

# Innovations in Data Dissemination at the Central Statistical Agency of Ethiopia

## Abstract

The Central Statistical Agency (CSA) of Ethiopia is responsible for providing accurate and timely statistical information for development planning and monitoring purposes. To fulfill its responsibility CSA has been engaged in utilizing Information and Communication Technology (ICT) to facilitate its data processing, archiving and dissemination system so that the required statistical information can be generated and reaches the users.

CSA is considered to be one of the leading institutions in Ethiopia in utilizing ICT to accomplish its basic tasks. The CSA started its computerized statistical data production by utilizing the IBM System/3 with 12k CPU. Currently the agency is handling its statistical data archiving and dissemination through high capacity servers and a reliable network infrastructure. The paper based dissemination and restricted access to the CSA data has undergone significant improvements. The DDI application has improved the metadata documentation as the CSA metadata is now archived and disseminated to an internationally accepted standard. Utilization of GIS in providing easy access to geographic data for decision makers has shown a significant improvement as well.

## Keywords:

Data Archiving, Data Dissemination, Geographic Information System (GIS)

## 1. Introduction

### 1.1 Country Profile

Geographically, Ethiopia is situated on the horn of Africa between 3 and 5 degrees north latitude and 33 and 48 degrees east longitude with a total area of 1.1 million square kilometers. Ethiopia is administratively sub-divided into nine regional states and two city administrations.

The topographic features of the country range from the highest peak at Ras Dashen, (4,550 meters above sea level), down to the Afar depression at 110 meters below sea level. The climatic condition of the country varies with the topography, ranging from 10 to 47 degrees Celsius. Moreover, Ethiopia is a home to about 80 ethnic groups that vary in size from less than 1000 to more than 18

by *Yakob Mudesir Seid*<sup>1</sup>

million persons. The third Population and Housing Census conducted in 2007 reveals that there are about 74 million inhabitants in the country. Among these the majority (84%) live in the rural areas. Ethiopia is the second largest country in Africa in terms of population size with diversified cultural, linguistic and ethnic compositions.

### 1.2 Rationale for Improving the Data Archiving and Dissemination System

We are in the era of globalization and, in order to properly plan progress and monitor development, every piece of information useful for policy making and strategic socio-economic decision should be readily available at the country, regional and international levels. Indicators such as the Millennium Development Goals (MDG), information on the country's Sustainable Development and Poverty Reduction Program (SDPRP), or data on the country's socio-economic characteristics should be easily accessible and available in a timely manner.

The Central Statistical Agency (CSA) of Ethiopia, as the official national data provider, has been producing, analyzing and disseminating information obtained from surveys, censuses and administrative records. This activity of the Agency started in 1960, when the then Central Statistical Office was established under the Ministry of Commerce, Industry and Tourism. Since then the office has passed through different organizational restructuring and the existing shape of the Agency was created in 2005 when the office was re-named the Central Statistical Agency of Ethiopia and became accountable to the Ministry of Finance and Economic Development.

Like statistical offices around the world, the Agency is mandated to conduct socioeconomic surveys and compile national statistics from administrative records. The agency has also been given a mandate to coordinate the national statistical system (NSS) of the country and be involved in statistical capacity building activities for various Ministries, Departments and Agencies. Since the launching of its Integrated Household Survey Program (IHSP) in the 1980's, the agency has been able to conduct about 11 different surveys annually. Moreover, the agency conducted three consecutive Population and Housing Censuses (1984,

1994 and 2007). The agency also conducted the first ever Agricultural Census in 2001/02.

To be an efficient statistical organization and to accomplish its goal of becoming a recognized statistical office in the region, the capacity of the activities of Agency has to be strengthened through the effective use of existing Information Communication Technology (ICT) for its data collection, data processing, analysis and dissemination. This implies proper utilization of ICT such as Internet, Geographic Information Systems (GIS), and electronic methods of data archiving and dissemination.

In addition, provision of data also involves harmonizing and integrating statistical data, filling the gap between data produced and data available, laying down efficient ICT infrastructure, improving the quality and comparability of data, solving the challenges emerging from data and metadata exchange and harmonizing different standards with the data management system. All these need the strengthening and proper utilization of ICT at the CSA.

Recognizing these facts, the CSA's top priority has become improving its data collection, management and dissemination system by making use of ICT. Accordingly, the Agency has given a higher profile to ICT in its organizational structure and set a vision towards the improvement of ICT capacity that pursues different action plans. One of these action plans involves setting up an IT based dissemination system for its survey microdata and related metadata. This involved the following:

- Development of an integrated Central Data Bank of survey and other data and an Ethiopian Socio-Economic Database for basic indicators;
- Development of database management systems;
- Website Development;
- CD-ROM publishing;
- A Comprehensive Program of Documentation of existing and new data especially data related to socio-economic indicators;

## **2. Historical Development of the ICT Infrastructure at the CSA**

### *2.1. The IBM Series:*

The modern data processing activity utilizing the ICT products at the CSA can be traced back some forty-three years. The first computer system used was the IBM System/3 with 12k CPU which accepted data from 96 column punch cards.

The CSA was utilizing these types of machines on a rental basis with a monthly rental value of 2,498 USD which was increased to 7300 USD in 1980's. Data from price surveys and small socioeconomic surveys were processed by this

system. It is recorded that about two million punch cards were utilized per annum to transcribe the data collected through survey questionnaires to process the data from surveys conducted at that time. The IBM machines utilized 1200 feet and 2400 feet tapes to store the electronic data

Evidences reveals that the CSA was one of the few institutions with this "fancy" technology at the time and various requests to use these machines for various purposes were made, including from the Economic Commission for Africa.

This IBM system was utilized for about sixteen years with its 12k CPU which is really difficult to imagine at this stage of development.

### *2.2. The Main Frame System*

The CSA migrated to an HP main frame system in 1982. This was realized by the financial assistance obtained from the UNDP and the machines were supplied by a Paris based company called SERIC. The initial cost for the HP3000/ Series 44 system was 241,433 USD and the system comprised the following:

- HP 3000/Series 44 system processor unit with 1MB memory
- Two 404 MB Removable Media Disc Drives
- Two 1600 bpi 45ips Tape Drives
- Two Line Printers 300 LPM
- COBOL, Fortran and RPG Basic compilers

This was a great improvement compared to the 12k CPU IBM machines and CSA has utilized this system to process various surveys including the 1984 Population and Housing Census. The HP3000/Series 44 system was upgraded to a HP3000/series 48 and this system then upgraded to a HP3000/Series 925 in 1989.

### *2.3 The Stand-Alone PC System:*

The introduction of a PC based system was not realized well until 1994 when the CSA was busy undertaking the second population and housing census. Documents reveal that the first set of PCs available in late 80's were:

1. IBM PS/2 MOD 80
2. HP VECTRA ES/12 MOD 21

During this time, there were only five PCs dedicated to senior programmers and access to those PCs were very limited.

The CSA decided to process the 1994 population and housing census on a PC system and the issue of abandoning the HP3000/Series925 main frame was seriously considered by the management. This did not receive a very warm

response from the data processing experts at that time, who feared the new environment. However, with financial assistance obtained from UNFPA more than 90 486 DX PCs were introduced for census data capturing. In addition, fifth generation computers with Pentium processors were used for the data processing activities of the census.

This exercise using a PC based system for census processing encouraged the CSA to implement PCs for all of its survey processing. Accordingly all of CSA's survey processes were migrated to a PC based system in 1995 and replace the HP main frame system. .

### 2.4 Network Environment

The PC system used until 2004 at the CSA was a standalone system and as a result resource sharing and efficient communication was a serious problem. Utilization of 1.44 MB floppy diskettes was considered an efficient means of transferring files or documents among professionals. In addition, there was no centralized management of the system which hindered the data security and management system. Thus arose the necessity of establishing a Local Area Network (LAN) and CSA established its LAN in 2004.

The CSA network at its inception connects the users at the two venues to a centralized data center and to the internet. The connection of the two venues was made with fiber optic cable using 100Base-FX standard.

The network that the CSA has currently is one that improved on its structure from the previous network setup.

The new network system in place also facilitates the back up system and these days the CSA is using HP-DL 380 G5 (Generation five) Tape drives and a HP Ultrium data cartridge with a capacity of 400 GB

### 3. Framework Set at the CSA to Improve the Dissemination System

To accomplish the above mentioned activities, the IT framework has been designed in accordance with international metadata recommendations and best practices in data archiving to facilitate data dissemination and metadata exchange at the global level. It has the following basic structure.

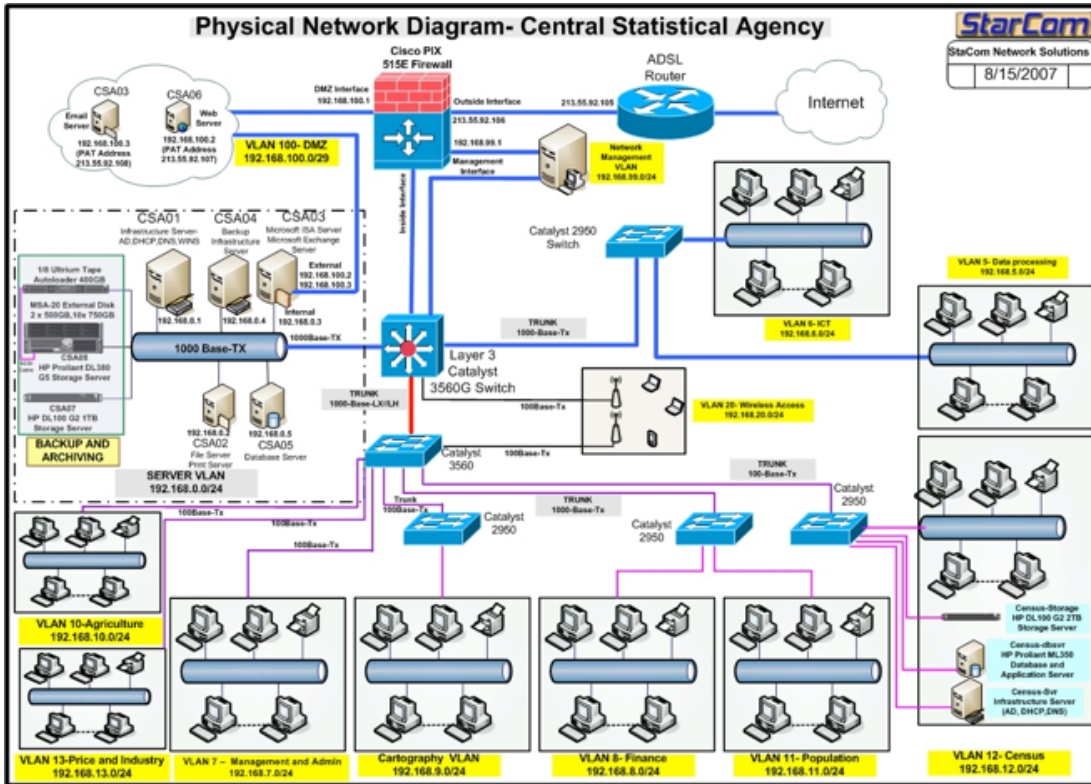
As shown in the diagram 2 (page 29) , the IT infrastructure is based on users/producers needs and has a data management framework which takes into consideration the establishment of task performing bodies, IT core architecture and data archiving and dissemination tools.

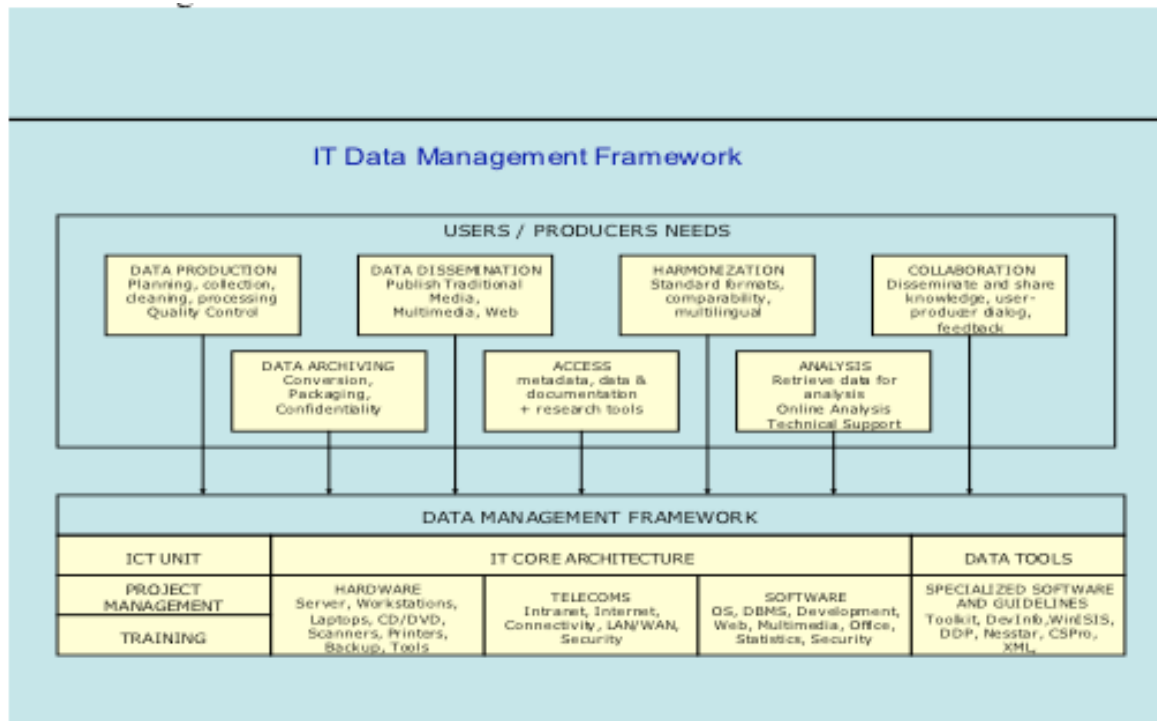
This IT based archiving and dissemination system is made possible by the establishment of a central databank to archive all documentation and microdata obtained from surveys and censuses and develop a user friendly system for its dissemination so as to provide easy access to the Agency's data for users. This in turn will help establish a

one-stop data shop for our users.

To achieve this, the project calls for the adoption of the standards and specialized data management tools. This includes specifications such as DDI, Dublin Core and SDMX and the use of tools like the International Household Survey Network's Microdata Management Toolkit and the UNICEF DevInfo package.

The data management framework is being put in place





from their success and failures and convinced staff that scanning could be applicable for the census data processing. Moreover, two types of pilot censuses were conducted using both conventional and OMR scanning of questionnaires with one scanner. The experience gained during the pilot censuses helped to convince the

takes into account global recommendations and best practices in microdata management and plans for step by step improvements to support high quality online access to metadata and aggregated data, online analysis and microdata access services.

#### 4. Utilization of New technologies for Data Capturing

##### 4.1 The Scanning Technology

One of the time consuming process in census data processing is data capturing. For instance, the data capturing activity of the first population and housing census of Ethiopia took about two years to complete using a main frame system and the second census took more than a year using a stand- alone PC system involving 180 data entry clerks using 90 PCs.

In order to capture the data and provide the result of the census as quickly as possible, scanning technology was implemented for the 2007 Population and Housing Census. The actual data capturing for the census of sedentary areas (except Somale and Afar Regions) was started on mid July, 2007 and the scanning work was completed in November 2007. This shows more than 95 percent of the population data was captured within four months.

Before implementing the scanning technology for the third Population and Housing Census of Ethiopia, information was obtained from the study tours made to National Statistics Offices in Tanzania and Ghana who utilize scanning for their Population and Housing census. This assisted the Central Statistical Agency of Ethiopia to learn

CSA that the technology could prove to be appropriate and effective for census data capturing, if properly planned and the necessary precautions could be taken.

The 2007 population and housing census of Ethiopia utilized 11 PS900 iM2 Scanners and printing and designing of census questionnaires in the local language was done by the supplier of the scanning machines. Technical assistance and on the job training was undertaken by the supplier and contributed to the success the new technology. The PS900 iM2 scanner used had the following features

- The ability to scan up to 8,000 Double sided Forms per hour
- Simultaneous data validation
- Bitonal & /or Greyscale Images & Clips (200dpi)
- 3 programmable output hoppers
- SOSKitW software utilities

The scanning process put in place for the census involved three major procedures, namely:

- Scanning / Data Capture
- Key-correction or validation of scanned data
- Export of the scanned and key-corrected data into ASCII or Text format suitable for processing

Major benefits obtained from scanning



- Significant decrease in time required to capture the data and minimization of non- sampling errors generated during data entry
- No need to store millions of forms for possible future reference since scanning captures the whole content of a questionnaire in an electronic format

Requirements for the effective utilization of scanning technology:

- Proper training in the use of both hardware and software to assist staff to “ own” the new technology
- A reliable Network System
- A well organized space for forms and data flow
- Proper file management and care in :
  - Checking batch (EA) IDs and orientation of forms
  - Proper recording of the in-coming and out-going questionnaires
  - Ensuring the consistency of EA codes and the quality of EA database
  - Close attention in detecting errors in the scanning process
- Ensuring that there is proper paper throughput

through the scanner

- Ensuring smooth running of the machines

#### 4.2 The PDAs

The CSA has deployed PDAs (Personal Digital Assistants) to facilitate the price data collection and electronic transfer from the field to its head office. The previous (?) price data management process at CSA was reviewed and compared with the new PDA system. The system satisfies both the statistical standards’ requirements for CPI and PPI calculation and also timeliness requirements for price data users. The full deployment of PDAs to 119 retail price markets in all 25 CSA branch offices should reduce data processing time by three weeks. In the first implementation phase price data from five CSA branch offices was transferred to head office within two working days after collection. With full deployment of PDAs CSA price data can be on-line within the same month in which data are collected. Excel PDA version software was customized for the application for price statistics, including the data entry form, with min-max validation tools to reduce data entry errors and the introduction of Amharic fonts. A PDA user guide has been fully developed in English and translated into Amharic. A total of 138 PDAs are being deployed (including back-up devices to replace those requiring maintenance).

The traditional paper based data collection system is being fully replaced by a digital system. In order to further reduce data processing time an automated excel spreadsheet has been developed. This facilitates market price data cleaning and further reduces the time from data collection to data dissemination on the web. A manual on the utilization of the automated spreadsheet has been developed for CSA subject matter specialists. Data quality is a critical issue to meet the statistical objectives for PPI and CPI calculation. Three comparative tests of data entry on PDAs by enumerators and desk top computers by data entry clerks have shown that data quality between different systems is comparable. The tests showed that data quality improves when the data is managed in conjunction with



data generation in the field.

## 5. Achievements with regard to the Dissemination System at the CSAA

In order to improve its data archiving and dissemination system described in its IT framework above, the CSA has been engaged in various IT related improvement activities. The following are the main achievements so far related to the IT based archiving and dissemination system since 2004.

### 5.1 Establishment of the Central Databank

A Central Databank has been established for the microdata and contains over 5000 data and documentation files covering about 71 surveys. Moreover, more than 48 surveys have been archived using the World Bank Microdata Management Toolkit, making the metadata compliant with the Data Documentation Initiative DDI-XML specifications, as recommended by the International Household Survey Network (IHSN).

### 5.2 Website Development

The CSA was able to develop its official website and use this website not only to provide adequate information about the Agency's activities but also to provide statistical data to the users. Most importantly, the CSA is using its website to:

- Disseminate national statistics and monthly CPI figures
- Provide access to all documentation

and related metadata for all of its surveys archived in the central data bank and

- Serve as a portal for other access points for CSA's data, like the ETHIOINFO database, the ENADA system, and the Price database

### 5.3 CD\_ROM Publishing

CD\_ROM products have been prepared and a CD based electronic dissemination system has been put in place. Accordingly the CSA is able to disseminate its survey results, documentation and metadata archived using the IHSN toolkit on CD-ROMs

### 5.4 The EthioInfo Database

EthioInfo is a customized adaptation of DevInfo, widely used user friendly software that helps to organize and present data in a results based environment with unique features linking to strategic monitoring and evaluation of policies such as MDG and National Poverty Reduction Strategies. As a common platform for indicators related to Human Development, EthioInfo facilitates indicator harmonization at global, regional and country levels by making statistics available to a wide audience. It allows



presentation of data through Tables, Graphs and Maps.

### 5.5 The ENADA System

The ENADA system is a web based application that allows the CSA to catalogue its data and release the survey metadata (information on the survey) into the public domain for data users to browse. The ENADA system is unique not only because it provides the tools

commodity price data in a given market in a particular year and month. This online system provides easy access to time series price data collected by the CSA.

### 6. Introducing the GIS system

Geographic Information Systems (GIS) is a collective term commonly used for computer systems that manipulate geographic data. These systems are implemented with

computer hardware and software functions for the following archiving activities of geographic data

- Acquisition & verification,
- Compilation,
- Storage,
- Updating & changing,
- Management & exchange,
- Manipulation,



to comprehensively catalogue the entire history of surveys conducted by the CSA using the international documentation standard known as the DDI, but also it provides full control of the microdata and allows implementation of a dissemination policy in a systematic and formalized manner. In fact, the data catalogue of the CSA becomes an effective portal into the rich treasure of survey data that the CSA archives without releasing any confidential information.

As explained above, at the heart of the ENADA system is the searchable catalogue. This search function allows a user to browse the entire catalogue down to the variable level using keywords across all the 71 surveys loaded on the system so far. The data user or researcher can then determine on the basis of the search results which information would be required to carry out their analysis and can proceed to apply for the microdata according to access policies set up by the CSA.

### 5.6 The Price Database

Price data dating back to 1997 are available online via the CSA's website. This assists data users to have access to

- Retrieval and presentation and
- Analysis & combination

The major advantages of GIS are:

- The ability of integration of data from many different sources.
- Identification of the spatial relationships between map features.
- integration, analysis, and visualization of the data;
- Organization of data for more sensitive and intelligent decision making. Showing patterns and trends; and finding solutions to problems.
- Linking of dissimilar data like Census, Demographic and Health Survey, Agricultural survey, Industry, Cartography data

Realizing the above facts, the CSA has been heavily engaged in utilization of GIS and in the last two years the following major activities were undertaken



:- The Production of an Atlas of the Ethiopian Rural Economy in collaboration with the Ethiopian Development Research Institute and the International Food Policy Research Institute

- The creation of a digitized Wereda (district) map that shows each EA within a Wereda including the following features

- Urban/Rural Kebele(dweller association) Boundaries

- Cultural and Natural Features

- Educational/Health Facilities

- Religion Centers

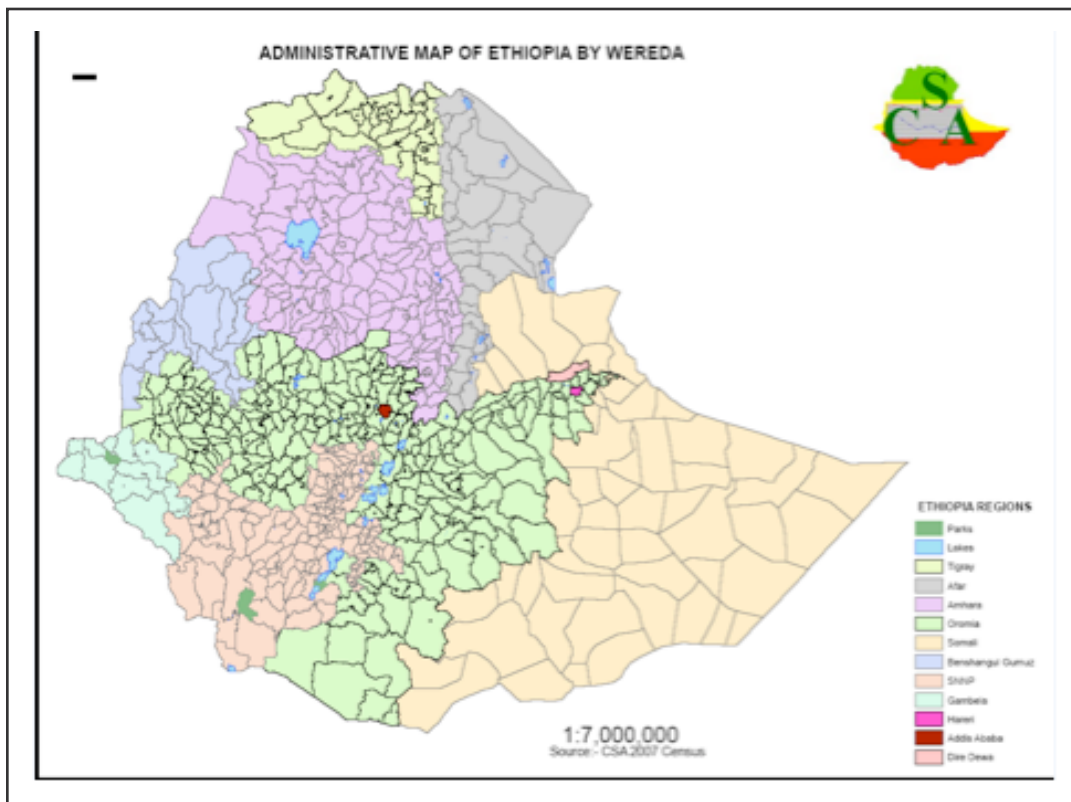
- Localities within the Wereda

- Roads of any type

- Rivers

- Rail ways

Rural Facilities Atlases are also being developed. These atlases use the data obtained during the cartographic work for the 2007 population and housing census. It includes a special distribution of basic facilities in the rural areas of the country. Among others, these facilities include schools, health posts, water points agricultural

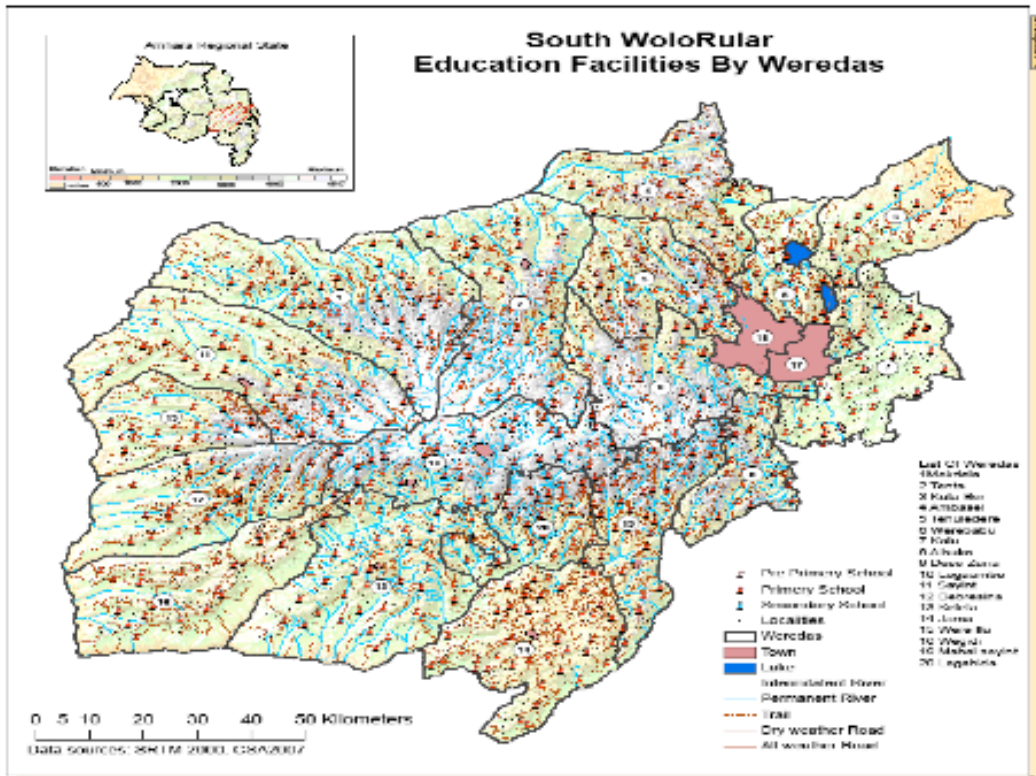




development offices and roads. These atlases are produced for all districts and contain rich geo-referenced data for development planning.

The atlases aim primarily to report at district level to give accurate information on the spatial pattern of basic facilities and services within the boundaries of each district. The atlases will also be of importance to the academic community in providing students and various researchers' adequate information on the level of services and facilities in rural parts of Ethiopia. Different areas of social science can find in these maps interesting topics to explore the development level of rural parts of the country and they will be provided with a good understanding of rural Ethiopian society's basic infrastructure.

### 6.1 Satellite Imagery



In 2000 the UN Statistics Division published a Handbook on Geographic Information Systems and Digital mapping to assist countries in their preparation for census activities and in particular in adopting GIS technologies for census mapping operations. The traditional role of maps in census and survey operations is restricted to supporting enumeration and the presentation of results in atlas form. However the accuracy, content and timeliness of satellite imagery makes it useful for pre-census activities such as ensuring consistency and facilitating census operations.

During the actual census these maps support data collection and monitor census activities and at the post-census stage they makes it easier to present, analyze and disseminate census results in the form of an atlas that will show the complete picture of areas of interest. the UN's belief is that adopting a digital approach through the use of satellite imagery will result in planners, policy makers and public administrators being able to make better-informed decisions and this is fully endorsed by the CSA. With multiple years of imagery collection, archived data will provide a large proportion of the required coverage from pre-existing data over a period of years ensuring that a viable product can be produced within a relatively short time period.

One of the advantages of satellite imagery is its political immunity and non-invasive sensing of any geographical area regardless of the conditions on the ground that

might otherwise dictate its accessibility, such as natural hazards, remoteness or hostilities. Whilst satellite imagery cannot measure population levels directly, it can help to identify populated areas for census planners to reduce the effect of pastoralism on the final figures.

Considering the advantage that these satellite images provide for the census, the Agency decided to utilize the 5 meter resolution SPOT 5 and 50 cm resolution Quickbird imageries for delineating Enumeration Area (EA) Boundaries in pastoralist areas. This was the first time in the history of the census in the country where pastoralist areas were delineated using EA maps like other sedentary

areas.

### 6.2 Multiple uses of the Satellite Imagery

As mentioned earlier, it is anticipated that such a product would be a valuable national asset, able to inform a whole range of other national, regional and local governance activities, as well as provide input to other plans such as support for food security, provision of health services, planning of housing, water supplies and transportation infrastructure to support the urbanization process in



Ethiopia's economy, the country lacks a reliable and up-to-date land cover database and an area sampling frame. These data are the baseline for agricultural censuses and are essential for generating reliable agricultural production estimates and forecasts. Crop acreage estimates are particularly

Ethiopia. Current mapping products are a fundamental part of creating a sustainable economic and social society as well as helping to address civil instability by providing accurate information about the natural and manmade environment.

Despite the importance of the agricultural sector to

important information for national institutions such as the CSA and the MoARD. Field surveys were carried out by the GOE in coordination with the FAO and WFP (Crop and Food Supply Assessment Missions - CFSAMs). The project is assisting the Government of Ethiopia (GoE) in the implementation of a robust statistical methodology



using the latest technology. Before the start of the project, the GoE acquired Landsat and SPOT 5 meter resolution satellite images for the entire country supplemented by Quick Bird/Ikonos images at 0.6 meter resolution for all urban districts. This investment provided an opportunity for the development of cutting edge approaches to improve agricultural statistics as well as construction of a critical land cover database and its sampling frame.

**References:**

[www.ihsn.org](http://www.ihsn.org)

[www.drs.org](http://www.drs.org)

[www.csa.gov.et](http://www.csa.gov.et)

**Note**

Yakob Mudesir Seid, Central Statistical Agency, P.O.Box 1. 1143, e-mail:yakobu@ethionet.et, Addis Ababa, Ethiopia.

The article is based upon a presentation at the IASSIST May 2009 conference in Tampere in the session on "Building Data Archives and User Communities: Greece, Estonia and Ethiopia".