THE SOUTH AFRICAN CENSUS 2001 SPATIAL INFORMATION SYSTEM DATA CAPTURE PROBLEMS[*]

Helena MARGEOT and Sewnath RAMJITH, South Africa

Key words: Census GIS, Census Mapping, Census data capture problems, spatial information system.

INTRODUCTION

October 10, 2001, the official census day in South Africa, marks the culmination of 18 months of intensive planning by Statistics South Africa (Stats SA) to conduct a census. A Geographic Information System (GIS) provided the backbone for the entire process and this paper will attempt to describe some of the technical process and highlight some of the practical limitations that became apparent during the life cycle of the project.

BACKGROUND

South Africa is a country with a total population of 41 million people (1996) living on a total surface area of some 1 220 000 sq km. The initial transition to democracy in 1994 resulted in the formation of 9 provinces. The final phase of this transition process was the creation of new local government structures, which was completed in 1999 when an independent body, the Municipal Demarcation Board (MDB) demarcated the country into 6 Metropolitan Councils, 47 District Council, 25 District Management Areas and 231 Local Councils.

POPULATION CENSUS IN SOUTH AFRICA

A *population census* is the process of counting the number of people, at a given point in time in a country, and collecting information about their demographic, social and economic characteristics. After data collection, the process includes the processing, analysis and dissemination of the information collected. *"The purpose of official statistics is to assist organs of state, businesses, other organisations or the public in planning, decision making or other actions; monitoring or assessment of policies, decision-making or other actions."* This extract from the South African Statistics Act No 6 of 1999 emphasises the need for an accurate and thorough coverage of the country and Stats SA has been given the responsibility of conducting a national population census every five years. This census data provides a persuasive and important tool for decision-making and for facilitating sustainable development in the country.

[*]) *The views expressed in this paper are those of the authors and do not necessary reflect those of Statistics SA.*

CENSUS 2001

Before enumeration for a Census to be undertaken, the country must be divided into small pieces of land; this process of dividing the country into these areas is called demarcation. The main objective of demarcation is to subdivide the country into small Enumeration Areas (EAs), each of which will be assigned a unique identification code on the basis of the country's administrative divisions.

Planning for Census 2001 began in November 1999 when the Minister of Finance approved funds to conduct a census in 2001. It was understood that an accurate and thorough census was needed which meant structuring the various census processes and verifying and correcting all EAs, nationally. Equally important was the need to perform the task as economically as possible within a very narrow time frame of 18 months. It was also decided to replace the old census methodology of using Photostat copies of 1:50 000 topographical and municipal maps, hard copy aerial photographs or sketches of the EAs as a base for enumeration with a GIS as the framework for producing high quality, accurate Enumeration maps. This spatial information system would correct the deficiencies of the former spatial data, minimise the volume of fieldwork required in the EA demarcation process and be used as a decision making tool by census data users in South Africa.

CENSUS 2001 PROCESSES

To successfully create and utilise this GIS for Census 2001, 5 key steps had to be implemented.

1. Commissioning a GIS. Central to a GIS is the acquisition of hardware, software and an operation system. The GIS was commissioned in June 2000 following a very rigorous tender process which involved benchmarking several solutions. The GIS consisted of a UNIX server with Oracle Spatial Cartridge ver. 8.6 as a database engine. Initially 1.7 terabytes of disk space was reserved for data storage, however this has been upgraded to 2.7 terabytes and will be upgraded to 10 terabytes within the next six months. Intergraph's Geomedia and Geomedia Professional version 4.0 was chosen as the front-end GIS software running on NT4.
Diagram 1 illustrates the GIS operational structure, starting with data obtained from various sources to assist in the Census processes, followed by the demarcation and data capture processes and the structures put in place for the analysis and dissemination of the Census data both at Head Office and Provincial levels.

2. ***Data Gathering.*** The 1996 Census database was used as the starting point to work from for Census 2001. The spatial database was captured after Census 1996 took place, therefore the 83 126 spatial EAs captured were imperfect. There was need to rectify inherent deficiencies and to update EA boundaries, place names and attributes of the various geographic entities.
To verify the and standardise past demarcation, relevant vector and raster backdrop data was obtained and assembled from several government departments, metropolitan councils, parastatals and private organisations. Only data that was deemed useful in facilitating this

process was obtained. As examples, EA boundaries do not cross natural features or administrative boundaries, so it was necessary to obtain Administrative boundary and topographical data. Coming from diverse sources, this data also varied in accuracy and format, which was problematic.

At the completion of the data gathering, a workflow process was designed and standards were set to ensure the correct datum, projection ellipsoid and data quality were adhered to. Finally, detailed metadata was captured for each data set.

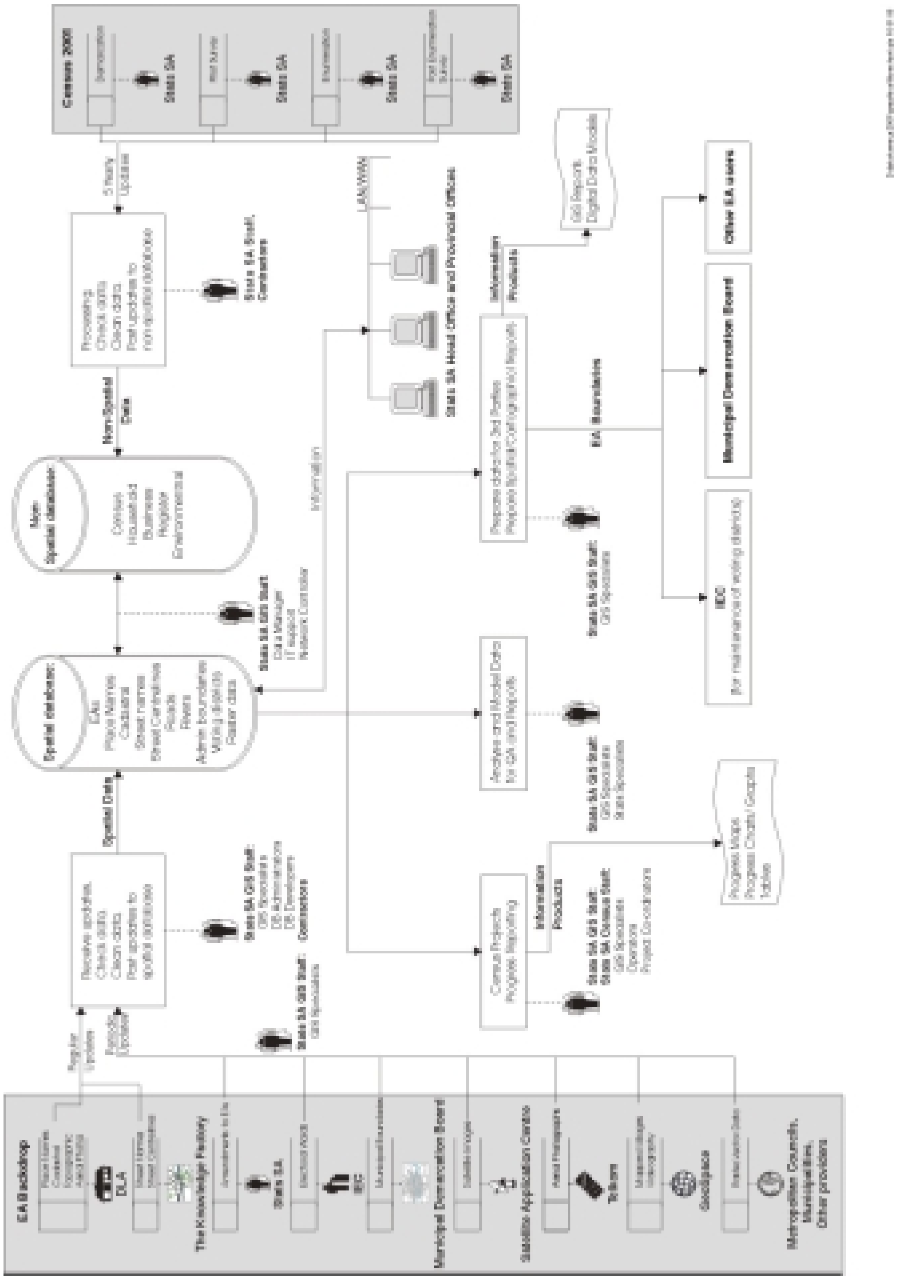


Diagram 1: Statistics South Africa’s GIS Operational Structure

3. Demarcation. The data assembled in Step 2 was then used to correct and re-define the 1996 EA boundaries, using strict demarcation rules and procedures. Stats SA staff were trained and the re-demarcation process was done at provincial level on paper maps, using the 1996 EA data and supplementary backdrop data available.

The process of evaluating the existing 1996 EA dataset started with office demarcation. The existing EA boundary was compared against all available backdrop data. A hardcopy colour A3 map of each EA was produced and the EA was assessed on screen in relation to the adjoining EAs. A count of visiting points/dwelling units was undertaken to verify that the count was within the norm and boundary alignment was verified. Recommendations as to whether the EA was acceptable, required field verification by the provincial staff or required additional imagery to be collected were made.

4. Digital Capture and Quality Assurance (QA). Stats SA did not have the capacity to complete this task on time for the census; as a result this step was tendered. Once the corrected and re-defined EA boundaries were captured digitally, a QA process was introduced to verify that the capture process met the standards laid out in the tender.

5*. Integration of Census 2001 in Database and Enumerator Map Production.* The final process was to integrate the 80 788 spatial 2001 EAs into the data base and produce Enumerator maps to be issued to enumerators for the census.

PROBLEMS ENCOUNTERED IN THE STAGES OF CORRECTING THE 1996 CENSUS DATABASE

1. Vector Data

Vector data was gathered from the following government and private sector sources (Table 1: Vector Data Sources and Coverage).

Datasets representing boundaries were of particular importance since one of the fundamental rules of EA demarcation is that an EA should not cross administrative or social boundaries. The land parcel data for the whole country obtained from the Surveyor-General (SG). was not always clean or topologically correct. Due to the manner in which this data was captured and stored (it was captured in haste to be used for election purposes in 1996), positional accuracy of boundaries was often problematic. Since each land parcel was captured individually from its relevant source diagrams, a common boundary between two adjacent land parcels could be duplicated as it was derived from source data of different degrees of accuracy and currency, resulting in two different lines that were not co-incident representing this common boundary. It was not the mandate of Stats SA to correct any of this data and the SG has since initiated its “Clean Sweep” operation to clean and structure its database. Unfortunately, the cleaned dataset, was not available during the life cycle of this project.

Furthermore, time did not allow for verification and comparisons between the different datasets. This resulted in datasets from different organisations, with varying degrees of accuracy and currency being simply housed as separate layers.
The various vector datasets thus obtained, were converted to WGS84 Lat/Long and stored in a separate warehouse in GeoMedia format.

Complicating matters further was the fact that these vector datasets were then extracted per local municipality (the working unit) from the Oracle warehouse and handed to the contractors in a GeoMedia MSAccess format to use in the EA capture process. Also, no records were kept of which dataset combination was used during demarcation, capture and printing; this led to different datasets being used by different contractors for the different phases of the project.

SOURCE	DATA TYPE	% COVERAGE
Surveyor General	Cadastre	Total RSA
Chief Directorate: Surveys and Mapping	Topographic Data Place Names	Total RSA
Municipal Demarcation Board	Municipal Boundaries	Total RSA
Independent Electoral Commission	Voting Districts Voting Wards	Total RSA
ESKOM **(Electricity Supply Commission)**	Tribal Village Boundaries Tribal Village Erf Boundaries	Former Homeland Areas
National Department of Agriculture	Farm Ownership Farm Boundaries	Total RSA
Metropolitan Councils **Town Councils**	Cadastre Road Centrelines Suburb Boundaries Institutional information	Metros, Localised Towns
Human Sciences Research Council (HSRC)	Place Names Schools/Police Stations	Total RSA
Private Enterprise	Suburb Boundaries Road Centrelines Farm Ownership National Address Register	Metros and 80 Main towns
Statistics SA	EA Boundaries Institutional Information Attribute Data per EA	Total RSA

Table 1: Vector Data Sources and Coverage

2. Imagery

Having high quality current backdrop data (Aerial Photography, Videography and Satellite Imagery) is of immense value when it comes to demarcating EAs. The following tables

(*Tables 2 and 3*) and map (*Map 1*) illustrate the type and coverage of raster data as actual area coverage.

Photo Source/ Scale	No. Images/ Map Sheets	Coverage – Km^2	% RSA Area	Purpose for Acquiring
Spot Images	128 scenes	474 525	**38.90**	Change Detection

Table 2: Coverage and Use of Spot Ortho-Rectified Images for Change Detection for Census 2001

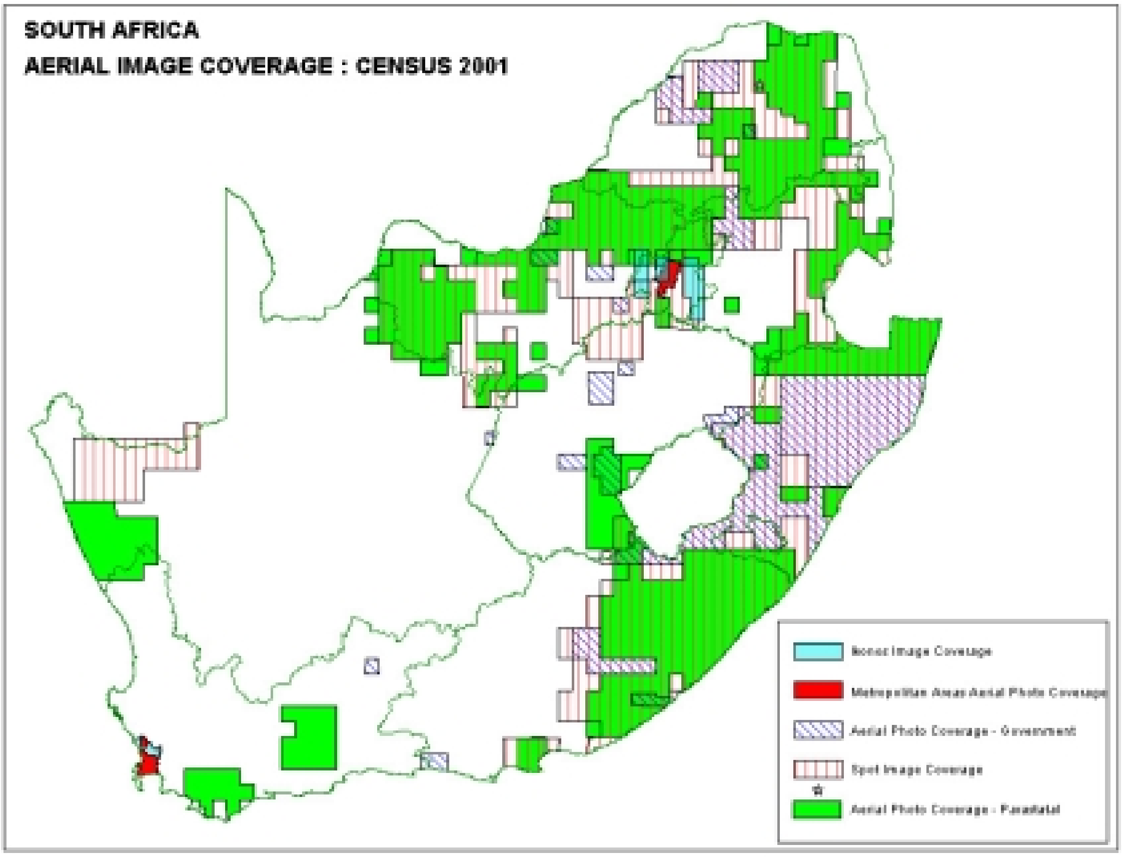
Photo Source/ Scale	No. Images/ Map Sheets	Coverage – Km^2	% RSA Area	Purpose for Acquiring
Ikonos	232	9 000	0.74	EA Demarcation/ Backdrop
1:10 000/12 000	332	4 102	0.33	“
1:20 000	1772	299 025	24.50	“
1:30 000	555	93 656	7.70	“
1:32 000	1275	34 425	2.90	“
Videography 1m/2m pixel size	7721	Not Known	RSA Squatter Areas	“
TOTAL			**36.33**	“

Table 3: Scale, Coverage and Use of Ortho-Rectified Imagery for Census 2001

Aerial coverage concentrated on areas of high population density, areas where population change occurred rapidly (squatters), and areas where there was minimal backdrop and no cadastral data (former tribal areas and homelands). The imagery used was not older than October 1999; the most recent sets were the satellite images and the videography (not older than November 2000 – July 2001). Table 4 gives an indication of the coverage with regard to total EAs and therefore, primarily areas of high concentrations of population.

	Aerial Photography: Government	Aerial Photography: Parastatal	Aerial Photography: Metropolitan Councils	Ikonos Imagery	Spot Images Imagery (Change Detection)	Videography (for squatters)
E. Cape	14.6	85.9			94.6	9.0
Free State	53.2	18.7			22.1	9.1
Gauteng		27.4	35.7	29.1		6.4
KwaZulu	43.1	51.9			99.7	36.6
Mpumalanga	35.7	44.9		2	68.5	19.9
North West	10.4	71.5			98.4	3.0
N. Cape	.2	0.1			.2	33.2
N. Province	7.3	75.6			98.4	10.6
W. Cape	0.7	2.3	60			2.8

Table 4 : Aerial Backdrop Coverage as a Percentage of Total EAs per Province


SOUTH AFRICA
AERIAL IMAGE COVERAGE : CENSUS 2001
Ikonos Image Coverage
Metropolitan Areas Aerial Photo Coverage
Aerial Photo Coverage - Government
Spot Image Coverage
Aerial Photo Coverage - Parastatal

Map 1: Ortho-rectified Aerial Photographic Coverage for Census 2001

The bulk of the aerial photography was stored in WGS84 projected on Transverse Mercator projection with standard meridians at 2^{o} intervals on the odd meridian. The videography was stored in WGS84 non-projected Lat/Long format.
A complete set of scanned 1:50 000 topographic maps (1 978 sheets) covering the whole country was also stored in this latter format. These scans were used as backdrop where no aerial imagery was present for the EA demarcation and EA map printing processes.

Common Problems Encountered with the Imagery are as follows:

File Size. An average file size is 400mb per tile. A municipality on average required 20 such tiles, which meant having at least 8 gb of data on the workstation thus requiring a workstation with a large storage capacity. It was not possible to read the data over the network due to the large file sizes.
File Formats: Different contractors required the images in various formats as the speed of operation on different software varied. Therefore, MrSID, Intergraph JPEG, Intergraph with overlays, TIFF with its associated world files and GeoTiff copies of the same image had to be generated and stored on the server. The storage space on the Stats SA server was taxed to its limits and at times did run out of space as each process phase (capture, QA, printing) had to be supplied simultaneously.

Data was also supplied in 8 bit, 11bit, and 16bit Greyscales as well as 8 and 16bit RGB layers. Exchanging data between the different GIS packages became quite difficult due to the limitations of some of the GIS packages.

Image Catalogues. Fortunately ArcView and most of the imagery packages today allows for an image catalogue file to be used. This meant that the user did not have to know the name of the sheet that covers his current view, as the catalogue is able to calculate and display the correct image.

Image quality: Statistics SA, in conjunction with the Chief Directorate-Surveys and Mapping funded the acquisition aerial photography for large areas of South Africa (*see column 1 of Table 1*). Due to the large-scale nature of the project and the limited time available, seven aerial photography companies were contracted to fly different areas. However, this resulted in images of varying quality being delivered to Stats SA. Further, the weather patterns often did not allow for optimum flying conditions, so compromises were made at times, images of inferior quality had to be used in parts of South Africa.

Ortho-rectification of Imagery. Due to the scale of the project and the strict time schedules, compromises were made in terms of the accuracy of the geo-referencing of some of the Ikonos satellite imagery and the videography. Time did not allow for the supplied imagery to be checked consequently errors only became apparent during the data capture phase when shifts in the imagery were recognised. In an effort not to delay the process, the correction of the imagery was carried out in-house with available software.

Printing of Maps using Imagery as backdrop: Imagery allows for the production of very informative and practical maps. Using imagery causes certain problems with the processing and printing of maps as various GIS packages handle image file formats differently. When maps are printed using MrSid format images as backdrop, large amounts of disk space are required since the image is first expanded to TIFF format before it can be processed. This process slowed production in some software and the images had to be converted to geo-tiff to improve processing times.

3. Other Processes

Data Conversions

Several GIS and Image processing packages were used during the Census 2001 project; however, GeoMedia was the preferred GIS package of Statistics SA. One of the selected contractors had already developed in-house skills and applications in other software from previous contracts and to a large degree this package were used to capture the spatial changes and produce the maps for Census2001. It therefore became necessary to convert the vector data to the other GIS platforms prior to work commencing. A decision was made to export the data to MSAccess format per municipality. This process was not automated because the specifications of what dataset to export per municipality varied.

Attribute Data Tool

A tool was developed in order to capture the attributes relating to each EA consistently. Due to the complex nature of the land categorization in SA and the database design for the Census project, it was important that the correct fields be populated and the integrity of this data be verified. This tool allowed for these objectives to be accomplished. Substantial amounts of attribute data was gathered about each EA and allied to this several checking applications were then developed to test the integrity of the data and report on exceptions.

The one drawback of this application was that there was no link between the spatial and non-spatial components of the data while capture was taking place as the attribute capture process did not operate directly from the Oracle database but from an extracted MSAccess database. This process was implemented as the capture environment was not always implemented in GeoMedia. As a result, an operator could not click on an EA in the GIS and get the attribute data to pop-up in a window.

Spatial Capture of EA Boundaries

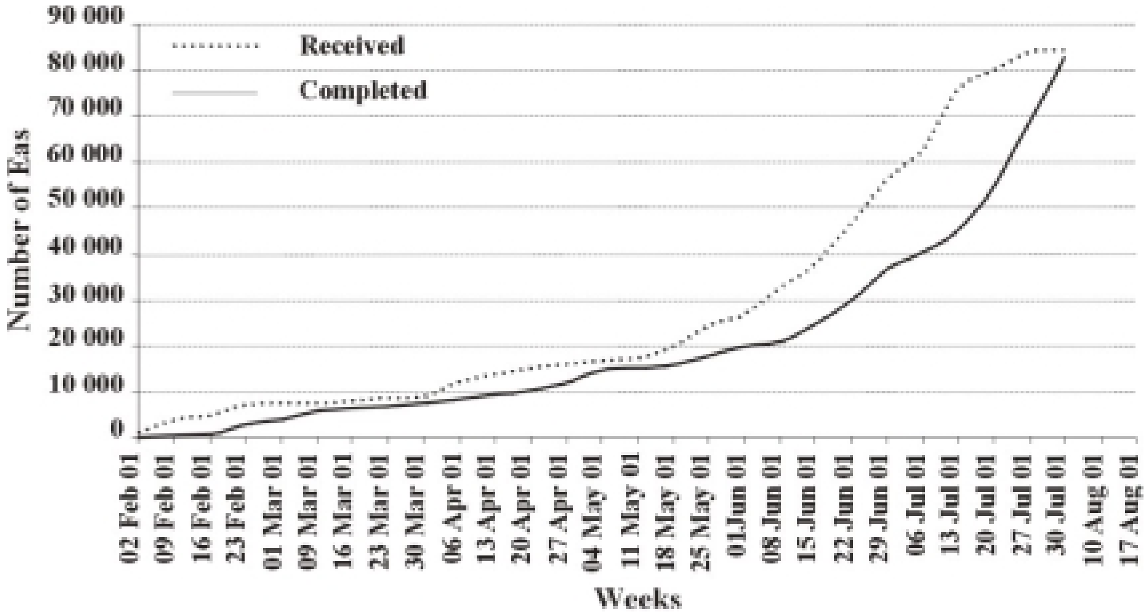
About 64.4% of the EA data was captured spatially using a specialised extension in ArcView. Once the data returned to Stats SA from the provincial offices in hardcopy format (maps of various sorts and attribute lists), the 1996 boundary data was edited and altered with all image and other vector backdrop visible on the screen. A drawback of ArcView 3.1 was that imagery from different projection settings could not be viewed simultaneously. The vector data had to be re-projected to Transverse Mercator or un-projected lat/long settings to deal with different image backdrops.

A specialised ArcView application was adapted for the capture process. It allowed for the easy snapping of lines and nodes to other vector backdrop coverage and had the ability to draw new “clean” boundary lines from existing lines and to snap to existing vertices on nodes. Once these changes were completed, each EA number was captured as a point feature. When the editing and capture process was complete, the final polygon data sets were built in ArcInfo. The spatial data was then compared to the attribute data, which was captured using a specific attribute data tool and the discrepancies were rectified. 64.4% of the EAs were captured in ArcInfo/ArcView and MSAccess database links and 35.6% were captured using GeoMedia and MSAccess database links.

Quality Assurance

This process was also outsourced by tender procedures.

Initially the scope of the tender allowed for the QA process to be done on a sample basis. However this was changed to 100% QA of all the capture as the demarcation rules were being applied in a subjective and varied manner. The capture of the attribute data with regard to place names and institutional information was sometimes incorrect and had to be corrected. Also, problems were encountered with scheduling the GPS and videography fieldwork to coincide with data capture delivery schedules. Again, weather was sometimes an adverse factor in allowing this work to proceed.



Graph Illustrating the Census 2001 QA Progress over Time

To complete the QA process on schedule, additional resources had to be implemented. The QA process itself had to be amended to allow for faster output. Some processes such as the populating of certain attribute fields were automated, dwelling unit counts were accepted without checking after July 2001 and the allocation of the same type of work to the same operators also speeded up production.

As a result of this effort, the QA process was completed during the week ending August 17 2001, with a total of 80 788 Eas for Census 2001 having been checked over a period of 45 000 hours by 40 contract staff.

CONCLUSION

A project of this magnitude requires careful planning and execution during all phases of its lifecycle. One of the greatest contributors to either the success or failure of a project of this nature is the human element factor. This is due to the subjective nature of demarcation and to the large number of people that are needed to be an integral part of this type of project.

GIS as a tool has proved to be of immense value when used effectively. Every part of the process has to be tested thoroughly so that all limitations could be identified and all technical problems could then rectified.

ACKNOWLEDGEMENTS:

To Statistics SA and the GIS management staff Sharthi Laldaparsad, Helene Verhoef, Jean Marie Hakizimana and Anne Marie Buitendach; Moses Sidindi and Francois Bezuidenhout in Census Mapping; To the contractors Africon (Francois Stander), Ikwezi (Brian Haupt), Data World, AfriGIS (Shanta Vadeveer), GeoSpace (Fanie Ferriera and Bernhard Jacob),

The Satellite Application Centre (Karen Wentzel) and the Chief Directorate: Surveys and Mapping (Don McIntyre and Michelle Denner)

CONTACT

Helena Margeot
Deputy Director – GIS
Statistics South Africa
Steyns Arcade
274 Pretorius Street, Private Bag X44
0001 Pretoria
SOUTH AFRICA
Email: helenam@statssa.pwv.gov.za

Sewnath Ramjith
Director
Data World (Pty) Ltd
364 Stamford Hill Road
Morningside
4001 Durban
SOUTH AFRICA
Email: ram@dataworld.co.za