in exchanging data with a number of other systems such as Databases, spreadsheets, statistical softwares, word processing, information presentation tools, etc.

Almost all staff members are using Microsoft Excel for various purposes such as doing some calculations, generating graphs, etc and sharing data with others.

10.1.4. Ethio-info/DevInfo

EthioInfo is in use in CSA at http://www.csa.gov.et/di5web/ . They are using it as a common platform for indicators related to Human Development, to facilitate data sharing and indicator harmonization at global, regional and country level by making statistics available to a wide audience. It allows presentation of data through Tables, Graphs and Maps.

End users can get screens whereby they can enter some parameters for searching and information presentations. For example for the industry database one can use http://www.csa.gov.et/di5web/devinfoapp.aspx

The site is using ACCESS Databases at the backend.ASP.NET programming language for the site at the front-end.

EthioInfo V 2.1 is the up-to-date version and contains the latest Ethiopia Demographic and Health Survey 2005 and also includes the following surveys;

- Welfare Monitoring Survey (WMS), 1996, 1998, 2000 & 2004
- Total Population for 2004 & 2005 (Population size, Area, & Density)
- Area and Production of temporary crops for 2004 & 2005
- Ethiopia Demographic and Health Survey 2000
- Household Income Consumption and Expenditure Survey (HICE), 1996, 2000

It has been found out that EthioInfo is a customized adaptation of DevInfo, a world wide used user friendly software that helps to organize, present data in a result based environment with unique features linking to strategic monitoring and evaluation of policies such as MDG, National Poverty Reduction Strategies.

DevInfo is a powerful database system that is used to compile and disseminate data on human development. The software package has evolved from a decade of innovations in database systems that support informed decision making and promote the use of data to advocate for human development.

DevInfo was developed by UNICEF in cooperation with the UN System to assist the UN and Member States in tracking progress toward the Millennium Development Goals (MDGs). In 2002, DevInfo was proposed as a standard software package for the whole UN System. Its specific purpose is to store existing data, identify gaps in the MDG indicators, provide a single entry point for data on the MDG indicators, and disseminate information simply and attractively.

DevInfo is claiming that it is an integrated desktop and web-enabled tool that supports both standard and user-defined indicators. The standard set of MDG indicators is at the core of the DevInfo package. In addition, at the regional and country levels, database administrators have the option to add local indicators to their databases. The software supports an unlimited number of levels of geographical coverage: from global level to regional, sub-regional, national and sub-national down to sub-district and village levels (including schools, health centers, water points).

Data from DevInfo can be exported to XLS, HTML, PDF, CSV and XML files and imported from spreadsheets in a standardized format. DevInfo also has a data exchange module for importing data from industry-standard statistics software packages such as SPSS, SAS, Stata, Redatam, and CSPro.

DevInfo is distributed royalty-free to all Member States and UN agencies for deployment on both desktops and the web. The user interface of the system and the contents of the databases it supports include country-specific branding and packaging options which have been designed to ensure broad ownership by national authorities. UNICEF has absolutely no restrictions on the database and its use.

The most common DevInfo users include UN country teams, national statistical offices, planning ministries, and district planners. Frequent users also comprise members of the media (for reporting and tracking human development data), educational institutions (for analyzing data and helping students gain data access), as well as DevInfo administrators (in particular for customizing the system or adding data through advanced database administration modules).

10.1.5 Adobe PDFMaker

CSA has adopted PDF to streamline document management and reduce reliance on paper. They are using it as the standard format for the electronic document management and dissemination of output of all surveys.

The PDF formats are generated by ICT development department. Reports and questionnaires are available in PDF format for surveys over Internet.

This format is used to keep file format by preserving the fonts, images, and layout of source documents created on a wide range of applications and platforms. PDF is the standard for the secure, reliable distribution and exchange of electronic documents and forms. Adobe PDF files are compact and complete, and can be shared, viewed, and printed by anyone with free Adobe Reader software.

The PDF documents can be opened either in Acrobat or in a web browser. In Windows, users can configure their web browser to open PDF documents.

All PDF documents which were made available to users over the website do not have a feedback collection form which is an electronic-based document that can collect data from a user and then send that data via email or the web to CSA.

This implies that electronic document generating staffs are not making use of some important features of the dissemination system. For instance, all PDF documents do not have feedback forms in the body of the document.

10.1.6 SPSS

This software is mainly being used by ICT development department to disseminate microdata and to generate statistical tables whenever requested by clients. For details on the software refer to the above section.

10.2 Planned reporting and dissemination systems

What does CSA has in hand for the future? Currently, they do not have additional reporting and dissemination systems in plan but they are ready to improve the currently in use systems based on feedback from users. For example they planned to:

- improve the quality of the website
- make the website dynamic
- make some microdata online

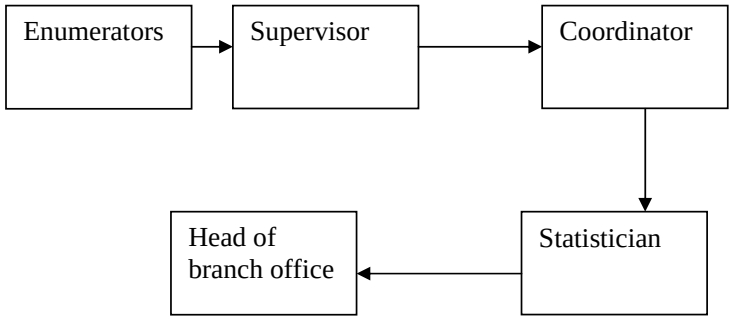
Whenever they are preparing reports they are using previously generated reports. This means for the regular reports they are strictly following what have been produced earlier so as to produce a report in hand. They are using previous report as a reference and as a guideline in order to generating the report in hand.

10.3 Existing reporting and dissemination procedures

10.3.1 Reporting at the Branch Office level

Reporting at the branch office level is from enumerator to supervisor; from supervisor to coordinator; and from coordinator to statisticians. Figure 1 depicts how reporting at branch office is being performed.

Figure 1: reporting at branch office



Enumerators are collecting data as per set guidelines and passover the filled in questionnaires to their respective supervisors.

The supervisors are collecting filled in questionnaires from enumerators for which they were assigned and send them to branch office or coordinators.

All completed survey questionnaires are sent to the HQ formally after checking for quality as discussed in report 1. They are sending hard copies by describing encountered problems and pending issues within a page of a covering letter. Generally reporting is performed as shown in figure 7 of report 1.

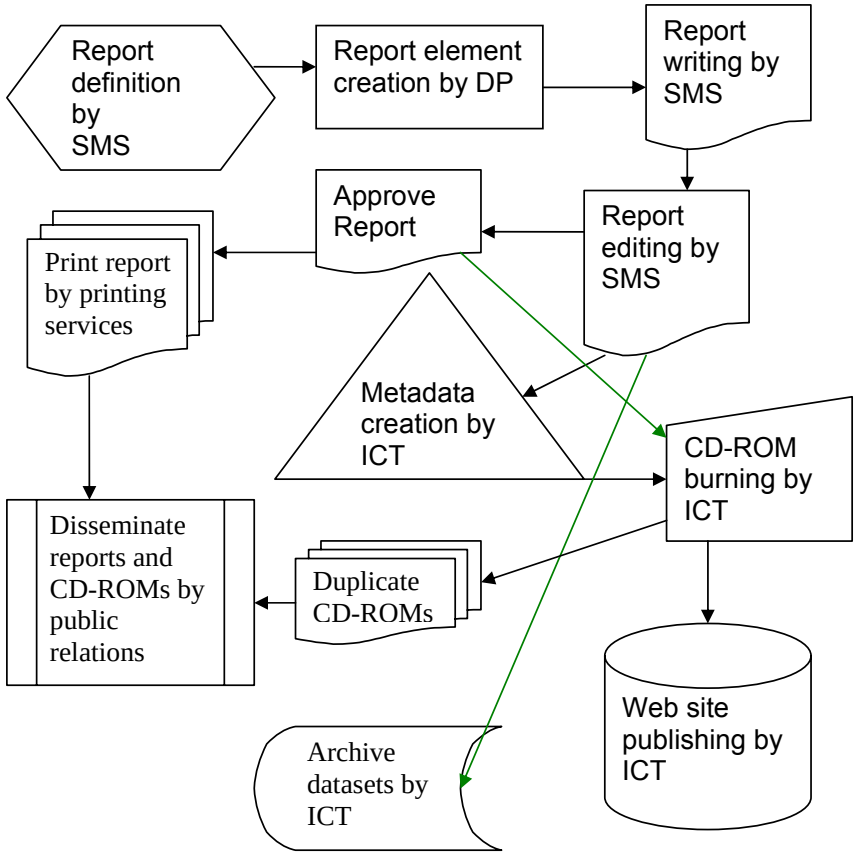
Price data is sent to the HQ by using disks but because of high staff turnover most of them have stopped sending data through disks. Almost all of the branch offices are sending price data in a hard copy.

10.3.2 Reporting at the Head Quarter level

Generally reports are initiated, prepared and disseminated at the head office. Since the target of a survey is to generate report, the report generation process

involves a lot of staff members from Agency directly and indirectly. An overview of the process is given in figure 2.

Figure 2: Report generation and dissemination process in CSA



The SMS are responsible to generate reports for their respective assignments at the Head quarter level for CSA. For example, the Natural Resources and Agricultural Statistics Department has produced analytical report only for the agricultural census to satisfy analytical needs of its clients. All other reports are mostly focusing on statistical reports with minor summary of reports.

The Department's personnel involved in a survey preparations to generating different types of reports at different degrees. It is critically involved in quality control of data and report preparations in collaboration with Data Processing Department.

Once a given report has been completed it will be sent to the management for approval and comment. After getting feedback from management it will be edited, finalized and sent away for sign-off.

10.3.3 Level of Reporting

Data items that are to be collected must be extrapolated from samples which take into account lower administrative units. As depicted in reports, most data are available at the country and regional levels and some are at Zone levels. Basically, the bottom-up development approach to which Ethiopia adheres to calls for the availability of data at grass roots (for the time being zonal level) level that can be channeled to the higher level is very important. This dissemination could be enhanced and put into practice gradually through RDBMS so as to get data at the zonal administrative levels in various formats.

10.3.4 Time of reporting

It has been identified that there were a lot of improvements in data collection, processing and reporting from time to time. Still it is important to devise mechanisms to further narrow down the period between data gathering, processing, reporting and dissemination of the results to users.

The output delivery time is not fast enough, for example, they need to generate production forecast report within less than a month, in the month of December because the government needs the information as early as possible. Therefore the government is insisting to get it as early as possible.

Average retail prices, CPI, producer price reports are generated monthly by the Household budget, welfare monitoring and price statistics Department. Except for the producers price report all are being generated on time as expected by users.

Except for the Production Forecast Survey all of the Agricultural sample surveys are expected to be processed at the head quarter within a time range of three months. The forecast survey is usually done within one and half months. Generally stating at the field level ¼ of the total time is required while ¾ of the total time is consumed at the head office.

The regular Large and Medium Manufacturing Industries Survey takes 30 to 45 days at field levels; seven to ten days at the branch level offices and five to six months at the head office.

Ethiopia Welfare Monitoring Survey needs about 10 to 15 days for data collection. At the head quarter it needs about 12 months to reach on report generation.

Household Income Consumption and Expenditure Survey needs about two months for data collection from the field. At the head office it needs about eighteen to twenty four months to reach on report generations.

Urban annual Employment Unemployment Survey takes fifteen days at field level, one week at branch office, and three to four months at the head office.

10.3.5 Level of classification in reporting

Although CSA has made efforts to conform its classifications to the international recommendations to satisfy the need of data users, it would be more usable, if it can make the level reporting more flexible.

10.4 Information dissemination from CSA

As indicated in figure 2, reports are being generated based on set formats and subjects. The formats have been set by subject matter specialists in collaboration with the Data Processing Department. Then, the Data Processing Department is generating tables which will be further analyzed by subject matter specialists. The tables are briefly analyzed and interpreted and reports are prepared by subject matter specialists. It had also been noted that different tables (statistical data) can be generated from raw data whenever requests are come in.

It has been identified that the SMS do not have a formal training on report writing so as to generate more attractive reports.

Here it has been identified that there is a formal procedure which was set by management for handover activities/reports from one department to another for example, Data Processing to ICT department. However the departments have not applied the set procedures because of different reasons. Rather transfers are being done through informal communications. For the time being this can be efficient but it is not long lasting. Although informal communications among departments is the supported environment, CSA needs to enhance and implement formal guidelines where certain activities handed over to the next level of activities.

Timing of dissemination

Since clients were identified before report generations it takes insignificant time for dissemination of both hardcopy and softcopies. Distribution of reports on demand also does not take so much time.

ICT needs four to five days for each of the surveys to publish website and produce CD-ROMs, if they receive complete information from the Data

Processing Department. The time can be extended unexpectedly, if electronic copy is lost or incomplete.

CSA Website

The CSA website is one of the richest website of the Ethiopian Institutions in terms of its content. The website really describes all the major activities and operational structure of CSA to anyone who is not familiar with CSA. Therefore CSA should continue with having up-to-date and current information and more user friendly information architecture on the website.

The information architecture of the website is not user-friendly particularly the home page and second levels of the site. In addition the survey information has been dumped on the website without optimizing it for the web. For instance, in survey information the header reads “CD-ROM” which needs to be edited (e.g http://www.csa.gov.et/text_files/Agricultural_sample_survey_2006/survey0/index.html).

Basic formats of the electronic copies came from SMS and Data Processing Departments. ICT is generating CD-ROMs and publish website based on templates created through IHSN publishing software. These templates are standardized through the World Bank and are compatible to the world wide acceptable standard called DDI.

GIS in CSA

Currently, GIS Infrastructure is available. Digitization of maps and development of coordinate database development are under process so as to incorporate various survey datasets over them and prepare the national data for the decisions and policy makers and other users of the data.

Currently GIS is being applied on population census data of 2007 by using Arc/GIs 9.2. Currently they procured the latest version of Arc/GIS.

The GIS team leader has also been involved in the Land cover mapping project.

There are 3 staff members who are new to GIs and ready to commence working on GIS.

GIS maps have been commenced to be generated. The plan is there to apply same on all surveys wherever applicable.

10.5 Verifications and sign-off of reports and datasets

In general, the management of CSA is responsible for the general verifications and sign-off the reports and datasets.

In particular, the DDG of Operation, Methodology and Data Processing is signing-off all the electronically disseminated reports. There are two category of out put which are being disseminated. The first one is aggregate information and the second categories are micro-data. The microdata are being distributed based on Data Access Policy which was approved by the Council of ministers in 2004.

It has noted that the DDG of Social Statistics is responsible to sign-off all hard copies.

10.6 Personnel involved in Reporting and dissemination systems

All survey reports are produced by respective SMS while disseminations are being done by two departments. The hard copies and CD-ROMs are disseminated by Public relation department while website is published and handled by the ICT Development Department.

SMS professional staff members who have specialties in geography, statistics, sociology, demography, and related fields are participating in report generation within their respective surveys.

Generally stating the line departments (SMS) are using same professionals for questionnaire preparation, quality control and report generation.

The ICT Department has one team which is fully dedicated to deal with information disseminations. This team has currently four professionals who are involved in information disseminations. Two of them are graduates in Mathematics and two of them are IT graduates.

11. Standards

In this section of the report we will focus on national and international standards which are being used by CSA and institutional standards which are specific to CSA.

The Industry, Trade and Services Department is using international standards. They are using UNSD (United Nations Statistical Divisions) Classifications or recommendations by considering national needs and requirements for the standards. The trade classifications are using ISIC and HS. HS code is the Harmonized Commodity Description and Coding System which is being issued by the World Customs Organizations. To identify the appropriate HS code for many commodities, the department is referring to the "HS Commodity Data

Base" a CD-ROM product issued by the World Customs Organization. In addition ILO standards are being used accordingly for Informal sector survey which is ad-hoc survey. Some of the stake-holders or institutions are challenging not to accept the adopted standards. These are micro and small enterprises.

The Natural Resources and Agricultural Statistics Department has set standards for the surveys which are taking place at different times. The standards include use of concepts, definitions, timing of survey, major variables to be included in a survey, etc. For example, agricultural data is being collected yearly which is internationally set frequency. Although some countries are collecting twice a year, CSA is following FAO recommendations which are being widely used in this category of the survey. In fact, the department is adding some variables which are unique to Ethiopia.

Manpower and social statistics department is following UN recommendations for DHS; ILO standards for National Labour Force and Urban annual Employment Unemployment Surveys.

The Household budget, welfare monitoring and price statistics Department is following international standards from ILO, IMF, and ISIC.

Institutional standards rely on training manuals and working manuals as a standard for data collection and keeping its quality for all surveys.

Although they are following internationally set standards, particularly there are implementation differences of codes from implementing department to another. The agency lacks standards at the institution level.

12. Identified gaps

This section outlines the gaps identified on some of the components of major activities in addition to some technical identification which are mentioned in the body of the report.

Social and technical gaps
The Division of Operation, Methodology and Data Processing is facing technical problems at all levels in its four departments. The problems do have different degrees. Some are critical and some are minor. These problems are being manifested in the quality of data and services.

Social and cultural problems are being manifested at data collection levels from respondents. For example, household survey has faced such problems in urban areas. The urban householders are not happy to give real data because they suspect that the data may be used for other purposes which may hurt them economically.

Non-responsiveness of respondents is also another problem.

Human resources gaps
Both lack of skilled human resources and high staff turn over are being revealed in the Division in general and they are highly affecting the ICT development and Data processing Departments in particular. The DDG has tried a number of options to minimize these problems. For instance, it has tried to have better salary scale. And also tried to adopt what ICTDA has implemented. In both cases the division is not successful at a moment, because from the government side both options has not taken or accepted to solve the stated problems.

Furthermore, it has been identified that staff turn over in hardship areas is higher than other areas.

The new organizational structure is calling for required qualification to accomplish certain activities within limited time. However, it is hard to get the set skill and experiences with the set salary.

Standardization gaps

It has been identified that coding of the items in some questionnaires are not complete. Some items are coded and some are not. And the entries of un-coded data were ignored at data entry level. This led to losing some information which was not coded. For example, Industry department is generating data by saying X% coded and Y% un-coded. There is no standard set which percent of un-coded is accepted. This could create a problem in migrating data to RDBMS database.

The Industry survey dataset is also missing textual data entry into the system. This shows that what has been collected is not fully available in the electronic format. This pushes Data migrating process to refer back to the hardcopy of questionnaires.

There is no set standard which will be followed in preparing a given survey except having a procedure. Therefore inconsistencies have been manifested among different datasets of a given category of survey. In fact, price data is relatively consistent and in a better position as compared to others.

Not having reliable office
There are office problems at some of the branch offices. The branch offices that are facing problems are 13 in number. The problem of these offices is that the landlords can ask for more rent payment anytime so that CSA will be forced to change to another one which can be affordable. In that case, the office may loose its system installations, if any.

Gap on dataset storage

The datasets have been organized in unrelated files. They lack relational links with one another. That means they have not been organized in databases.

Gaps on analysis

- Getting data over range of time and space is not easily available under the current situations
- There is no Interactive Analysis on the electronically available datasets
- The skills of using available software systems is not the maximum, i.e. they are not using full features of the available systems
- Strategy is not in place to make use of the existing data as easily as possible

Gaps on dataset security and safety

- There is no backup strategy or policy
- There is no disaster recovery policy
- There is no electronic information preservation policy

Part Three: RDBMS and Detailed Recommendations

Table of contents **Page**

Part Three: RDBMS and Detailed Recommendations

1. RDBMS for CSA

1.1 Benefits of RDBMS

What are benefits of having well organized, structured and related datasets? What is a relationship between such organized datasets and application of RDBMS?

Why CSA does need Relational Database Management System? The Ethiopian Government is annually investing about 70 million ETB on CSA out of which about 38 million ETB per year on surveys or datasets which is about 59% of the total budget of CSA.

Eight out of the 12 major surveys are related to food security data as shown in table 1. And 31 datasets out of 44 are food security related except price datasets.

It is trivial that the country will not allocate such amount of money if the surveys are not useful for its economic development.

It is also obvious that value of data or information is increasing with its use as opposed to other resources. Value of information can be increased if its use is increased from time to time. Therefore, in order to make all the datasets more useful through their lifetime, we need to organize the datasets over a time range so that one can get real picture of a given situation.

In order to get such pictures we need to relate the datasets of a given survey in particular and of different surveys in general in a meaningful manner. This can be done by using the widely used RDBMS instead of the current flat file structure.

Major benefits of RDBMS are:

- Exhaustively using datasets which the government or other body has invested on
- Saving money and enhancing quality by re-using existing datasets
- Bring in positive change in planning, policy making and research results
- It can be resulted in fulfilling the statistical data requirements essential for planning, policy formulation, monitoring and evaluation, socio-economic policy analysis, food security and research activities in general.
- It can contribute a lot in setting up systems and mechanisms to ensure a sustainable flow of statistical data in Ethiopia and thereby wherever possible bridge over the existing statistical data gap.

Some of implied benefits/functions of RDBMS are:

- Online access to micro-data based on the existing data access policy

- Online access to Meta-data as required
- Ability to locate Data Sets online
- Ability to locate Questions within Data Sets.
- Online analytic capabilities. Users will have privileges to produce different reports vertically and horizontally on-line. This is complimentary to browsing option given below
- Search Tools
    - Data Sets
    - Questions/Variables
- Browsing Options
    - Subject, datasets title, time, space, etc.
- Exchange of data (export) as required by users of data
- Historical data are more secure in a relational database than in separate files which can be lost or deleted

1.2. Do we need Open source or commercial software in CSA?

Comparisons of OSS and Proprietary systems are given in the following sub-sections.

1.2.1 Advantages and disadvantages of OSS

Advantages of OSS

Motivations for using and developing OSS are mixed, ranging from philosophical and ethical reasons to pure practical issues. Some of the practical advantages are:

- The availability of the source code and the right to modify it is very important. It enables the unlimited tuning and improvement of a software product. It also makes it possible to port the code to new hardware, to adapt it to changing conditions, and to reach on a detailed understanding of how the system works.
- There is no one with the power to restrict in a unilateral way how the OSS is used, even in a retroactive way. For instance, when a proprietary software vendor decides not to upgrade some software product for some old platform. In this case, users can only stick to the old version of the software, or switch to another product.
- There is no single entity on which the future of the OSS depends.
- There are no black boxes possible in OSS. This point makes open source to be considered by many experts as one of the necessary conditions for dependable applications.
- There is always the possibility of forking (creating an alternative code) base if the current one is in some way perceived as wrongly managed.
- No per-copy fees can be asked for modified versions, and anyone can use the current code base to start new projects.

- Usually OSS is delivered when the development team feels that its quality is good enough. There is no single commercial entity of OSS pushing for precise delivery dates or features that must be supported.

Disadvantages of OSS

OSS development models lead also to the perception of some disadvantages. Some of them are:

- There is no guarantee that OSS development will happen. It is not possible to know if a project will ever reach a usable stage, and even if it reaches it, it may die later if there is not enough interest.
- It is sometimes difficult to know if an OSS project exists, and its current status.
- Documentation and user manuals can be hard to follow for non-techies.
- The OSS programs themselves may be extremely powerful once learned, but are not always as intuitive as they could be.
- Customization costs development time and money. There are limited financial incentives for improvements and innovations.
- Since most businesses operate on proprietary programs, so sharing information or documents might be difficult.

1.2.2 Advantages and disadvantages of Proprietary Software

Advantages of Proprietary Software

- Proprietary software exists to generate revenue.
- The proprietary software provides the vendor a guaranteed income which can be used to better service their customers. A proprietary software company usually listens to the needs of their customers, and responds accordingly.
- If a problem arises with proprietary software, then the company can usually fix it more quickly with the help of consultant than waiting for developer's community.
- It guarantees structured innovation, which is innovation that is planned within a single responsible organization.
- Proprietary software licenses provides protection for intellectual property

Disadvantages of Proprietary Software

- It is expensive to license and maintain proprietary software.
- It depends on proprietary or closed standards. Many technical specifications that are sometimes considered standards are proprietary rather than being open.
- Proprietary software has little or no local support. If the software manufacturer decides to discontinue development of the product, no one

has the right to take the program and continue development on it, effectively killing its usability in the market.
- The quality of the software depends entirely on the owner organization
- The local software industry is not developed because of major market share is owned by proprietary software.

1.3. How do we select an appropriate RDBMS

The following are some of the key issues we need to address in selecting an appropriate RDBMS for CSA. These are performance, reliability, scalability, platforms, standard, management, support, maturity, expertise, Total Cost of Ownership and spatial functionality. These parameters are considered by using the following tables.

The tables are used to compare general and technical information for the four widely used relational database management systems globally. We should keep in mind that all comparisons are based on the stable versions without any add-ons, extensions or external programs.

Table 2 : Maintainers, license and version in comparison

RDBMS	Maintainer	First public release date	version	License
Microsoft SQL Server	Microsoft	1989	9.00.3042 (2005 SP2)	Proprietary
MySQL	Sun Microsystems	November 1996	5.0.51	[v]GPL or proprietary
PostgreSQL	PostgreSQL Global Development Group	June 1989	8.3.3	BSD (open source and free)
Oracle	Oracle Corporation	November 1979	11g Release 1 (September 2007)	Proprietary

[v] Please note that MySQL has two versions. One version with less features is open and free of charge. But if some one needs full features of the system it is not for free.

Table 3: Limitations of RDBMS

RDBMS	Max DB size	Max table size	Max row size	Max columns per row [vi]	Max Blob/Clob size	Max CHAR size
Microsoft SQL Server	524,258 TB (32,767 files * 16 TB max file size)	524,258 TB	8060 B (Sql Server 2000)	1024	2 GB	8000 B
MySql 5	Unlimited	2 GB (Win32 FAT32) to 16 TB (Solaris)	64 KB	3398	4 GB (longtext, longblob)	64 KB (text)
Oracle	Unlimited (4 GB * block size per tablespace)	4 GB * block size (with BIGFILE tablespace)	Unlimited	1000	4 GB (or max datafile size for platform)	4000 B
PostgreSQL	Unlimited	32 TB	1.6 TB	250-1600 depending on type	1 GB (text, bytea) - stored inline	1 GB

Table 4: Basic features of the open sources RDBMS

Features	MySQL	PostgreSQL
Maximum Number of Joined Tables	61	152
Maximum Logical Operators in WHERE clause	AND: ≥ 5354 OR: ≥ 5354 RANDOM: ≥ 5354	AND: ≥ 1006 OR: ≥ 5008 RANDOM: ≥ 5008
SQL-Standards	Several differences between MySQL and standard SQL. INNER/OUTER join SQL-92 syntax supported. No full outer join.	Subset of both the SQL-92 and SQL-99 standards. INNER/OUTER join SQL-92 syntax supported.

[vi] - Number of fields per table of the given database

Features	MySQL	PostgreSQL
Database Links	Not supported	Supported
Online Backup	Only for commercial version	Supported
Online Reorganization	No index rebuild. Analyze can cause instabilities	Indexes can be added to and removed from tables at any time.
Security	Proprietary authentication protocol using user/password combination. No support for Kerberos, LDAP, ... Built-in SSL support	Kerberos for authentication Built-in SSL, PAM, MD5 and SSH support
Load Balancing	No parallel query, partial table scans in clusters	Supported SELECT load balancing between two nodes through add-on module
Limits per Table	Number of rows: Limited by maximum table space Indexes per table: 64	Number of rows: Limited by table size Indexes per table: Unlimited

Table 5: Initial Cost of RDBMSs

RDBMS	Price per 5 license and/or per CPU	Support
Microsoft SQL Server	2K USD	
MySQL	0.00 USD for free version and commercial version costs from 600 to 5000.00 USD	Free support for the free version and commercial support available for the commercial version.
Oracle	40K USD	
PostgreSQL	0.00 USD	Free and commercial support available.

Table 6: More basic features for Oracle and MS-SQL server

Features	Oracle	MS-SQL server
ANSI-SQL compatibility	ANSI-SQL 92	ANSI-SQL 92
Operating system portability	Unix, and Windows	Only Windows
Backup and recovery facility	Online, incremental, point-in-time back up and recovery	Online, incremental, point-in-time back up and recovery

Network support for wide range of protocols	TCP/IP, SPX/IPX and all popular protocols	TCP/IP, SPX/IPX and all popular protocols
Support distributed database replication	yes	yes
Data import and export	ASCII, DBF, WKS data loading	ASCII, Data loading
Spatial analysis feature	Yes, it supports	Yes, it supports
Security	Less secured	More secured

Major Spatial Data Features MS-SQL server 2008

It implements Round Earth solutions with the geography data type. Use latitude and longitude coordinates to define areas on the Earth's surface. Implement Flat Earth solutions with the geometry data type. Store polygons, points, and lines that are associated with projected planar surfaces and naturally planar data, such as interior spaces. Microsoft SQL Server 2008 delivers comprehensive spatial support that enables organizations to seamlessly consume, use, and extend location-based data through spatial-enabled applications, ultimately helping end users make better decisions. Specifically the system has the following features.

- Use the new **GEOGRAPHY data type** to store geodetic spatial data and perform operations on it.
- Use the new **GEOMETRY data type** to store planar spatial data and perform
  operations on it.
- Take advantage of new **spatial indexes** for high-performance queries.
- Use the new spatial results tab to quickly and easily see **spatial query results** directly from within SQL Server Management Studio through support for spatial standards and specifications.
- Extend spatial data capabilities by building or integrating **location**

Partner Ecosystem

Approximately 15,000 ISVs support MS-SQL Server. MySQL has approximately 70 developers and 50 support staff.

PostgreSQL community is also estimated not to be less than the one of MySQL and Oracle community is also following MS-SQL server.

Therefore in terms of partner ecosystem all of them are in a good condition.

Which RDBMS do we put on the top from open sources for CSA?

Based on the aforementioned information we are in a position to select the best one from the open source RDBMSs.

MySQL

a. Performance of MySQL is very fast for common DB operations.

b. We can enjoy large community support for MySQL. In fact nearly every problem we face has been seen by someone else which makes it literally great.

c. The latest version 6+ has triggers, replication, and high-availability on UNIX with heartbeat, stored procedures and views.

PostgreSQL
PostgreSQL is a powerful, open source relational database management system. It has been in development for more than 20 years and has a strong reputation for excellent architecture and world-class reliability, data integrity, and correctness.

PostgreSQL is an enterprise-class database that boasts sophisticated features, such as Multi-Version Concurrency Control (MVCC), point-in-time recovery, tablespaces, asynchronous replication, nested transactions (savepoints), online/hot backups, a sophisticated query planner/optimizer, and write-ahead logging for fault tolerance. It supports international character sets, multi-byte character encodings, and Unicode, and it is locale-aware for sorting, case-sensitivity, and formatting. PostgreSQL is highly scalable, both in the quantity of data it can manage and in the number of concurrent users it can accommodate.

PostGreSQL is a high-end Oracle like database, in fact at installation of the instance, it requests the type of installation you need (oracle type or ...), the web based monitoring and logging is excellent. It is really a high-end of the currently available free RDBMSs. As of recent time it is ready for windows. It is truly open source RDBMS.

There is a fear that MySQL is retreating from Open source community to free software. For instance, MySQL AB is the one in open source and the enterprise is taking most features for cost. In addition documentation for Version 6 of MySQL is not under GNU license.

Table 7: support for spatial functions

MySQL	PostreSQL
OGC mostly only MBR (bounding box functions) few true spatial relation functions, 2D only	Over 300 functions and operators, no geodetic support except for point-2-point non-indexed distance functions, custom PostGIS for 2D and some 3D, some MM support of circular strings and compound curves

Table 8: Supported Geometry Types

MySQL	PostreSQL
2D, can store 3D 4D(e.g. M and Z) but no functions do anything with those) - Polygon, Point, LineString, MultiPoint, MultiPolygon,MultiLineString, GeometryCollection	2D, some 3D, 4D (support for storing Z,M but most spatial functions ignore the higher dimensions) and some curve support - Polygon, Point, LineString, MultiPoint, MultiPolygon, MultiLineString, GeometryCollection, CircularString, CompoundCurve, CurvePolygon, MultiCurve, MultiSurface

The spatial feature also pushes us to select PostgreSQL over MySQL.

Therefore from the open source RDBMS, PostGreSQL is the appropriate selection for CSA.

Which RDBMS do we put on the top from commercial systems for CSA?

Microsoft SQL Server is the leading according to the 2007 Database and Data Access, Integration and Reporting Study, completed by BZ Research in late June 2007, 74.7 percent of enterprises are using SQL Server. This is slightly lower than the 76.4 percent reported in a comparable July 2006 study, but it's still significantly higher than the other popular databases.

The same study showed that the other top databases, in terms of use, are Oracle 54.5 percent in 2007.

As indicated in the above tables technically MS-SQL server can fit in the set requirements.

In addition there are more Ethiopian ICT professionals who are working on MS-SQL server than the once who are working on Oracle.

Security of MS-SQL server has been found out to stronger than security system of Oracle. Same flaws were repeated in Oracle for a number of releases.

In terms their costs MS-SQL server is cheaper than Oracle and it's affordable for CSA.

Total Cost of Ownership (TCO)

SQL Server delivers high quality at a low TCO by providing CSA with a comprehensive data platform solution out of the box, with no need for expensive add-ons.

- How much does MS-SQL cost? Do we have free release or only commercial? Yes, we do have free versions with less features
- A set of world-class tools and an integrated debugging environment help reduce development costs.
- SQL Server Management Studio, which is designed to help create a self-managing system, helps reduce staffing costs.
- Reduced TCO and faster development time with the common engineering strategy implemented across Windows Server products. Plus, Microsoft offers a variety of SQL Server licensing and pricing options, with each one providing robust support.

Please note that Oracle has no free version software.

Therefore from the commercially available and widely used RDBMSs, MS-SQL server is recommend mainly by considering availability of experts nationally, internationally, and affordability.

Then, which RDBMS is suitable for CSA; MS-SQL server or PostgreSQL?

PostgreSQL or MS-SQL server

We selected PostgreSQL from the open source RDBMS and MS-SQL server from commercial RDBMS. Which one do we select from the two? Why? Which one can give more benefit to CSA?

As depicted above user base of MS-SQL server is leading all over the world and in Ethiopia by using visual basic programming languages. Second group of

experts are working on MySQL particularly in the area of higher learning institutes. There are very few who are working on Oracle and PostgreSQL in Ethiopia. Most of the professionals who are working on Oracle are working for private IT companies in Ethiopia.

Conclusion: Therefore the latest version of MS-SQL server is recommended to be used by CSA.

1.4 ICT Human Resources for the proposed system

Major skills required for the proposed system are: system analyst, information architect, database administrator and database programmers.

The ICT human resources with the above skills are found in three different clusters. These are:

- International expertise (East Africa or neighboring developing countries)
- National expertise
- Internal (within CSA)

Is there any need to use International expertise? The availability of professionals we have in East Africa is more or less similar in specialization. However there are some more specialized private companies who specialized in different systems of the state of the art of the technology in East Africa than what we have in Ethiopia.

Here we can also understand that there is an economic integration in east Africa through different Regional Economic Communities (RECs) setup like the Common Market for Eastern and Southern Africa (COMESA), the Inter-Governmental Authority on Development (IGAD), the East African Community (EAC). This can also tell us that CSA could see not only what is available nationally but also what are available in East Africa.

What are chances of maintaining ICT technical staff at the Agency level?

Availability of relevant human resources and salary scale in Ethiopia can be grouped into three major groups. Minimum salary is paid by public institutions which are categorized under Civil Service Agency. CSA's staff members are categorized in this group.

The medium paying institutions are NGOs, private ICT companies, banking and development agencies of the government.

The highly paying category is UN and International institutions. It should be noted that most of technically advanced professionals are currently working for UN, International institutions, International NGOs, and private companies. Currently, the Ethiopian ICT experts have started migrating to abroad.

From this situation we can easily understand that if CSA is trying to retain particularly its ICT staff, it is almost impossible or the chance of their stay is low. This shows that CSA need not purely rely on internal staff member at anytime.

Thus, the sustainability of the proposed system can be realized if CSA can consider the involvement of professionals we have in the country and East Africa. This may help CSA even to attract more sponsors of some of the surveys.

Table 9: Situation of using different categories of expertise

Employing expertise	**Likelihood/ Availability**	**Knowledge/Experience**	**Cost**
Internal expertise	low	Low	Less
National expertise (Company or individual)	Medium	Medium	Medium
International expertise (Company or individual)	High	High	High

Although managing internal expertise is more convenient for CSA retaining the local experts and having sustainable professionals is less probable. National and/or International experts are by far reliable and more sustainable.

Pros and cons having expertise from different categories or one category shall be assessed in depth and reach on decision by CSA.

1.5 Estimated cost of the proposed system

In a daily business, information and communications technology is still seen as a main cost. This, however, means ignoring the value added by ICT. Focusing on cost alone is supported by the fact that IT costs are more or less transparent, while IT value-added cannot as easily be shown and managed as depicted in this report.

Cost of the newly proposed system

Table[vii] 10: Hardware required at the Head quarter

Items	Available quantity	Missed quantity	Total quantity
PCs	All	Nill	All
DB server	2	0	2
File server	1	0	1
Web server	1	0	1
FTP server	1	0	1
Backup server	1	0	1
Connection to Internet	1	0	1

This table shows as that at the initial stage of the system implementation additional hardware is not expected as a result of RDBMS. This is true if and only if the hardware can be procured according the Network Master Plan.

Table 11: Hardware required at the Branch offices

Items	Available quantity	Missing	Total quantity
PCS	5	0	5
DB server	1	0	1
Connection to HQ	Nill	required	

For the branch office we can also use the proposed server as a Database server and 5 PCs proposed by the Network Master Plan.

Table 12: Initial software cost

Items	Quantity	Estimate price	Remark
MS-SQL server	5 license	2000 USD	Latest version is required
System development tool	1 or 2 licenses	1000USD	Complete Visual studio (latest version)
Total cost of the software			**3000USD**

The price given in the table 12 is one time cost. There is no need for software maintenance fee unless the Agency decided to upgrade to a new version of the software.

[vii] Here it is assumed that the proposal of the Network Master Plan will be implemented before the system development is taking place.

Table 13: System development and training cost

Items	Quantity	Estimate total price	Remark
System development surveys	10 surveys	50K USD	If out sourced to a national company or expert
Migration of datasets	10	10KUSD	This cost shall be attached with system development
Training on MS-SQL server DBA	1	1K USD	On-site group training for about 20 hours
Training on how to use the developed system	1	200 USD	Two days training.
Total cost of system development and training			**61,200USD**

As indicated in the above table, the total cost of development is estimated to be 61K USD. Here it assumed that each of the surveys costs about 6K USD at an average.

Development cost – The cost of the system development is the only added significant cost to CSA. Since the added value of the new system is significantly enhancing the use of survey data at various levels, this cost may be covered by donors such as survey sponsors.

The quality of the database management often translates directly into client satisfaction and having the database is one of your most important competitive advantages. Hence, companies go to great expense to carefully protect their mission-critical data resources. As database administrator (DBA) staffing costs rise, many companies realize that they cannot justify a full-time database administrator, especially when senior DBA salaries approach 120K USD per year in USA and Europe.

In today's connected world, it can be fiscally prudent to hire remote DBA specialists. But more important is the potential for loss of institutional knowledge. The average attrition of a database administrator is less than 5 years, and using a remote DBA provider ensure continuity of coverage and no loss of database support. How much does it cost in Ethiopia and East Africa? It is estimated that senior DBA expert costs about 24K USD in Ethiopia and 45K USD in East Africa per year full time. In fact, if only sometime has been leased this cost can be reduced significantly.

Generally, it is suggested to make use of three categories of experts by setting appropriate modalities.

Table 14: Summary of comparison of estimated costs for the existing and proposed systems

Major entities	**Estimate cost of existing system in ETB**	**Estimate cost of RDBMS in USD**
Development	0	61K
Workstations/PCs	2,500,000	00
Servers (Hardware)	400,000	00
Software Tools	0.0	3K
RDBMS	0.00	2K
Operating system software	Not available	same
Human resources for the system development		
Internal	2,365,715 per month	Same or less if outsourced
National	Not used	It depends on the level of involvement
International	Not used	It depends on the level of involvement
Printing	150,391 per month	same
Transport	510,778 per month	Lesser[viii]
Maintenance	Not used	4K – whoever has developed it

Please note that the above estimate cost of system development is given for system development (database design and program development), migration of datasets to the new system, training on the new system has given for national expert or company. If it is given to international expert this cost can be more than the estimated one.

Implications of the new system

CSA would benefit from the new organizational structure if it can slightly modify the existing one to have for example data entry personnel at all branch offices and second level of quality control at the head office as discussed earlier.

[viii] There is a cost reduction in transportation. This reduction is occurred because there is no need to transport filled questionnaires from branch office to the HQ. This cost includes perdium of drivers and transportations (fuel cost and cars).

What will be a relationship of SMS, DP and ICT? These three Departments are sharing the same central database and working on it from different angles accordingly. Particularly DP and ICT should operate more closely.

The RDBMS has positive impact on different tasks of different departments by avoiding duplication of efforts. In addition validation of data can be done differently by using internal validation by the system, for example by using histogram, or other options.

Based on these principles we can come up with the following modifications on the institutional setup in survey information management.

Table 15: Institutional set up for the RDBMS

Major activity	Actor/s for the existing system	Actor/s for the new system
1. Plan Questionnaire	CSA, NGO, Public institutions	same
2. Questionnaire preparation/design	Subject matter specialist, Data processing department	same
3. Questionnaire dispatch	Field operation department, Subject matter specialist	same
4. Questionnaire filling	Enumerators, technical assistance from Subject matter specialist	same
5. Questionnaire retrieval/collection	Field operations department	same
6. Questionnaire storage	HQ: Data processing department through the Editing, coding and documentation team -- documentation	Branch office – HQ can also store it as a custody
7. Editing and coding	HQ: Data processing department through the Editing, coding and documentation team and Subject matter specialist	Branch Office: similar professionals
8. Data entry	HQ: Keying by Data Processing Department through data entry, cleaning and computer operation and checked by Subject matter specialist	Branch office: similar professionals
9. Data cleaning (first level)	HQ: Data Processing Department through data entry, cleaning and computer operation and Subject matter specialist	Branch office: similar professionals
10. Second level quality control	HQ: Subject Matter Technicians	HQ: Subject Matter Technicians
11. Reformatting data	HQ: Data Processing Department	HQ: Data Processing and fairly

Major activity	Actor/s for the existing system	Actor/s for the new system
	through Systems and programming team	automated
12. Tabulation and Analytical Output	HQ: Data Processing Department through Systems and programming team/ Subject Matter Specialists	Fairly Automated after the first few times
13. Data preparation/conversion	Data Processing Department through Systems and programming team	Fairly automated
14. System documentation	Data Processing Department through Systems and programming team	Can be done either by DP or ICT
15. System backup	Prepared by Data Processing and sent to ICT department	Can be done either by DP or ICT
16. Preparing documents in appropriate format	ICT	ICT
17. Ad hoc Analysis	Currently not available	Designated Users with appropriate access to data and analytical tools
18. Electronic Information dissemination	ICT	ICT
19. Hard copies and CD-ROMs disseminations	Public relations	Public relations

Formatted: Bullets and Numbering

Formatted: Bullets and Numbering

As it has been shown in table 15, the new system avoids manual activity transfer from one to the other department in survey dataset management. In addition at least an average of one week for report generation is reduced because it overcomes transportation of filled questionnaires from branch offices. Most importantly data availability will be by far speeded up for urgent use of survey results.

1.6 Conversion to the proposed system

The system conversion shall be implemented on phase basis. The phases should be based on priority/category order. In addition both existing and proposed system should run in parallel till full system conversion of a given survey takes places.

Most of the survey datasets can be migrated to relational DBMS while few of them are not cost effective to be migrated to RDBMS (In fact, the price data are already being converted into an initial MS SQL database). The ones which are not recommended are Agricultural enumeration survey and population & housing census which are being undertaken every ten years [if CSA needs to have them in RDBMS technically it can be done]. The third one which is not recommended to be used over RDBMS is Child Labour Survey. There is no indication weather the Child Labour survey can be continued or not.

Based on the need of CSA and benefits which can be obtained from the conversion the surveys for which CSA needs to develop RDBMS is suggested to be in the following priority order:

- Price Surveys (underway in MS SQL)
- Annual Agricultural Sample Survey
- Livestock and Livestock Characteristics Survey
- Large and Medium Manufacturing Industries Survey
- National Labour Force Survey
- Urban annual Employment and Unemployment Survey
- Ethiopia Welfare Monitoring Survey
- Demographic and Health Survey
- Household Income Consumption and Expenditure Survey
- Agricultural enumeration survey and
- population & housing census

These can also be further grouped in three major categories of priorities. These are:

Category I: monthly datasets

- Price surveys

Category II: yearly datasets

- Annual Agricultural Sample Survey
- Livestock and Livestock Characteristics Survey

- Large and Medium Manufacturing Industries Survey
- Urban annual Employment and Unemployment Survey

Category III: Less frequent datasets

- National Labour Force Survey
- Ethiopia Welfare Monitoring Survey
- Demographic and Health Survey
- Household Income Consumption and Expenditure Survey
- Agricultural enumeration survey and
- population & housing census

The development and conversion timing of system depends on the type of employed experts and/or availability of fund and commitment of the senior managements of CSA.

Data transmission from branch office to the head quarter

- It is more secure, if we do have one working or temporary server at the head office whereby data entry can be done online. Once everything is completed for a given dataset we need to publish it on the productions server for final use. In the future, whenever technology and connectivity allows, it is suggested that databases should be replicated at each branch offices DB servers.
- As an option if the Agency insists to have the existing system, CSPro can be used as usual and the clean and final datasets can be migrated to the production server. This approach can be used only at the initial stage of ISIS. Eventually it should give a way to the use of template for the respective database. Then after, CSPro will give its way to the new system.

System maintenance: For the system maintenance there should be an agreement with system developer to provide support services for sometime, a year or two. Note that maintenances can be done by using any of the three categories of expertise.

2. High level design of Integrated Survey Information System (ISIS)

An Integrated Survey Information System is generally any kind of computing system that is of "enterprise class". This means typically offering high quality of service, dealing with large volumes of data and capable of supporting CSA.

The Integrated Survey Information System will provide a technology platform that enables the Agency to integrate and coordinate their business processes. It provides a single system that is central to the organization and ensures that information can be shared across all functional levels and management

hierarchies. This system is invaluable in eliminating the problem of information fragmentation caused by multiple information systems in the agency, by creating a standard data structure and central databases.

2.1 Server/service structure

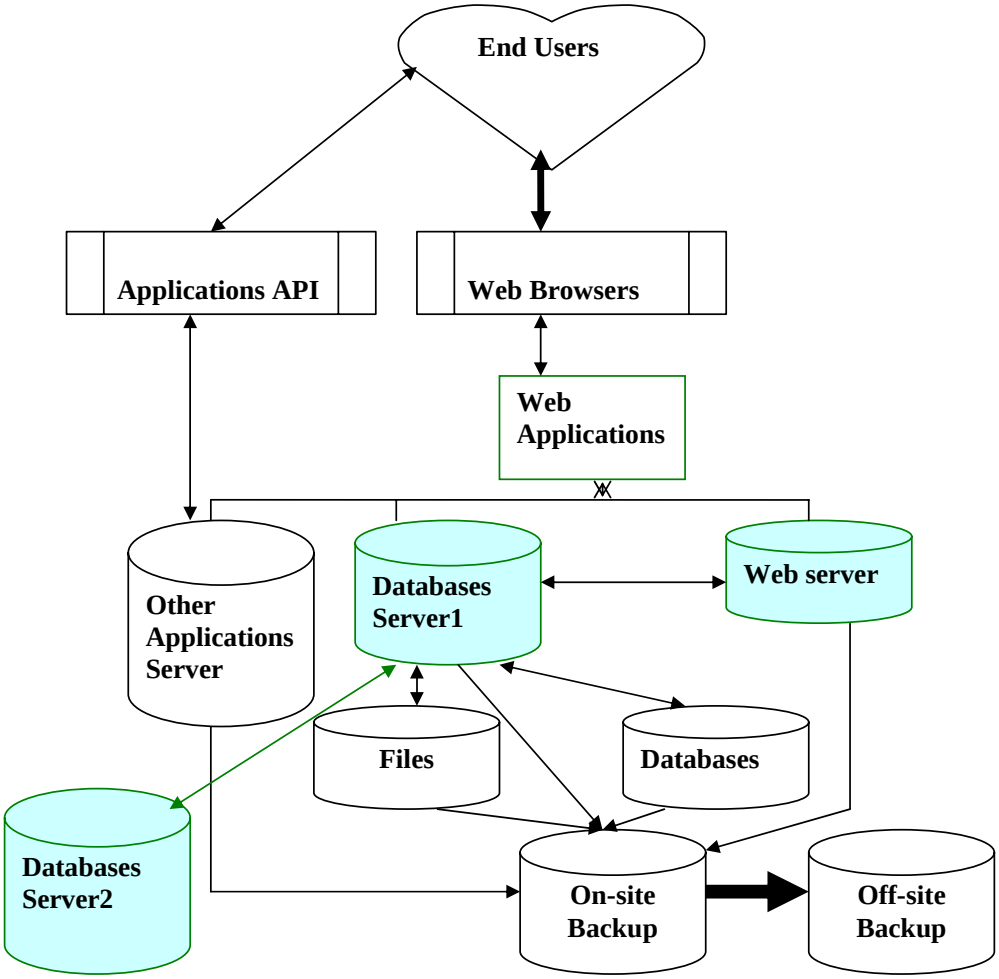


Figure 1: Higher level design of the ISIS at the head office

The proposed system is based on web-integrated collaboration systems and tools. It also shows only high level design.

2.2 How can each of the survey related to one another in RDBMS?

It is difficult and challenging to standardize the details of analysis datasets across studies. The study designs and scope of the studies vary. The standardization outlines here do not mandate that all derived variables and the “contents” of each of the analysis datasets be identical across all studies. Rather they intend to standardize the file “structure”, such as one observation per subject for demographic analysis dataset. This also means that the field names and file names should be consistent across studies. The exact variables and the definition of those variables of each analysis dataset may be study specific, but it is required that the relevant information related to a common analysis purpose be included together. For example, demographic related data should be included in demographic analysis dataset, and efficiency related data should be included in efficacy analysis dataset, and so on. Finally, it is important that the attributes of key common variables and data code lists be consistent across studies as much as possible.

The standardization and consistency of analysis datasets across studies allows users to search information more quickly since all related data is grouped together. With standardization, programmers need only to acquire the basic understanding of the analysis dataset structure and the scope of the analysis datasets. After they are trained on the standard analysis dataset structures, they can then apply the same understanding in programming analysis datasets for any survey. This significantly reduces learning curve that study specific analysis dataset customization would otherwise create. As a result, the standardization saves programming time, increases productivity and allows CSA more resource allocation flexibility.

Generally stating the datasets of a given survey can be treated on year or time interval basis.

Such action can help CSA data management system to be used easily. Import and export features can be implemented very easily. GIS and other graphical data presentations tools can share same micro data for different purposes.

How do we share common enteritis across different surveys? As mentioned in the previous report (report1), we need to have all common entities in one common place and use a connection string to share from one common point. The updating or administering the content can be done through a single and assigned section or team.

It should be noted that Data Tabulation Model (DTM) should be devised for each of the surveys before implementing the proposed system. This can help to analyze sequential datasets of a given survey over an applicable space.

Here it should also be noted that the end users (unauthorized users) should not be allowed to have direct access to microdata except their summary information so as to comply with micro data access policy.

The microdata in RDBMS should have batch import and export facility to common standard system so that it can be used for archiving and dissemination purpose for example by using IHSN Microdata Management Toolkit.

Finally, as a standard to the new RDBMS clients should be able to use SPSS, Excel, Word and even PowerPoint to extract data from MS-SQL Server based on their selection or presences. In fact, limitations of each of these systems should be considered before extracting data for further or different analysis.

2.3 Sample entry points for Ethiopian ISIS (Portal)

The portal is expected to have the following major entry points in addition to functions listed in section 3.1. The suggested entries in bold fonts are leading topics with others under each of them.

Agriculture: Land use, Area production and yield, Farm management and practice, Livestock breeding, Forestry, Fishing

Industry: Large, medium, small, Industrial production, industry distribution,

National/regional Accounts: Gross National Domestic Product, Gross Regional product

External trade: Foreign exchange, Balance of payment

Construction: Housing, Water supply service, etc.

Employment and unemployment: Employment, unemployment, labor, wages

Health: Demographic and health situation by nation, regions, zone, etc

Wholesale & retail trade, Service: Over various commodities

Household income and expenditure: Over region, Zone, etc

Population: Total, Amhara, Oromia, Tgiray, etc.. Sex, Religion, etc.

Price: Beverages, non-beverages, markets, commodities, etc

Major functions required from the system are:

- Search: simple and advanced searches
- Export and import functions: export micro data to SPSS, IHSN; export statistical output or tables to MS-Excel; import from CSPro based on users roles
- Provide comparison data/information over time series of a given survey at different levels, such as at country level, regional level, zone level. Generally based on available data it should be able to give situation of data over possible spaces.
- Viewing micro data through GIS, possibly through GIEWS Workstation of FAO
- Generating graphs-can be incorporated with other features
- Generating tables, etc- can be incorporated with other features

The entire implemented functions should be **role based** administration and use of the portal for the two category of users; CSA and public users. Public users should have a very limited to reading plus limited access to micro-data. In addition the CSA staff members should have different types of privilege e.g. administrators full control over the system, data entry will have only data entry access, quality controller should have a role of applying raising factor accordingly, etc.

2.4 Other important considerations

2.4.1 Backup and arching strategy (outline)

Data disasters are very infrequent, but they destroy both critical as well trivial data indiscriminately. Therefore preparing data disaster recovery procedure is most. For the recovery system backup is a possible and applicable solution. The backup devices includes DLT (Digital linear Tape), DAT, DVD, CD-ROM, etc.

Magnetic Media

Magnetic media, including diskettes, fixed disks, and tapes are subject to corruption. The information on these media are required by the application of magnetic fields, and are subject to disruption by other magnetic influences. Therefore, these media must be kept in a place that will diminish the possibility of magnetic interference.

How many generations of the backup tapes to retain and for how long? Modern high-density backup tapes have a limited number of re-uses and some manufactures recommend that tapes should be re-used no more than six times and then should be archived or discarded. Therefore, every DLT (Digital linear Tape) is used for six times based on the recommendation of the manufacturer. Daily backup tapes are rotated weekly while weekly tapes are monthly. Monthly tapes are stores for a year while yearly tapes are stored permanently. Others use incremental backup system of two DAT tapes at a time.

Storage of backup tapes

It is strongly recommended that backups should be stored onsite and offsite. The storage location of offsite backup should be distant enough from the organization that any foreseeable catastrophe would be unlikely to make both the computer location and backup location inaccessible. Therefore, most of the backups on site backup are kept in a locked fireproof cabinet in the data center and the offsite backup is arranged in a distant location where the data are stored in safe custody with proper temperature. The tapes should be clearly labeled.

Table 16: Who is taking what backup, when?

	Item	**When to be taken**	**No of backups**	**Retention period**	**Who is taking backups**	**Remark**
1	Data					
2	Program and data					
3	System[9]					

Procedures for approving restoration of data

[9] Also system backups are taken whenever new version is installed, or any major change is done to system configuration or new patches are added.

Occasionally, it may be necessary to restore data from a backup. This may happen for any number of reasons, from hard disk failure to apparent corruption of a file to major operator or user error. Therefore, restoration is strictly controlled and is not to be at the discretion of an operator. Any restoration process that affects the service requires approval from senior management and senior ICT management.

This will be done after assessing the reasons, alternatives, data that have to be reentered, and downtime of the system.

2.4.2 Requirement for Intranet for collaborative work

Major functions of the proposed Intranet

- Acts as an organizational communication tool helping the Agency communicate to its employees more effectively

- Acts as an administration tool helping CSA staff from both the branch office and head office in Addis Ababa campuses gain access to resources for the services sections e.g. Human Resource documents e.g. Leave forms, policies etc, Information and Communications Technology manuals and policies, Finance information etc.

- Assist staff members gain access to Knowledge Management tools and external databases.

Benefits of Intranet

- Helps CSA staff stay in touch with what's happening in the organization.

- Helps the Communication department communicate organizational news, changes and policies more effectively to CSA staff all over the country

- Helps all staff gain access to the latest institutional information and any news release in the Agency

For instance, the ICT department will have its own knowledge data system about the know-how of its ICT people. The contents of this system can be updated once a year. Every ICT person updates his or her own data in their own section on the Intranet. After this, IT teams go together through the data of the team

members. This way the contents of the system will be up-to-date and easily accessible and useable.

2.4.3 Requirement of MIS

MIS can help CSA to enhance its productivity. Therefore it is suggested to have appropriate MIS for the institute. This platform should also be linked to the Intranet of the Agency.

This system should be integrated with the proposed Intranet. Then, the staff members will get all required management information on the fingertips.

For instance, if some asks for budget breakdown over staff category, survey, stationary, transportation, etc of the budget year he/she should be able to get it in seconds.

Annex

TOR for the Development and Implementation of RDBMS

I. Background

The Ethiopian Central Statistics Agency (CSA) has decided to develop and implement a Relational Database Management System (RDBMS) to centralize existing and future survey datasets. A centralized relational database management system will streamline existing procedures and methods, promote standards, and provide the ability of statistical and spatial analysis across time and subject matters.

This implemented centralized relational database management system will be able to entertain data entry, processing, storage, analysis, and dissemination as easily as possible.

The suggested RDBMS for the purpose is MS-SQL server 2008. The programming language used should be in .NET environment.

II. Tasks of the project

Main objective of the project is to use the RDBMS study as the strategy to implement a CSA Relational Database Management System.

The specific objectives of the project are:

1. To develop a project work plan outlining the specific tasks after consultation with CSA management and staff.
2. To revisit and re-define in detail the process and capacity requirements of the

eventual system and to ensure CSA management have all appropriate details and understand the implications of the new system.

3. To design a system for the current surveys which has more than one datasets that CSA collect regularly, and in the future
    - Conceptual database design: Identifications of enteritis, relationships and attributes
    - Logical database design: design of relations
    - Physical database design: Physically implementing in MS-SQL RDBMS

For Eight food security related surveys undertake the following activiteis

4. To design data entry, data cleaning, data view screens
5. To generate various automated statistical outputs (graphs, charts, tables) in a user friendly manner
6. To enable manual statistical analysis, via customized or third party software (Excel, SPSS, CountryStat, other)
7. To have appropriate import and export features and ready to import from CSPro, and other standard systems and export to SPSS, MS-Excel, MS-Word, IHSN and other Web-based dissemination tools.
8. To implement efficient access security system both at database (back-end) and front-end levels
9. To implement a data backup system to ensure against data loss and corruption.
10. To implement and setup the developed system on the central database server and make sure that data entry can be done from branch offices both online and offline
11. To document system development processes and to document all programming codes to the standard
12. To provide all source codes both on the database server and on CDs in a meaning full manner
13. To work with internally assigned CSA programmers (at least 2) during all stages of the development to ensure proper internal understanding of the system development.
14. To provide technical support for an identified period of time to ensure CSA are able to own the system and take it forward.

III Features of the RDBMS

Some of the functions of RDBMS are:

- Online and offline data entry from HQ and Branch offices
- Users roles based data entry, processing and management (e.g. Internet users, internal browsers, data entry, data cleaning, etc)
- Online access to micro-data based on the existing data access policy
- Online access to Meta-data as required
- Ability to locate Datasets online
- Standard validation system accordingly
- Ability to locate Questions within Datasets.
- Ability and flexibility to apply raising factors by authorized user
- Ability and flexibility to entertain unique surveys whenever undertaken
- Ability to generate PDF output
- Ability to print required output
- Generate standard tables, graphs for
- Online and flexible analytic capabilities.

- Statistical reports over series of time, space and subjects accordingly
- Search Tools
    - Dataset
    - Questions/Variables/metadata
    - Documents associated with datasets
- Browsing options
    - Subject, datasets title, time, space, etc.
- Exchange of data (export and import) as required by users of data

IV. Human resources and Deliverables

Two external consultants (one project manager, one database programmer) and at least two internal staff (database programmers) are required. The deliverables for each are included below:

Project Manager (external): Under the supervision of CSA Deputy Director, the consultant will document and present the system development to CSA and stakeholders. Upon receiving feedback from the interested parties, amendments and modification request will also be included in the system. Specific deliverables include the following:

1. Develop detailed work plan and timeline once on board.
2. Revisit and Revise system specifics document to present to CSA
3. Design the database/s with full descriptions
4. Oversee all elements of database development
5. Ensure external programmer is working with internal programmers to train and hand over system
6. Ensure management and technical staff (HQ and Branch Offices) are fully aware of the process and end results
7. Keep in regular communication and provide status reports to CSA management on a regular basis.
8. Document process as required
9. Assist in developing user guides to ensure system is well understood by users.
10. Other tasks as defined by CSA management

Database Programmer (external): Under the supervision of the Project Manager, the consultant will design database and develop all necessary codes for the completion of a central relational database management system for CSA. The consultant will work within the guidelines of the RDBMS study proposal and ensure CSA programmers are involved in every step during the development. Specific tasks include:

1. Review and assist in revising the RDBMS design document
2. Design and Code RDBMS as identified above in tasks and features
3. Provide both source code and complied codes after each of the components have been developed.
4. Provide system documentation
5. Assist in writing reports on all developed sub-systems, files, and associated documentation on a CD with explanatory notes on all files provided.
6. Other tasks as requested by Project Manager.

Database Programmers (3 CSA staff) with the following tasks:

1. Assist in design phase
2. Assist in code development – all aspects as outlined above
3. Ensure proper understanding of code and how to add future data sets to RDBMS
4. Ensure proper understanding of code and how to program additional modules for import, cleaning, analysis, export, etc
5. Assist in preparing documents for users and technical staff
6. Assist in presenting status to management and any internal/external presentations as required

V. Experiences and skill requirements

Given the complex nature of the work, the external consultants must have extensive expertise in the development of RDBMS for mid and large institutions. The consultants must also have previous experience in assessing and developing statistical and spatial (GIS) analysis database systems.

Project Manager

- Minimum 3 years project management in information systems related projects.
- Minimum of 5 years experience managing database systems for mid/large business out of which 3 years are in MS-SQL and/or 2 years in related RDBMS
- Minimum of 5 years experience managing projects related to database systems, information systems, web development, and internet security is required.
- Extensive knowledge in recent developments in information technology and computer-based systems training is essential.
- Experience both in public and private sectors
- Fluency in Amharic and English for verbal and written communication

Database Programmer

- Minimum of 5 years experience building RDBMS systems for mid/large business out of which 3 years are in MS-SQL and/or 2 years in related RDBMS
- Minimum of 5 years experience in developing complex computer applications - design, implementation and maintenance by using different programming languages
- Minimum of 5 years experience in on-line systems, web development, and internet security is required.
- Extensive knowledge in recent developments in information technology and computer-based systems training is essential.
- Experience both in public and private sectors
- Fluency in Amharic and English for verbal and written communication

VI. Duration

The project should be completed within 10 to 12 months.

- Project Manager for an initial time of 4 to 5 months (spread over the first 10 month period) and 1 month (part time for 12 months as support)
- Database Programmer for an initial time of 10 months (full time) and 2 months (part time for 12 months as support).
- 2 Internal Programmers for an initial time of 8 months each and eventual full ownership of the system.

References

1. http://www.csa.gov.et/ , June 19, 2008 [CSA Official website]
2. System study of Wide Area Network, 2008 [consultancy report]
3. Thomas Gabrielle; Information Management Final Report Mission 1, 2008 [mission report]
4. Thomas Gabrielle; Ethiopia CPI Methodology, 2008 [Mission report]
5. Directive no. 1 /2004: Directive issued to establish procedures for accessing raw data to users (http://www.csa.gov.et/text_files/directives.htm)
6. Proclamation no. 442/2005: a proclamation to establish the central statistics agency
7. http://www.dba-oracle.com/oracle_news/news_oracle_prices_2007.htm
8. http://seclists.org/bugtraq/2005/Oct/0056.html
9. http://www.microsoft.com/sqlserver/2008/en/us/spatial-data.aspx