The integration of statistical processes through IT support, methodological support and standards development

Pistorius, Marlize
Statistics South Africa
170 Andries Street
Pretoria, 0001, South Africa
E-mail: marlizep@statssa.gov.za

Arrow, Jairo
Statistics South Africa
170 Andries Street
Pretoria, 0001, South Africa
E-mail: jairoa@statssa.gov.za

Introduction

This paper presents the architectural concepts, tools and the methodology used for the implementation of an integrated information system for Statistics South Africa (Stats SA). The paper focuses on the integration between systems and standards developments and methodological support during the design and implementation of statistical production processes.

The focus will be on standardised sampling methodologies for both household and business surveys. A system design which can be either generic and a metadata driven database or a hybrid method between a normalised and metadata driven database will be examined.

An array of integrated IT systems tools are used to support and optimise the underlying statistical production processes. During the analysis phase data stored in SQL2005 are made available to the Statistical Analysis Software (SAS)® platform. The SAS® repository provides the ability to access data from all data sources from one application while being able to perform data analysis and publish from the same application. The entire SAS® platform runs on Virtual servers which ensure that access to the server are always guaranteed.

Edit and imputation rules are captured with a decision-table tool LogicGem. The rules are then converted to SAS code. Through the use of SAS Enterprise guide generic methodologies can be applied. Imputation, editing and non-response adjustments are carried out with help of CLAN. Statistical Macro Extensions (StatMX) is used for weighting and estimation. Finally certain Microsoft Office Add-Ins are used as a part of creating statistical publications. Dissemination of management information is allowed through the Web-Tier which is a presentation layer and it allows users to explore data.

The potential of integration can only be exploited through a holistic approach focusing on the integration of statistical data and the synergy of statistical processes using information and technology.

Traditional Statistical Information Systems

Owing to the fact that most surveys tend to focus on the variables within their universe, the information systems that are used as part of the statistical process are also generally built to consider only the variables in the survey. This is also known as the vertical design. The end result is slow but steady gravitation towards silos. This is despite the fact that some respondents are sampled across multiple surveys and the data required across these surveys is also common.

The side effects of this, perhaps immeasurable are immediately visible. The vertical design has negative impacts on the survey processes within a NSO but also provided the NSO with opportunities for improvement, in the past the following were observed:

- Although the information systems are capable of leveraging off collected data from other surveys they can't because of lack of transparency created by the silos. This increases the burden on the respondents. It also duplicates resources required during the collection process which invariably increases the cost of collecting a single piece of data (cost of data).
- Data confrontation and validation between multiple surveys on common variables is rarely done. This, perhaps, has the most impact because the same respondents when viewed over time might show varying reported values for the same variable.
- Processes and information system enhancements tend to be survey specific. There was very little cross pollination of discoveries and improvements made between surveys.
- Data structures varied between the different surveys. They were generally designed to support a specific information system. Downstream usage of the data was not considered. The information system and its database were tightly coupled at the expense of downstream processes Perhaps a coronary of this is that when statisticians work on surveys other than their own, they have to first understand the data structure. This overhead can be avoided.
- The development time of systems was long and a lot of development resources were allocated to maintenance work of older systems.
- Most of the processes were prone to human intervention.
- Missed opportunity of harmonising questions, standards and methodologies across surveys.

Generic Statistical Information Systems

The traditional design has been widely used in Stats SA. Over the past 3 years there has been a steady migration towards a more flexible design that addresses the issues experienced with the traditional design, a more service oriented architecture and design. The "generic" design is a design that uses a generic and metadata driven database. It is also known as the horizontal design.

The design hinges on/is underpinned by the fact that most surveys carry common objects, business and technical metadata definitions.

Because the over-arching structure is similar across **most** surveys, a metadata driven database structure built around common survey components creates a robust platform that allows new surveys to be provisioned/brought online with very little, if any, system development. Turn-around time for system development is, therefore, vastly improved. The generic design also introduces and enables processes to be

automated, but retains the checks and balances, thus reducing the burden on IT support.

The design offers more flexibility, higher information system functionality/component re-use, increased integration and reduced cost. The continuous re-use of previously designed functionality/component yields conditions of “economies of scale” on system development projects where each “unit” of software is produced a lot quicker and cheaper than during the last or first project. Overall, we are/have lowered the marginal cost of system development. The remaining cost in projects arises from integrating (assembling) previously developed components and testing their combined usage.

Functionality re-use is done in both information systems (capturing systems) and analysis processes (using macros). A formal System Development Life Cycle (SDLC) has been adopted and formalized to ensure that processes and procedures are adhered to, and all the necessary documentation is place. The underlying principle for such a development is standardisation of surveys processes, classifications, concepts and definitions.

Registers within Stats SA

Before a generic design can be applied, registers need to be in place. The registers are at the heart of the automation of processes.

The primary function for a register is to keep record of all the objects that could be used in the surveys. There are many types of registers including Master Sample for household-based surveys, Business Register (BR) for business surveys, Population Registers, metadata and variable registers (also known as data glossaries).

The Master Sample contain frequently updated listings of the primary sampling units (PSUs) which is made up of subset of census enumeration areas (EAs) defined in the population Census of 2001 and the BR consist of businesses that are registered for Value Added Tax (VAT) and Income Tax (IT).The BR data are sourced from government departments that have the mandate to maintain the registers. These data are the primary source for sampling and retrospective or reconstructive work that might be required in the future. The technology is in place to hold all the required data in electronic format, to provide the means to transfer data securely between agencies and systems within Stats SA, and to store and analyse data from different sources on comparable basis.

Metadata registers contain data about data. The metadata helps the consumer/user of the data understand it better or the processor (statistician/methodologist) process the data.

The registers can be included in the following:

- Business/Descriptive metadata – Data that describes to data/information consumers what the variables are and how they are derived/created. Turnover or Average number of individuals per household.
- Technical metadata – Data that describes when the variable was created, how the data is stored and in what format, e.g. Turnover: Datatype is Money and stored in Microsoft SQL in database X.

Variable register is a catalogue of all data elements, containing their names, structures (basic metadata), and information about their usage (which surveys use them).

A key requirement for object registers is that it should keep track of objects over time. As objects get retired, they should remain on the register but in a retired state. The object's metadata should show the commission date and decommission date. This necessarily implies that the format that the data are stored in needs to be able to cater for distinguishing active records from previous versions of the same record.

Some of the data collected during the collection phase is used to further improve the quality of the data contained in the registers. Certain criteria need to be met before data is used. Part of this is defining Master Data Management (MDM) processes that can clearly identify the surviving record/master record.

Sampling

In the traditional process, multiple survey specific samples are drawn and made available to the different surveys. This is possibly where the silos are created.

In business register every statistical unit is attached to a unique permanent random number uniformly distributed between 0 and 1. For every unit, the same random number is used on each sample occasion when drawing stratified Simple Random Sampling (SRS). The SRS is applied within each stratum using a **starting point** which is any number between (0,1) selected to be different across surveys, to avoid an overlap of sampled units.

An overlap of sampled units between surveys can only happen in fully enumerated strata and not in the sampled strata as a result of sampling using different starting points. This is used to spread the response burden amongst sampled units.

In household-based surveys, a two-stage sampling scheme is used where PSU are selected in the first stage and Dwelling Units (DUs) are in second stage. There should not be an overlap of sampled DUs within PSU across different surveys; this is due to sampling of DUs using systematic sampling with different starting points.

The burden placed on each reporting unit (respondent) is not visible to the survey manager because each sample is seen as an independent process.

The following tables depict how each typical independent business survey sample would look like:

Monthly Retail Survey	**Name**	**Industry**	**Stratum**	**Enumeration status**
Sample Unit 1	MGX LTD PTY	62	1	Fully enumerated
Sample Unit 2	MTT LTD PTY	62	1	Fully enumerated
Sample Unit 5	BEUT LTD PTY	62	2	Sampled unit

Monthly Financial Services	**Name**	**Industry**	**Stratum**	**Enumeration status**
Sample Unit 3	BEEE consulting	8	1	Fully enumerated
Sample Unit 6	Plastics services LTD PTY	8	1	Fully enumerated
Sample Unit 7	COCOC LTD PTY	8	2	Sampled unit

Annual Employment Survey	**Name**	**Industry**	**Stratum**	**Enumeration status**
Sample Unit 1	MGX LTD PTY	62	1	Fully enumerated
Sample Unit 2	MTT LTD PTY	62	1	Fully enumerated
Sample Unit 3	BEEE consulting	8	1	Fully enumerated
Sample Unit 6	Plastics services LTD PTY	8	1	Fully enumerated
Sample Unit 8	AGA services	8	2	Sampled unit
Sample Unit 4	PGMT construction PTY	5	2	Sampled unit

From the table above, it is not immediately visible that Sample Units 1, 2, and 3 are in both monthly and annual surveys and this is because they are fully enumerated, while sampled units are only sampled once, this is due to the use of different starting points across surveys.

The following table is a generic design which carries information for multiple surveys:

Sample Unit #	**Monthly Retail Survey**	**Monthly Financial Services**	**Annual Employment Survey**
Sample Unit 1	Y		Y
Sample Unit 2	Y		Y
Sample Unit 3		Y	Y
Sample Unit 4			Y
Sample Unit 5	Y		
Sample Unit 6		Y	Y
Sample Unit 7		Y	
Sample Unit 8			Y

Possibly a more efficient way to store the table above is to insert the stratum number in the Survey column.

Sample Unit #	**Monthly Retail Survey**	**Monthly Financial Services**	**Annual Employment Survey**
Sample Unit 1	1		1
Sample Unit 2	1		1
Sample Unit 3		1	1
Sample Unit 4			2
Sample Unit 5	2		
Sample Unit 6		1	1
Sample Unit 7		2	
Sample Unit 8			2

This indicates which surveys each unit is in while also indicating its relevance (stratum) without requiring an extra field for that.

The table is stored as a sample register which can be supplemented by other registers (population register with contact details and other attributes) which when put together will enable statisticians, methodologists and data collectors to perform their functions.

This immediately creates transparency across multiple surveys while also reducing the storage required to save samples for each survey.

Processing (IT Systems)

Infrastructure and technology is the foundation that enables everything discussed in this paper.

End users at Stats SA require continuous availability of all information systems and also require access while they are away from the office, therefore to satisfy this demand, certain users have been granted remote access to internal resources while out of the office using Microsoft Terminal Services and access to statistical information systems is made in a secure manner without any restrictions. This infrastructure allows people to perform their functions irrespective of time or distance from the office.

As servers age they are migrated to a virtualised server environment. The flexibility and redundancy offered by the virtual platform enhance server administrator's ability to respond to threats such as hardware failure with very little human intervention.

To limit the risk of data loss and increase data security, data storage has been centralised with the use of a SAN (Storage Area Network). In order to utilise the storage efficiently while taking organisational requirements into consideration, infrequently accessed data gets archived to an archiving solution or slower storage.

Backup solution management also becomes more predictable and easier with the use of centralised storage.

Raw survey data is stored in a RDMS (MS SQL Server), Users capture/edit data with in-house/custom developed systems. Systems are developed in VB/C# using the .Net framework, most systems are then hosted on an application server (IIS) having client access from a web browser. Security and access to respective databases/data sets are controlled by either domain or ASP authentication.

The processing part will be discussed according to its various phases, namely:

- Design and Build;
- Collection (Capturing); and
- Processing and Analyse.

Design and Build: Given the robustness of the upstream process mentioned (Registers with metadata and dynamic samples), the design and build process of the IT systems becomes relatively easy. The design and build process has universal compatibility to all surveys that have the correct metadata and objects defined in the object register. Data validation is done with key enforcement in the application.

The database is the combination of the Object Register; specifically the data glossary which specifies which survey component is used by which survey, and the samples drawn. This ultimately ties a set of survey objects (questions) to a set of sampled units for each survey.

Central to this design is governance. This ensures that each individual will act according to the designated role and that objects will interact in the desired manner.

Deciding on the system/database design architecture will depend on many factors including:

- Size of survey (Questionnaire and Sample Size);
- Method of Processing (Manual capture, OCR etc.);
- Expected IO operations;
- System performance expectations; and
- Purpose of system (System requirement).

One of three possible system/database architectures is currently implemented:-

1. Traditional relation database design (normalised);
2. Full generic design (meta data and survey data in row based combined design); or
3. Hybrid design (Meta data in row based design and survey data in relation design. Binding is done on meta data level).

Survey 1 has sample unit 1 in its sample. If question 5, is not asked in survey 1 then sample unit 1 cannot be allowed to answer question 5 for survey 1. This kind of design reduces human intervention. How humans interact with the survey will be according to the set metadata and registers.

Collection: The processes used by Stats SA to capture information received from the field are a hybrid of manual end user capturing systems and a form reading (character recognition) systems. Both of which use the dynamic database designs.

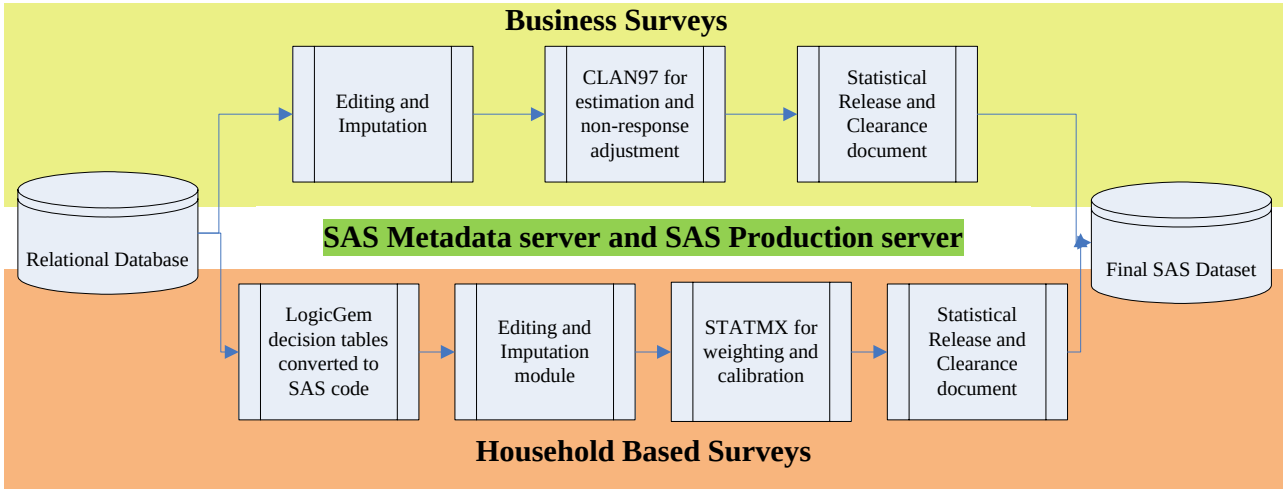
Collecting data digitally are currently being investigated and tested by various divisions within Stats SA, the following are currently being considered: Electronic forms, handheld devices (PDA's, Tablets, Smart phones) and respondent online capturing vis. the internet.

When a pre-determined phase is reached during the collection period, the capturing process is stopped. To enforce this, the metadata of survey is modified by the process owners to a state that does not allow any further inserts or modifications into the database. This "frozen in time" state ensures consistency in the results during the processing and analyses phase and allows retrospective analyses in the future.

Processing and Analysis: Previously in order to process and analyse the collected/captured data, each statistician had to assemble the different data streams for each part of their analysis. This posed challenges for business continuity and is flawed with human intervention.

The processing and analysis is carried out in the SAS® platform which is a client-server environment. The objects available in this platform are sourced from the MS SQL Server® using Open Database Connectivity (ODBC) connections. The SQL Server data sets are considered the Master data sets are is preserved for any future needs. The ODBC connections have "read-only" rights.

Figure 1 – Statistics South Africa's Architecture



Different tools (all running on the SAS platform) are used for Processing and Analysis for Household-based surveys and Business Surveys.

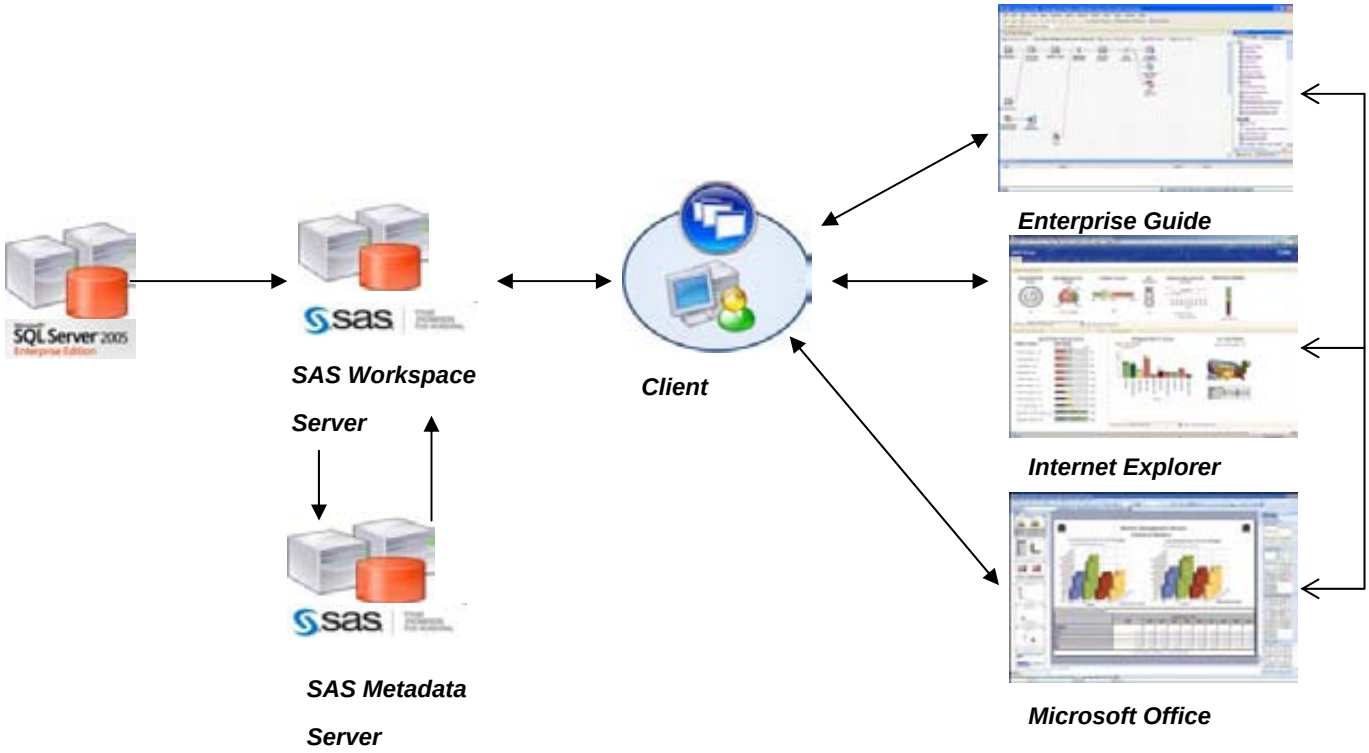
In Household-based surveys, edit and imputation rules are captured with a decision-table tool LogicGem. The rules are then converted to SAS code. Through the use of SAS Enterprise guide generic methodologies can be applied. Statistical Macro Extensions (StatMX) is used for weighting and calibration. In business surveys, estimation and non-response adjustments are carried out with help of CLAN97. CLAN97 is a SAS-program designed to compute point- and standard error estimates in sample surveys. Imputation and editing are automated in SAS; however, some manual interventions are still required for some of the cases.

The SAS® platform provides the statisticians the ability to access data from all data sources from one application while being able to perform data analysis and publish from the same application.

The platform also allows collaboration on projects/process. All the processes created and used by the statisticians are stored on the servers. This reduces issues associated with version control. A process can be accessed by a defined group of users, each with their rights to "read-only", "modify" or even "delete". This ensures consistency across all surveys and business continuity.

Two tiers/servers are available for statisticians at different phases of their analysis. The Development tier is used to develop and refine the process. As the process matures, it is migrated to the Production tier. The production tier allows very little, if any, modification to process and not all the statisticians have access to this tier.

Figure 2 – Statistics South Africa processing and analysis phase



The SAS® platform enables the statisticians to interrogate the data, analyse it and publish their findings. The platform also gives the statisticians tools Extract, Transform and Load (ETL) tools to scrub their data sets.

The platform uses SAS Enterprise Guide as its primary client tool. This application is the default tool for data access and analyses. The results or discoveries made in Enterprise Guide can be shared or made available to a wide audience in the organisation through common enterprise tools such as MS Internet Explorer® and MS Office®.

SAS offers a MS Office® add-in that allows data analysed in Enterprise Guide to be made available to MS Office®. The add-in enables statistical graphs, basic procedures and exploring of OLAP Cubes to be done in MS Office®. Users create graphs and table layouts once. For future publications, they simply refresh the tables and the latest data is retrieved from the data sources in a 'dissemination ready' format.

MS Internet Explorer® is used for presentation purposes. Executives use this tier to view dashboards and do basic analysis. The tier has been connected to ESRI ArcGIS® which enables Geo-Analysis. This gives the users the ability to have not only view their data but to see the geographical distribution and how to set or measure each geographical region.

The platform offers reporting flexibility in order to suit different and user requirements.

Dissemination

The time spent on the dissemination process is reduced by the SAS' MS Office® add-ins ability to produce 'dissemination ready' documents.

As part of the internal dissemination process, an internally shared data store was created for all users to save their final publication data sets in SAS or XML format. The data sets are used as downstream inputs by other statisticians.

External users gain access to disseminated data through the Stats SA website (http://www.statssa.gov.za) at the embargoed time. The website provides a wide variety of publication formats. The default format is in Adobe Acrobat® (PDF). Other formats such as SAS, PX Web, ASCII, Excel and Really Simple Syndication (RSS) feeds are available. XML format is being investigated as an output format for external users.

Methodological documents are also available on the website.

Concluding Remarks

In this paper we have presented the architectural concepts, tools and the methodology used for the implementation of an integrated information system for Statistics South Africa (Stats SA). The focus has been on the integration between systems and standards developments and methodological support during the design and implementation of statistical production processes.

We have shown how an array of integrated IT systems tools can be used to support and optimise the underlying statistical production processes

We have argued that the potential of integration can only be exploited through a holistic approach, focusing on the integration of statistical data and the synergy of statistical processes using information and technology.

REFERENCES

1. United Nations, 1999, Information Systems Architecture for national and international statistical offices, Guidelines and Recommendations, Conference of European Statisticians, Statistical Standards and Studies, No.51, Geneva
2. Gregory E. FARMAKIS, Yorgos KAPETANAKIS, George A. PETRAKOS, Michalis A. PETRAKOS, Architecture and Design of a Flexible Integrated Information System for Official Statistics Surveys, based on Structural Survey Metadata.