UNITED NATIONS STATISTICAL COMMISSION
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Workshop on the Common Metadata Framework
(Vienna, Austria, 4-6 July 2007)

CASE STUDY - STATISTICS SOUTH AFRICA[1]

1. Introduction

1.1 Organization Details

Name: Statistics South Africa

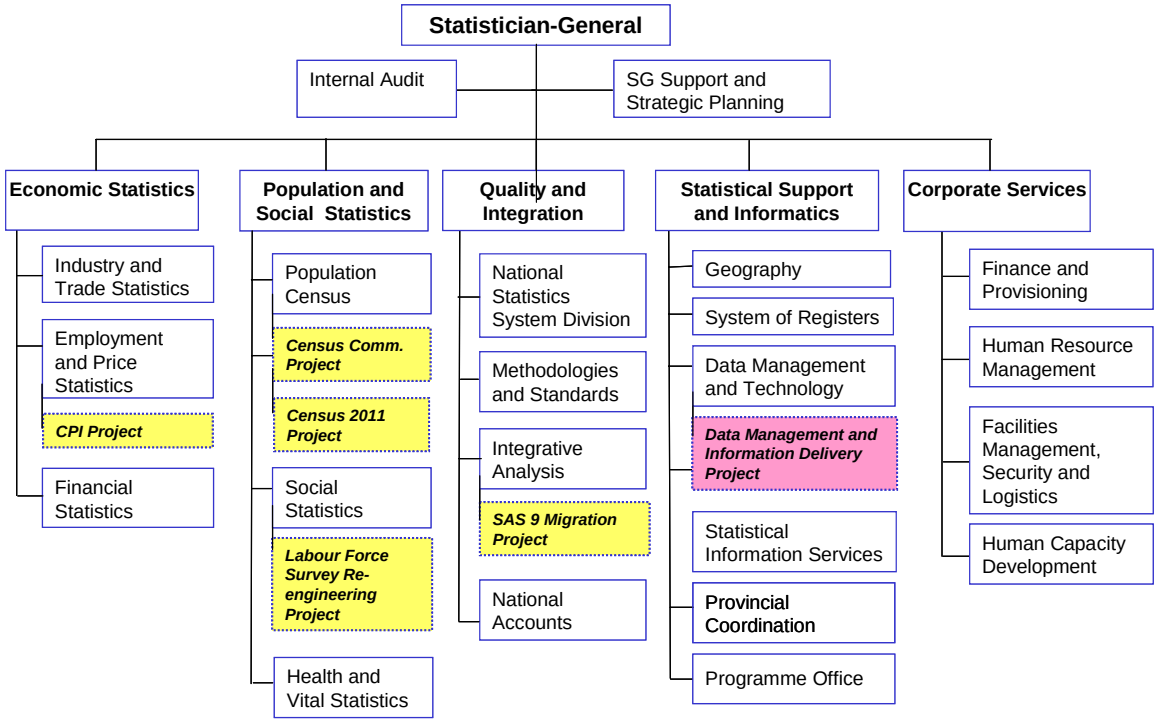
Contact Details:

Physical Address:
De Bruyn Park
170 Andries Street
Pretoria
0001
South Africa
Mailing Address:
Private Bag x44
Pretoria
0001
South Africa

Telephone Number: +27-12-310-8911
Facsimile Number: +27-12-321-7381

[1] Prepared by Sibongile Madonsela (SibongileMA@statssa.gov.za) , Matile Malimabe (MatileM@statssa.gov.za) Bubele Vakalisa (BubeleV@statssa.gov.za) and Ashwell Jenneker (AshwellJ@statssa.gov.za) all from Statistics South Africa (www.statssa.gov.za)

1.2 High-Level Organisational Structure



Number of Staff: ± 2,000

Figure 1: Stats SA Organization Chart

The Data Management and Information Delivery (DMID) project (magenta shaded box) is located within the Data Management and Technology Division (DMT)
The yellow shaded boxes indicate the ongoing projects that are concurrent with the DMID project.

The following chart shows how the DMID project is structured:

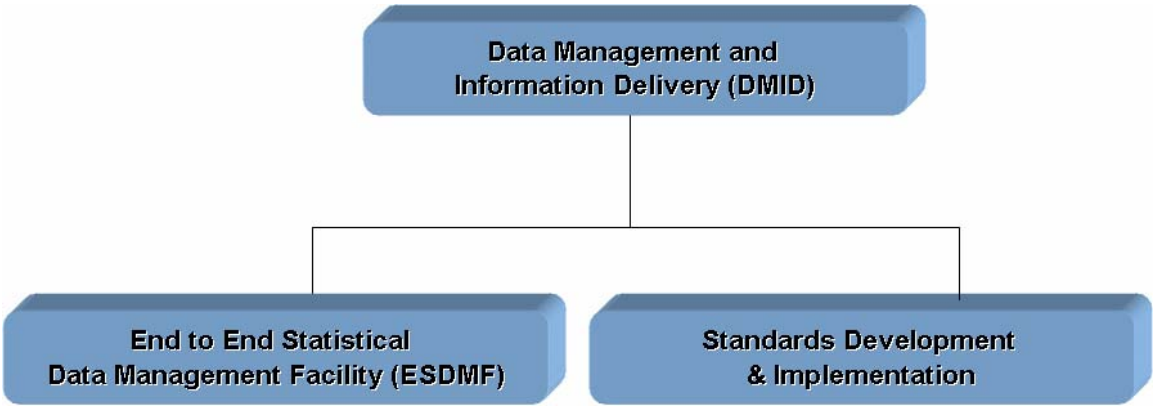


Figure 2: The DMID Project Structure

The following chart show how the DMID project is structured, including the supplier's resources:

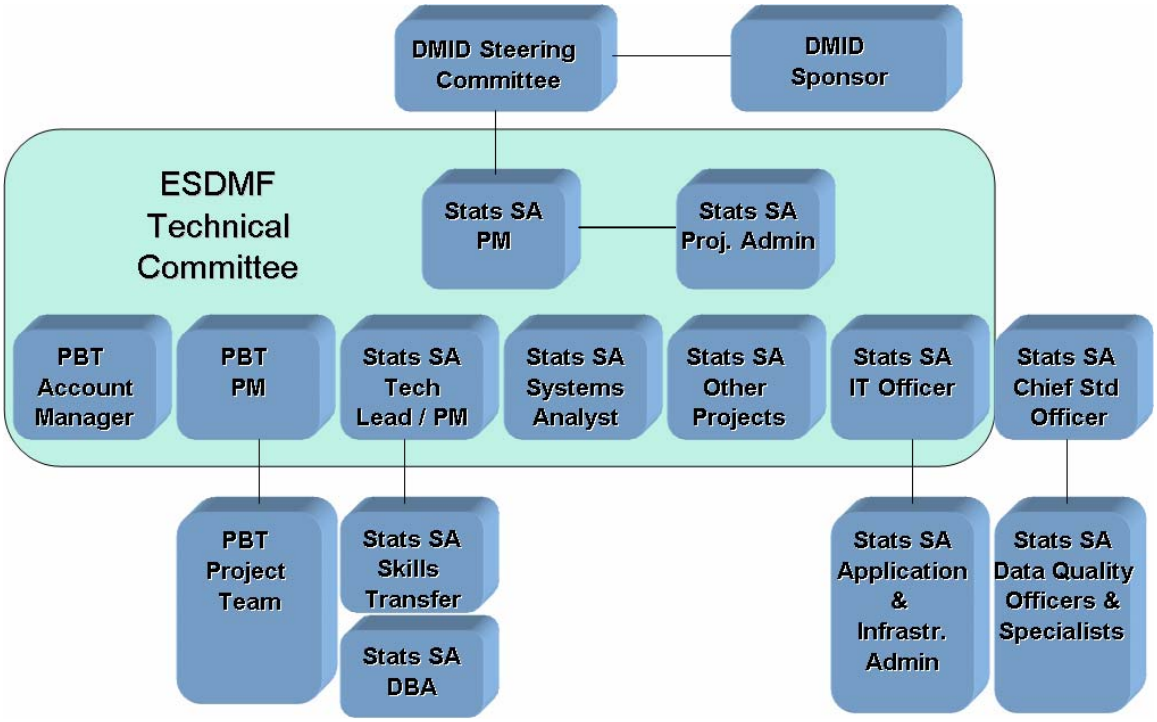


Figure 3: DMID Organization Chart

Prescient Business Technologies (PBT) - is the name of the supplier to the DMID project, developing the ESDMF System.
ESDMF – End to end Statistical Data Management Facility.
PM – Project Manager.

Number of staff:

- **Stats SA:** 1 Project Manager, 1 Technical Lead/Project Manager, 7 Developers, 1 Chief Standards Officer, 6 Data Quality officers/specialists, 3 Methodologists, 1 Systems Analyst, DBA (as needed), Network Support Technician (as needed) (20 total – excluding "as needed")
- **PBT:** 1 Project Manager, 1 Technical Lead (50%), 1 Architect, 1 Business Analyst (50%), 1 Release Coordinator, 1 Trainer, 1 Organisational Change Management Lead, 3 Developers, 1 Account Manager (10 total)

1.3 Strategy

Statistics South Africa's development of the metadata management system has its origins in the organisation's requirement to develop a data warehouse. The idea of a data warehouse came about because the organisation wanted to improve the quality of the statistics produced. It was believed that the data warehouse would play a major role in positioning the organisation within its vision of becoming the "preferred supplier of quality statistics". To begin our data warehouse initiative, we paid exploratory visits to various statistical organisations that had embarked on data warehouse developments in order to learn from their experiences. These visits taught us that a number of things about the complexities, difficulties and peculiarities of developing a data warehouse. In particular, our visit to the Australian Bureau of Statistics showed us that for a data warehouse to have any chance of succeeding in a statistical organisation, it needs to have a strong foundation of standards and policies that govern the statistical production processes. Standardisation of concepts and their definitions, as well as classifications of the terms of the actual survey process, were all found to be necessary for the

production of quality statistics. For it to be successful, a data warehouse also needs to operate in this environment.

A formal process for standardization was developed through consultation with standards experts standards development and implementation lifecycle was developed to monitor the standardization process. The following is the standards development lifecycle.

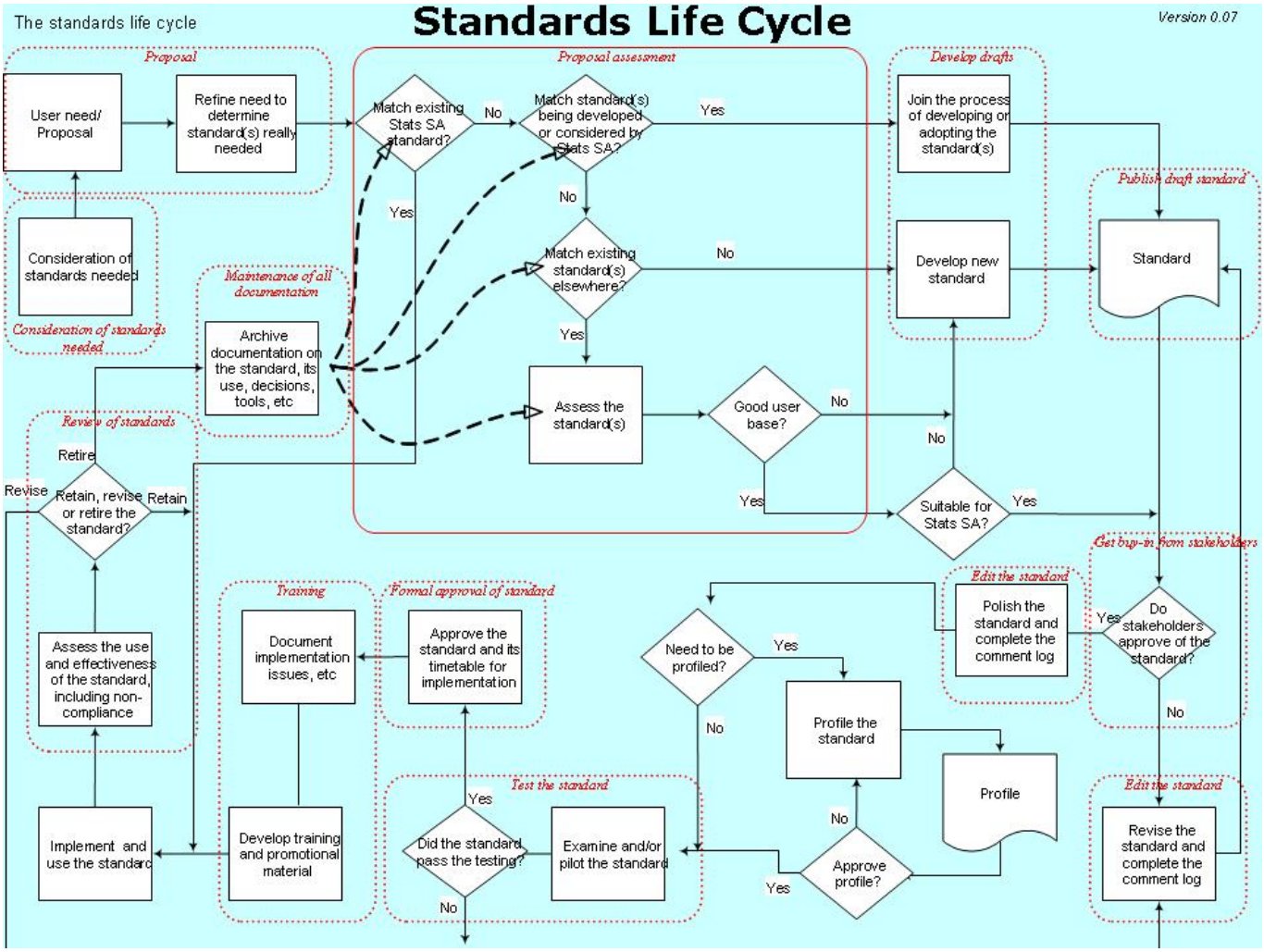


Figure 4: Standards Lifecycle

The next step for us was to investigate the strength of our standards and policy foundation. Upon this investigation, a number of gaps were identified. Chief among these was the lack of standard metadata in the organisation. The need for standardisation of metadata necessitated the development of a metadata management system. However, this had to form a good mix with all the other identified ingredients necessary for the production of quality statistics.

Strategically, our metadata management system forms part of a larger system of applications called the End-to-end Statistical Data Management Facility (ESDMF). As an end-to-end system, the ESDMF will consist of tools and applications to support the whole statistical production process. Within this facility exists a metadata subsystem (refer to figure 5), which plays a central role as the ESDMF was conceived to be metadata driven. In a statistical organisation, a metadata driven system is inevitable because metadata is used and generated at every stage of the statistical production process.

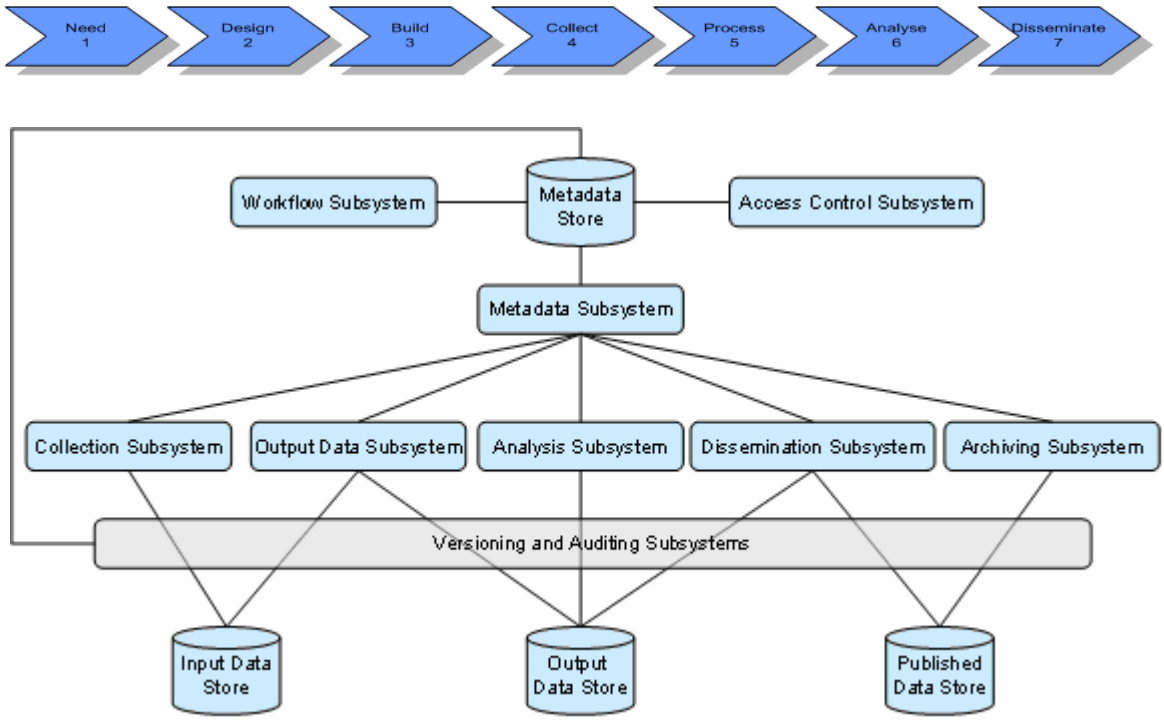


Figure 5: Conceptual components of the ESDMF

As a data factory, a statistical organisation needs to organise and package data in ways that make it useful to the end user. Produced data must also meet certain minimum quality standards. To satisfy both these requirements, use of metadata is invoked. In packaging its data and statistical products, a statistical organisation must ensure that they are attached with metadata for ease of analysis and interpretation by their users. Metadata also play a key role in ensuring that the end products of this data factory are of good quality. Such metadata includes descriptions of concepts used in the organisation, classifications of these concepts, methodologies and business rules. These are all necessary metadata to ensure that products are of good quality.

The development of a metadata management system was informed by the following principles:

- *Maintenance of trust in official statistics:* Descriptions of data collection methods, data processing, and storage needed form part how statistical data are presented to the end user. When presented like this, statistical data and products engender trust to the users.

- *Facilitation of correct interpretation of statistical data:* Metadata accompanying datasets and other statistical products.

- *Quality of statistics:* Standard metadata contributes to the improvement of a number of quality dimensions. Standardisation of concepts and their definitions and classifications are essential ingredients of standardized metadata.

1.4 Programme Providing Frame for Stats SA Projects

The work of all Stats SA components is mapped out in the organisation's Work Programme. Organisational units must support the following strategic themes to advance the work of the organisation:

- *Providing Relevant Statistical Information to meet user Needs*

- *Enhancing the Quality of Products and Services*

- *Developing and Promoting Statistical Coordination and Partnerships*

- *Building Human Capacity*

This project is aimed at supporting the strategic theme "Enhancing the Quality of Products and Services". Within the DMID project, the metadata management system, more than any of its components, addresses this strategic theme.

1.5 Overall Project Objective

Statistics South Africa's metadata management system therefore forms part of the organisation's broader objective to continuously improve the quality of its products. As the driver of the overall facility, the metadata management system is the first deliverables of the DMID project. The metadata management system is also divided into smaller logical units based on the organisation's classification of its metadata. Survey metadata, consisting of elements for providing the overall description of a statistical survey is the first of these metadata deliverables. The survey metadata component is fashioned along the lines of Statistics Canada's Integrated Metadata Database (IMDB) Metastat.

Following the survey metadata component will be the definitional metadata component. This will incorporate into the metadata management system the standardised organisation-wide concepts and their definitions and classifications as well as other components that form part of definitional metadata.

1.6 Lessons Learned

The supplier had a difficult time understanding the business of Stats SA, which is statistical production processes. Additionally, the goal of the project is to improve quality, which will help support the vision of Stats SA "to be the preferred supplier of quality statistics". Even in the face of this vision, the supplier failed to recognize that quality was a primary business objective.

Under pressure of meeting the deliverables, the supplier ignored the Skills Transfer Plan, with the result that the Stats SA developers were not involved in the final design and development of the system.

For a project of this magnitude (three years), we decided to break down the deliverables into twelve phases. Each phase was planned to be three months long in duration. Also, each phase was planned to be a complete deliverable in its own right, even though the next phase was planned to build on the previous phases. The first phase was delivered late mainly due to the lack of understanding that the supplier demonstrated. The key is that clear understanding of the requirements is very important in meeting the deliverables as well as milestones for those deliverables.

2. The Statistical Metadata Systems and the Statistical Cycle

The essence of Stats SA's meta-information system is captured by how the organisation uses the metadata. Metadata is used internal to the organisation to enable statistical production processes. This means that metadata is used during various stages of statistical production as essential input to production processes. However, the production processes in turn, produce metadata. This metadata is also important in documenting the trail of activities during the statistical production process. The documentation of production activities informs related metadata issues such as the assessment of data quality and its interpretation.

2.1 Categories of Metadata

Because of this diversity of metadata usage, it was decided that contents of the meta-information system should be aligned with these usage activities. The natural progression of this decision was to undertake a project to classify all of the organisation's metadata. The following is a list of the categories of metadata adopted by Stats SA:

- **Survey Metadata**

Often referred to as dataset metadata, Survey metadata is used to describe, access and update dataset, data structures. Stats SA chose to call this type of metadata *survey* rather than *dataset* because some of the metadata such as information about "the population which the data describe" refer to the broader aspects of the survey, and not only the dataset.

- **Definitional Metadata**

  This is metadata describing the concepts used in producing statistical data. These concepts are often encapsulated into measurement variables used to collect statistical data. Descriptive text is used to define individual concepts, however the concepts are further grouped into logical topics. These main topics are effectively classifications of data. Hence, included in Stats SA's package of definitional metadata classifications drawn from different study domains.

- **Methodological Metadata**

  These metadata relate to the procedures by which data are collected and processed. These may include Sampling, Collection methods, Editing processes, etc

- **System Metadata**

  System metadata refers to active metadata used to drive automated operations. Some of the examples of system metadata are:

  - Publication or dataset identifiers date of last update
  - File size
  - Mapping between logical names and physical names of files
  - Dataset input flows
  - Access methods to databases
  - Coordinates as kept in metadata store
  - Table and column definitions schema and mappings of data

- **Operational Metadata**

  This is metadata arising from and summarising the results of implementing the procedures. Examples include Respondent burden, Response rates, Edit failure rates, Costs and other quality and performance indicators, etc

The different components of Stats SA's meta-information system are logically grouped according to these categories of metadata. This means that the database for the meta-information system has different data structures corresponding to these metadata categories. We have recently (June 2007) finished developing the first metadata component, the survey metadata capturing tool, which is the subject of this case study.

2.2 How Metadata Fit into Other Organisational Systems

As already stated, the development of Statistics South Africa's metadata management system (Meta-Information system) is part of a larger system, the ESDMF. The central components of the ESDMF will follow the completion of the meta-information system, because the ESDMF is driven by the metadata. Although the ESDMF is a new system, it is merely a means to centralize the organisation's disparate statistical information systems. Figure 6 below shows the conceptual ESDMF subsystems and how they are placed relative to other organizational subsystems. The metadata subsystem supports the entire statistical cycle.

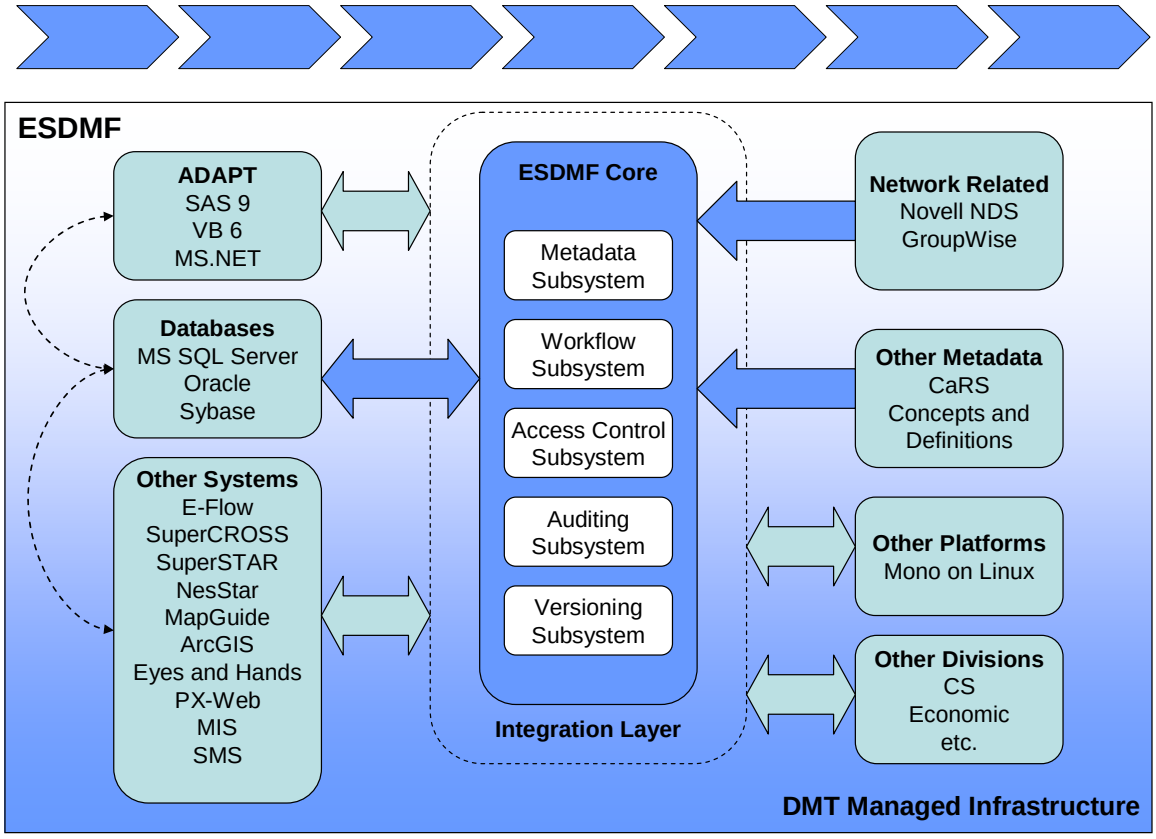


Figure 6: Conceptual components for the ESDMF in relation to other subsystems

2.3 Overview of the Process Model

Also shown in figure 6 are the main phases of Statistics South Africa's statistical cycle, internally referred to as the statistical value chain (SVC). The development of the SVC forms part of the organisation's standardisation of processes and systems. The SVC will provide broad guidelines for survey areas.

It was decided that the organisation should not re-invent the wheel in its development of the SVC. Therefore in developing the survey cycle, the starting point was to study how national statistics organizations (NSOs) similar to ours had logically structured their survey cycles. Another guiding principle was that Stats SA was not re-engineering the way it conducts surveys, but wanted to map and structure it as much as possible. We had to adapt survey cycles from organisations whose survey operations mimicked those of us. This work was done in 2005. Having made all considerations, we found the Statistics New Zealand's business process model for survey cycles to be the most aligned to our survey operations.

2.4 How Stats SA's Statistical Process Maps into Common Metadata Framework (CMF) Lifecycle Model

Resulting form the process stated above, Stats SA's survey cycle consists of the following phases:

1) Need

Although Stats SA already produces statistics that satisfy needs of a large and diverse group of the country's socio-political landscape, requirements for new or supplementary statistical products often arise. Requests for such new products may emanate from a number of sources, including government departments, private business and other stakeholders. When this happens, the first step in statistical production is to understand the need for the required statistics, i.e., what the required statistics are going to be used for in concrete terms by their users. Often for new projects, a lot of detailed information about what is needed is not at first clear and in some cases, not present, even from the perspective of the initiators

of the project. It is therefore important to go through a process of refining the understanding the information (statistics) needs to be addressed by the results of the project.

2) Design

The design phase consists of preparing ground for the execution of a statistical production project. This stage is reached when either a new project has been given the go ahead or a frequent on-going project is about to begin. In the case of on-going surveys, the design is usually in place already, but in a few cases it might need to be altered to cater for additional requirements or special circumstances.

3) Build

The build phase puts together all the pieces of the infrastructure for a statistical production project. These include the computer system, scanners, printing out of questionnaires, etc. The build phase also includes the procurement of the pieces of the infrastructure as necessary. The amount of work done in this phase also varies for new and on-going surveys.

4) Collect

Although the term collection may have different meanings in a statistical organization, it is used in the SVC to refer to both *direct* and *administrative* methods of data collection. The direct collection method refers to data collection in which Stats SA sources data directly from the respondents. In administrative collection, data are drawn from databases of other organizations which in turn source them from their respondents. It is important to note that this phase has a major influence on the activities of the design phase.

5) Process

The Process phase includes capturing collected data into databases so that data processing may be done. Data processing is necessitated by a number of issues. Chief among these is a fact that the data collection process is fraught with errors. The process phase is undergone to remove these data validity errors so as to improve data quality, and to package the data for use by analysis tools.

6) Analyse

After data have been cleaned during the Process phase, they are now ready for manipulation using analytical tools. This is the analysis done by domain experts to get insight into the meaning of the data. Further data quality enhancements may be done at this phase.

7) Disseminate

Stats SA collects data in order to produce statistics to be used by different stakeholders in the country including the general South African public. This means that the organization has to have ways of giving these communities access to the data and the resultant statistics. The Disseminate phase formalizes the steps Stats SA needs to go through in order to distribute information to the different communities as well as give them access to data repositories.

Stats SA uses a number of dissemination methods to ensure that the data produced by the organization is accessible to the widest user community. These include: electronic (e.g. via the internet), printed output and compact disks.

Table 1 below shows the mapping between Stats SA's survey cycle and the METIS cycle:

METIS	Stats SA
Survey planning and design	Need and Design Phases
Survey preparation	Part of Design Phase
Data collection	Collection Phase
Input processing	Processing Phase
Derivation, Estimation, Aggregation	Processing Phase
Analysis	Analysis Phase
Dissemination	Dissemination Phase
Post Survey Evaluation	

Table 1: METIS Cycle vs. Stats SA's Cycle

Post Survey Evaluation is currently done outside the statistical cycle. It is performed only for the large surveys such as the population census and the community survey.

3. Statistical Metadata in each Phase of the Statistical Cycle

Metadata are used and/or produced in each phase of the statistical value chain. This strong link between the between the SVC and metadata informs all the development of the metadata subsystem.

3.1 Stats SA's Statistical Value Chain

Statistics South Africa's core areas, i.e., those divisions in the organization responsible for the production of statistics, have up to now operated using different approaches. Although it is generally understood in the organization that there are many commonalities in the way different divisions conduct their work, no attempt has been made to formalize a standard statistical production process for the entire organization. The development of the SVC for the organization is a move to correct this situation. The SVC is a generalisation of the activities that need to take place from the beginning to the end of a statistical production process.

Stats SA envisions its statistical cycle along the lines of Michael Porter's Value Chain Model[2]. Hence we refer to our statistical cycle as the SVC. The value chain categorizes value adding activities of an organization. Figure 7 below is a schematic diagram of the main phases of Stats SA's SVC.

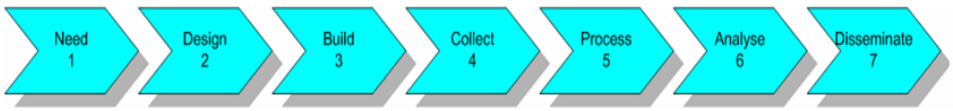


Figure 7: High level phases of Stats SA's Statistical Value Chain

The SVC was designed to be general, catering for most scenarios of statistical production. For example, it is clear that not all the phases of the value chain will be used by all surveys. Figure 8 below shows a flowchart of statistical production within the context of the SVC. It can be seen that old frequent surveys might not follow the same path as new frequent or once off surveys.

[2] Michael Porter explained this model in his 1985 book, *"Competitive Advantage: Creating and Sustaining Superior Performance"*

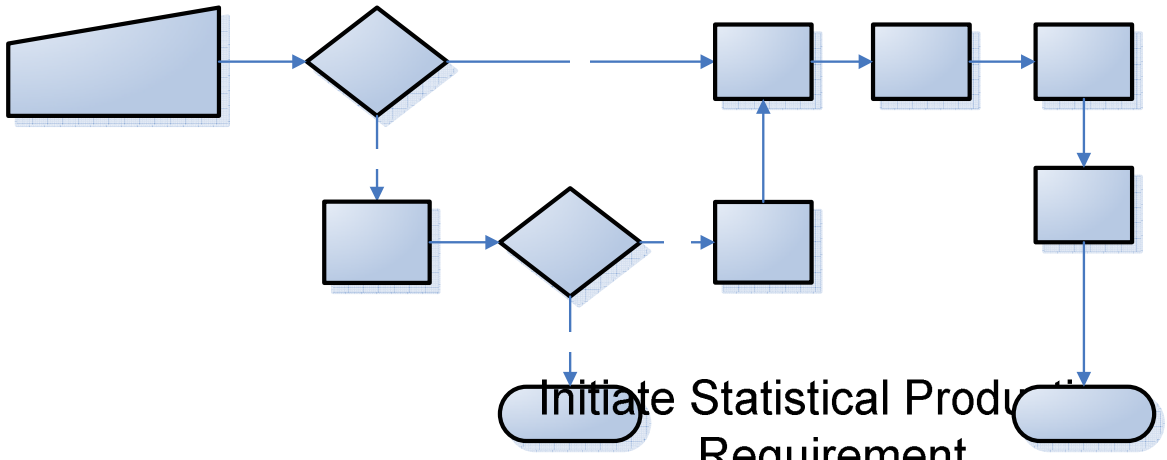


Figure 8: Flowchart of a statistical production using phases of the SVC

A high level description of the main phases of the SVC was given in section 2.4 above. In this section we give a detailed view of the activities involved in each phase.

3.1.1 Need Phase

The Need phase consists of the following activities:

- **Determine the need**

  The objectives and purpose for doing the particular survey or research must be defined. This starts with conducting interviews with the organisation or individual(s) requesting the new survey. This is an iterative process that concludes with a definition of a *statement of need*.

- **Determine Information Requirements**

  A need for a survey or study is triggered by requirements for information that solves a given problem. A clear determination of the nature and extent of this information or data is needed. This is done through consultations with domain experts from the community in need of the information.

- **Develop Budget and Plan**

  Similar to any project that requires resources, a statistical production project has to have a cost-benefit analysis as a foundation of its business case. During this phase, only a high level plan is produced.

- **Obtain Financial Support**

  Generally, Stats SA's projects are big and critical; thus they need huge financial investments. Because the government pays for them, an intensive process of budget approval has to be undertaken in order to ensure accountability.

- **Ministerial Approval**

  Stats SA projects are funded by the National Treasury under the Ministry of Finance. For large projects to go ahead, ministerial approval is required.

3.1.2 Design Phase

The following activities are contained in the Design phase:

- **Develop Detailed Project Plan**

  The output of the Need phase consists of high level aspects of the proposed survey. All Stats SA's surveys must go through detailed planning. For new priority projects, the responsibility for such planning lies with the organisation's Programme Office. The Programme Office has the overall responsibility for running the project to completion, after which, the future running of the project (in the case of frequent surveys) is handed over to the survey area.

- **Develop Survey Methodology**

  The goal of the survey methodology is to ensure that the statistics collected during the survey are reliable and representative of the survey's target population. For existing surveys, the survey methodology is often already in place. For new and re-engineered surveys, new survey methodologies are developed.

- **Design and Test Questionnaires**

  Questionnaire design is aimed at ensuring that the required information from a survey is realized. It consists of getting both the content and the layout of the questionnaire correct. This process is iterative between constructing survey questions and testing whether the responses to the questions asked address the problem the survey is intended to solve.

  Questionnaire testing is initially done "behind-the-glass", during which employees of the organisation are randomly selected for participation. Thereafter, pilot tests are conducted on the field to small population groups in the same way the actual survey will be conducted.

- **Design Operational Requirements**

  Survey operations are concerned with the tasks of getting data from respondents or other data sources. Operational requirements must detail all the technical and logistical issues that need to be sorted out in order to have a successful survey. These vary from resource issues to technologies needed to conduct the survey.

- **Design Computer System**

  The system to be used during the statistical production process consists of many related sub-systems that may be implemented through computer technology. Data collected during a statistical survey is captured in computer system for processing. A number of technologies are required to ensure that data are moved from their sources of collection to the computer.

3.1.3 Build Phase

Activities contained in the Build phase are as follows:

- **Build a Collection Vehicle**

  Stats SA collects statistical data through one of the following survey methods:

  - Sample survey using questionnaires
  - Administrative surveys, using IT communications methods to access data stored in other organisations' databases.

  Building a collection vehicle consists of ensuring, through building customised or procuring all the necessary infrastructure and items for the conduction of a survey.

- **Build a Technology Solution**

  A technology solution should include all the technological components required to support the entire SVC. These may include hardware such as scanners and Optical Character Recognition (OCR) tools for capturing questionnaire-based data, database management systems, data analysis tools and information dissemination tools.

- **Test Technology Solution**

  Before a technology solution is put into production, it must be tested by the prospective users. This is to ensure that the functionality required by the users is included in the system. Also, issues concerning ease of use, integration of systems are also addressed. At a technical level, the testing of the system may lead to the identification of system bugs that may have been missed during the technical tests done by the developers.

- **Implement Solution**

  The implementation of a solution means that it is deemed ready to be used to perform productive work. Therefore, users get to be trained on how to use the system and thereafter certain people are granted access rights to the system.

3.1.4 Collect Phase

Contained in the Collect phase are the following activities:

- **Manage Respondents**

  Enumerators must be highly trained so that they are able to explain to the respondents the reasons for collecting data and how they were chosen to be part of the survey and the way such information is planned to be used to improve functions of the agency and improve standards of living; whether responses to the collection of information are voluntary or mandatory (citing authority: Statistics Act); the nature and extent of confidentiality to be provided (citing authority: Statistics Act); an estimate of the average respondent burden together with a request that the public direct to the agency any comments concerning the accuracy of this burden estimate and any suggestions for reducing this burden. Respondent management must be done in ways that reduce the burden of survey on the respondent. Burden reduction includes ensuring that re-visits to respondents are kept at minimum and the questionnaire need to be of reasonable length.

- **Post Out**

  Post Out refers to the process of notifying respondents by sending letters via the post detailing this information. Administrative data does not have this requirement, though legal arrangements are put in place in advance e.g. Memorandum of Agreement, Service Level Agreements etc., for the other party to be able to provide the data. When a survey is conducted by enumerators visiting respondents, the respondents must be notified by Stats SA about the pending survey. This notification must include information such as the objective of the survey, the date(s) when the enumerators will be visiting, etc.

- **Acquire Data**

  Data acquisition at Stats SA includes both the direct (e.g. Sample Surveys and Census) and administrative methods. In most direct acquisitions, data are captured on paper based questionnaires. In a few other cases, electronic media may be used. Figure 9 below shows a flowchart of how Stats SA acquires its data.

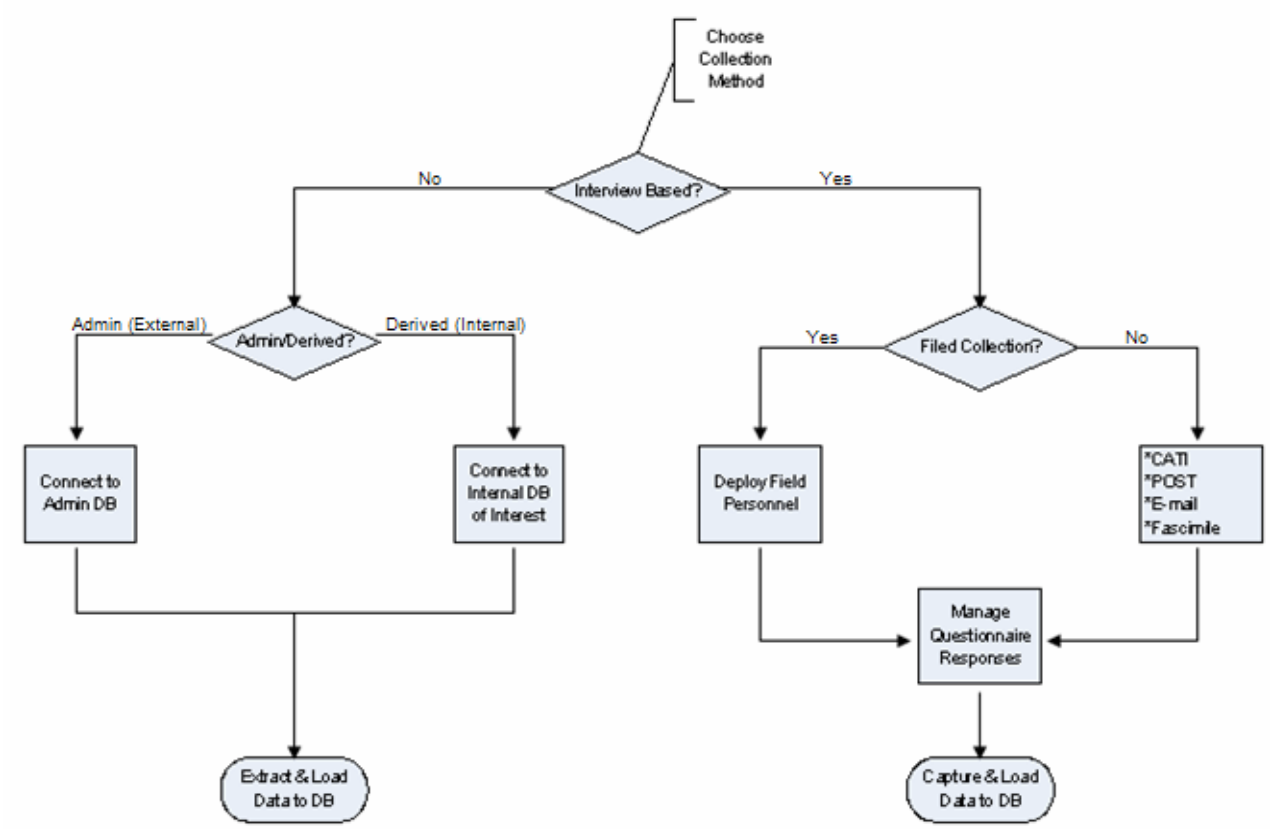


Figure 9: Flowchart of how statistical data are collected at Stats SA

- **Close off Collection**

  The collection period is usually specified at the design stage of the survey. The end of the last day of the defined collection automatically ushers in the closure of field collection of data.

3.1.5 Process Phase

The Process phase consists of performing the following activities:

- **Capturing Data into Electronic Form**

  This applies only to questionnaire based collection methods. Questionnaires are either scanned or manually entered by data capturers into computer databases. Data collected from other electronic systems might only need to be transformed into Stats SA's data formats.

- **Perform Macro Edits**

  Macro edits detect individual errors by: (1) checks on aggregated data, (2) checks applied to the whole body of records. The checks are typically based on the models, either graphical or numerical formulae that determine the impact of specific fields in individual records on the aggregate estimates.

- **Rum Imputation/Estimation**

  Item non-response may result in missing values in a survey dataset. Statistical organizations use imputation methods to calculate estimate values to fill in the missing values. Imputation is implemented using mathematical algorithms through computer programs.

Estimation of missing values should not be confused with the overall statistical estimates which form the main goal of a survey. Statistical estimates are calculated by aggregating all of the collected data. These are often called macro data, and are contrasted with micro data, which are detailed data collected from the respondents.

- **Produce Datasets**

  The primary output of the processing are "clean" datasets that are ready to be analysed. Analysis tools can only process data whose formats and structure they understand. Part of producing datasets is to package them into structures and formats that conform to Stats SA's analysis packages.

3.1.6 Analyse Phase

Statistical data analysis consists of the following activities:

- **Produce Statistical results**

  This is the process where results are produced based on the processing that was done on the data. The ultimate goal of any survey is to produce statistical estimates of the characteristics of the statistical unit of interest.

- **Validate Statistical Results**

  This is where estimates are assessed against expectations, comparing data with the one from previous period, and assessing quality measures to ensure good quality data.

- **Interpret Statistical Results**

  Numbers are meaningless if they are presented without any explanation accompanying them. This is one quality dimension that we cater at Stats SA, that all data that get released should be accompanied by the corresponding metadata.

- **Prepare Content for Dissemination**

  This is the process where actual particular measures are taken to ensure that content from the survey does not disclose information concerning any identifiable respondent. This includes: a) for micro data: remove respondent, content reduction, content modification, b) for tabular data: sensitive cells correction methods such as cell collapsing or suppressing by data providers.

- **Perform Quality Control**

  This process entails making sure that all quality measures in SASQAF have been implemented correctly and the results thereof are known.

3.1.7 Disseminate Phase

- **Receive and Validate Content**

  During this process, the dissemination team goes through a checklist of what was supposed to be accomplished and whether it was done accordingly and correctly. The content received by the team consists of macro and micro data, and other products such as published reports.

- **Manage Dissemination Repositories**

Data to be disseminated are kept in databases (dissemination repositories), from which they are extracted when disseminated. These repositories store datasets (including both micro and macro data), reports and other documents.

- **Pre-release for Publishing**

  This process entails preparations before releasing regarding tables, corporate formatting standards, electronic distribution and hard copy outputs

- **Manage First Release**

  This is where distribution media are managed and controlled in order to ensure that different categories of users of statistical information get access to relevant information. Release timelines are handled within this process.

- **Handle Customers**

  Handling customers is part of customer relationship and stakeholder management. A system to handle customer enquiries exists. Stats SA's Support and Informatics Services unit handles customer enquiries, categorises main users and other users, consult users to determine needs and make sure data is distributed timely to users.

3.2 Metadata Description Matrix

The implemented Survey Metadata Capture Tool of the ESDMF captures the following metadata:

Descriptions are provided for section headings.

1. **Active Metadata Set**
   The file identifier and status of the current/active metadata set is displayed immediately under this section. In other words, the metadata set that the user is currently capturing, editing or viewing.
2. **Overview**
   The elements accessible from this section collectively provide a brief description of the survey.
   The *Overview* section comprises the following items:
   Survey/Series Status
   Objective
   Abstract
   History
   Target Population
   Main Topic
   Main Users
3. **Generic Information**
   The elements accessible from this section collectively provide generic information about the survey time frames.
   The *Generic Information* section comprises the following items:
   Survey Frequency
   Series Time Frames
4. **Primary Data Source**
   The elements accessible from this section describe external inputs to the survey.
   The *Primary Data Source* section comprises the following item:
   External and Internal Data Sources
5. **Methodology**
   The elements accessible from this section collectively describe the activities conducted and the methods and processes used which are specific to the survey.
   The *Methodology* section comprises the following items:
   Survey Population
   Instrument Design
   Sample Design
   Collection
   Error Detection/Editing
   Imputation
   Estimation
   Quality Evaluation
   Disclosure/Confidentiality Control
   Seasonal and Working Day Adjustment
   Revisions
   Data Item/Variables
   Dissemination
6. **Data Quality Report**
   The element accessible from this section provides a hyperlink to the data quality report for the data release.
   The *Data Quality Report* section comprises the following items:
   Relevance
   Accuracy
   Accessibility
   Interpretability
   Coherence

Methodological Soundness
Timeliness
Integrity

7. **Documentation**
   The elements accessible from this section provide hyperlinks to additional documentation related to the survey.
   The *Documentation* section comprises the following item:
   Documentation
8. **Contact**
   The elements accessible from this section provide information concerning the contact person who will manage enquiries related to the data or information produced by the survey.
   The *Contact* section comprises the following item:
   Contact Person
9. **Loaded Metadata Sets**
   This section lists the file identifiers and statuses of metadata sets created by the current user. It enables the current user to switch between his/her metadata sets.

.
Table 2 below shows the metadata captured with the Metadata Capture Tool against the Statistical Value Chain, with example for each stage of the SVC.

Group	Description	Statistical Value Chain	Examples	Quality Dimensions
Survey Overview	*Brief overview about the survey that highlights the background, purpose, history and usage*	Need	Title of survey, Series status, Objective of survey, Keywords, Main users and usage	Accessibility
		Build	Metadata file identifier, Metadata version	
		Design	Target population, Main topics	
Survey Time Frames	*Information about time frames that the life cycle of the survey will be managed*	Need	Frequency of series, start date of survey, end date of survey	Timeliness
		Design	Reference period, collection period, product release date	
Type of Survey	*Classification of a survey according to its statistical activity that involves collection, compilation and publication of statistical data measuring characteristics of a population*	Design	Derived, Direct (e.g. Sample or Census) and Administrative	Methodological soundness
Primary Data Source	*Information that gives a description about or identifies the administrative data source*	Design	Administrative data information (i.e. title of survey from primary data source, primary data source description, contact person from primary data source)	Pre-requisite
Methodology	*Information about processes that are put in place and methods used to collect, process, analyse and publish statistical release*	Design	Survey population, instrument design, Collection, Editing/Error detection, Imputation, Estimation, Disclosure/Confidentiality control, seasonal adjustments, revisions, Data variables, Dissemination	Methodological soundness, Integrity and Accessibility
Data Quality Report	*Information about quality measures used and the errors obtained as a result of executing the statistical processes*			Accuracy
Design			Sampling errors and Non-	

Group	Description	Statistical Value Chain	Examples	Quality Dimensions
			sampling errors	
Documentation	*Attach any documents with extra information related to specific section of the template*			Interpretability
Contact	*Any additional documents that describe the concepts and definitions, methods and data quality applying to the specific survey*			Accessibility

Table 2: Relationships between various categories of metadata inputs and different phases of the SVC

The following table shows the stage of the SVC at which metadata is used:

Group	Statistical Value Chain	Examples	Quality Dimensions
Survey Overview	Build	Metadata File Identifier, Metadata version	Accessibility
	Collection	Objective, Main topics	
	Dissemination	Title of survey, Series number, Series status, Abstract, History of survey, Keywords, Users and usage	
Survey Time Frame	Collection	Collection period, reference period	Timeliness
	Dissemination	Frequency of series, Start date of survey, Product release date, End date of survey	
Type Of Survey	Collection	Derived, Direct (e.g. Sample or Census) and Administrative	Methodological soundness, Integrity, Accessibility
Primary Data Source	Collection	Administrative data information (e.g. title of survey from primary data source, primary data source description, contact person from primary data source)	
Methodology	Collection	Survey population, Instrument design, Sample design, collection, Quality evaluation, Data variables,	Methodological soundness, Integrity and Accessibility
	Process	Quality evaluation, Data Editing, Imputation, Seasonal adjustment, Revisions, Data variables	
	Analysis	Quality evaluation, Estimation, Data variables	
	Dissemination	Quality evaluation, Disclosure/Confidentiality Control, Dissemination methods	
Data Quality Report	Process	Sampling errors and Non-sampling errors	Accuracy
Documentation	Dissemination	Documentation	Interpretability
Contact	Dissemination	Contacts	Accessibility

Table 3: Metadata produced with groups of metadata with examples for each group

4. Systems and Design Issues

The design of the system will conform to Stats SA's Enterprise architecture. One of the main components of this enterprise architecture is the IT architecture. The software architecture of all the applications developed in the ESDMF is dictated by the IT architecture as document in the Stats SA ICT strategy..

4.1 Overview of Stats SA's IT Architecture

The Stats SA's IT environment, within which the ESDMF is developed, requires systems to adhere to the following architectural principles:

- **Integration**
  The system must integrate with other organizational systems. API's will be built for various applications that need to connect to the ESDMF. However, most of the connection is expected to be at a data level. With the exception of SAS, the organisation uses relational databases. Integration at this level is attained using ODBC connection. SAS supports ODBC and in addition to that, has native support for various databases.
- **Interoperability**
  To ensure interoperability, the ESDMF uses Java as a development standard because of its platform independence. The development of the system as a web application also means that only a web browser is needed to access the application.
- **Modularity**
  The development of all the components of the ESDMF is based on the organisational requirement for building modular systems that allow ease of management and flexibility. The metadata management system is modularised according to the different categories of metadata.
- **Scalability**
  Stats SA's computer applications have to be built such that they can scale up to accommodate the inevitability of growth of an organization. Both the database designs and storage hardware for all the components of the ESDMF are developed to cater for such growth.
- **Flexibility**
  Applications must meet the diverse needs of Stats SA. These needs change with time, and new ones are also discovered. Development of flexible applications that may be easily changed or added to is vital. Part of the insistence on the use of object oriented programming was informed by the need for flexibility. This will minimize "spaghetti programming" associated with large software projects.

4.2 IT Infrastructure Specification

The metadata management system is deployed in an IT infrastructure with a set of minimum specifications. These minimum specifications list the hardware items needed to run the system without going into details of the hardware items themselves.

- **Operating System(s)**
  Desktops are in Microsoft Windows. The application is deployed in an Open Source operating system (Novell SuSe Linux).
- **Computer Network**
  The network architecture is based on open protocols and industry standards. It allows remote access to some employees. This supports both local area (LAN) and wide area (WAN) networks.
- **Computer Servers**
  The system is developed as a client-server application. This means that there is a need for powerful computer servers capable of handling intensive processing.
- **Storage**
  Because of the vastness of data to be generated and/or captured in the system, there is need for a well-managed storage system. The Storage Area Network (SAN) is the technology used at Stats SA to provide storage management.

A. Development Environment			
Function	**Make/Model**	**Operating System/ Database Engine**	**Comment**
Application Server	HP BL45p Quad processor 4 GB RAM 2 x 72 GB HDD	SuSe Linux Ver. 10	Make/Model exceeds recommendation
Database Server	HP BL45p Quad processor 16 GB RAM 2 x 72 GB HDD	Oracle 10g *or* Sybase ASE and Sybase IQ Unix/Linux/Windows	Make/Model exceeds recommendation
Build Server	HP DL 320 Dual processor 2 GB RAM 2 x 72 GB HDD	SuSe Linux Ver. 10	Make/Model exceeds recommendation
B. User Acceptance Test (UAT) Environment			
Application Servers	2 x HP BL45p Quad processor 8 GB Ram 2 x 72 GB HDD	SuSe Linux Ver. 10	Make and model exceeds recommendation
Database Servers	2 x HP BL45p Quad processor 32 GB Ram 2 x 72 GB HDD	Oracle 10g *or* Sybase ASE and Sybase IQ Linux	
C. Production Environment			
Application Servers	2 x HP BL45p Quad processor 8 GB Ram 2 x 72 GB HDD	SuSe Linux Ver. 10	Make and model exceeds recommendation
Database Servers	2 x HP BL45p Quad processor 32 GB Ram 2 x 72 GB HDD	Oracle 10g *or* Sybase ASE and Sybase IQ Linux	

Table 4: Hardware and software specifications for the ESDMF infrastructure

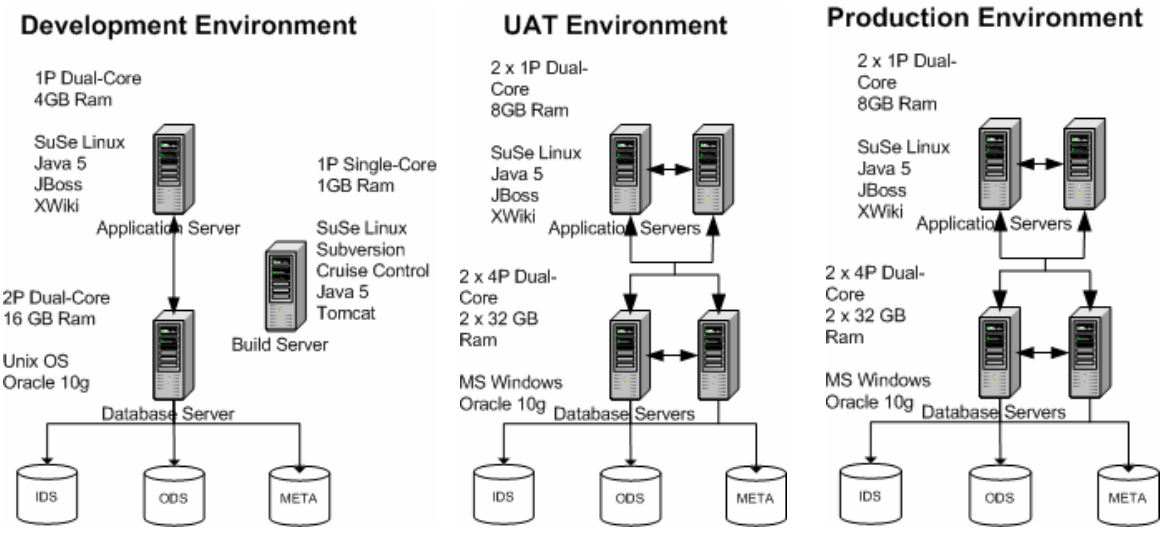


Figure 10: Hardware and software specifications for the ESDMF infrastructure

4.3 Components of Metadata Management Application

The application is web-based and developed in Java. Tomcat is used to implement Java Servlet API and HTTP functionality. The following are physical divisions of the application:

- **User Interface (UI)**
  The user interfaces for all the metadata management system applications is web-based. This allows us to quickly deploy the tool to users in the organization. Client workstations only need to have a web-browser to access server based applications. The main supported web-browsers are Microsoft Internet Explorer and Firefox.

- **Database**
  The application is supported by a relational database management system (RDBMS). Stats SA uses a variety of RDBMS engines. The RDBMS engine of choice for this project is Sybase 12.5.x. The project is currently using the open source RDBMS, MySQL.

- **Business logic**
  The business logic controlling the interaction between the UI and the underlying database is coded using Java server side scripting. There is also business logic coded using stored procedures. This mostly performs housekeeping within the database.

- **Application/Web Server**
  The application is served to the client via Tomcat, which processes Java code. Tomcat also handles HTTP calls from the web browser.

4.4 Tools for Metadata Management

The developed metadata management application allows Stats SA staff members to perform a number of tasks in the metadata management process. The application groups these tasks into three modules or tools.

- **Administration module**
  This module is used to manage users of the system, make changes to certain categories of captured metadata and other housekeeping activities. The administration module will also be used to administer other categories of metadata.

- **Metadata Capturing and Editing module**
  The survey metadata will be continually captured by the originating components whenever an instance of a given survey is required. The metadata captured here is specific to the instance of a survey. This module allows the users to capture and edit survey metadata into the system. A special user role, the Approver, is given permissions to approve all the captured survey metadata, at which point it is exposed for use in the organisation.

- **Query and Reporting Module**
  The metadata repository is query-able and therefore can be reported on. A metadata report is used as one of the ways to document survey data. This may happen in two situations. In the first situation, an internal user may want to view captured metadata. Producing a report of this metadata provides a structured way of viewing this metadata. Another way of viewing metadata is to use the “View Metadata” functionality of the Metadata Capturing and Editing module.

The following screenshots show how the different modules can be accessed from the user interface:

Figure 11: Different modules of the tool

Figure 12: Survey information page with navigation on the right hand side

4.5 How Meta-Information System Integrates to Other Stats SA Applications

Although this feature has not been implemented yet, the metadata management system, like the rest of the ESDMF, is designed link with other statistical processing applications and data repositories. In the immediate future, the repository allows access via the following two methods:

- **ODBC Connection by SAS**

  SAS can extract metadata from the repository for use as input to data processing and analysis activities of statistical production. At this stage, an Open Database Connectivity (ODBC) connection will provide SAS with ability to access the database. When our database is migrated from MySQL to Sybase, there will be an option to use SAS Access to Sybase.

- **APIs**

  Application Programming Interfaces (APIs) will be developed for each application that needs to exchange information with the metadata system. At this initial stage of the project no application uses the metadata management system in this way and therefore no API has yet been developed.

4.6 Metadata Revisions and Version Control

Metadata is expected to change due to revisions of concepts and their definitions, changes to classifications, business rules and user requirements. Sometimes more than one version of certain metadata used for the same purpose may exist at the same time.

In the current Survey Metadata tool the “Edit” functionality of the application allows for the revision of captured Survey metadata. These revisions may only be performed by users with requisite permissions. For changes to be effected, revised/edited metadata must be approved by an assigned *Approver*. Survey metadata can only have a single version. This means that the *Edit* process serves to update the metadata repository.

Version control will be introduced when metadata categories with metadata that can have more than one version are incrementally built into the system.

It is important to note that version control will be built into every aspect of the ESDMF.

4.7 Outsourced vs. In-house Development

The development of all of the ESDMF, including the metadata management system, is outsourced. Two issues influenced the decision to outsource. These were: the fact that Stats SA does not have enough skilled resources and the need to have views which would not be obscured by prior opinions of a statistical environment. This scenario requires that the outsourced resources invest a lot of time in understanding the organisation and analysing the requirements.

It is important to note that we conducted two stages of outsourcing. In the first stage we outsourced the task of gathering the requirements for the whole of the ESDMF. These requirements contain details of each of the components of the ESDMF, including the metadata management system. The second stage is the development of the system. The two tasks were done by two different organisations. This separation of tasks was done in order to maintain the focus on requirements gathering. In this development model, the development team mainly verifies existing requirements.

5. Organizational and cultural issues

Organizational and cultural issues

5.1 Roles in metadata/statistical lifecycle management

In order to understand the user requirements, we engaged the survey divisions as pilot groups. We involved them in verifying our understanding of the requirements, which was used to design and implement the system. These pilot groups were also involved during User Acceptance Testing (UAT).

The Survey Metadata Capture Tool can be used by different users depending on the roles that they were assigned. For example, a Capturer could capture metadata but this must be approved by an Approver, who is usually the supervisor or manager. There is also a role of viewer, whereby metadata could be viewed but the rights are restricted. For example, a viewer cannot edit, change or approve metadata.

The network infrastructure for both development and user environments is supported by the IT department. This includes configuring the environments as well as housing the different servers in the data centre of the organization. The databases are also managed by the IT department. The ESDMF is based on the Linux open source operating system. Because the IT department does not have the skills to service and maintain this environment, we have outsourced these services from a private company. However, this is done in conjunction with the IT department, who are in the process of raising their skill level in order to be able to support the ESDMF in the Linux environment.

During User Acceptance Testing (UAT) any identified defects were logged on the CA Unicentre system, which is used for IT help desk support. With the help of the IT help desk technicians, we were able to customise the system so that the unique categories of defects for the ESDMF system could be recorded.

The IT procurement group was used to procure all the hardware and software used in the development and deployment of the system.

The development of the ESDMF was not done in isolation of the existing projects within Stats SA. For example, the following projects were ongoing and in parallel with the development of the ESDMF:

- SAS 9 migration
- Re-engineering of other surveys
- Community Survey 2007
- Census 2011

Some members of these other projects were also involved in the development of the requirements and review of the architecture of the ESDMF. The goal is to ensure that we do not do things in isolation so that we can share our knowledge and ease the integration of the new system into existing systems.

Staff from the Methodology and Standards division was seconded to the ESDMF project. Their role was to develop policies, procedures and standards for the system. Our development process is that policies are developed and approved. Thereafter, the procedures and standards are developed. So, for each phase, the policies are used to develop and implement the system deliverables for that phase.

For example, for the first phase, we developed a policy for Data Quality and a policy for Metadata. As a result, Phase One was focused on capturing metadata (Metadata policy) in order to ensure quality of the output product (Data Quality policy). For the Second Phase, we already have approved policies for Concepts and Definitions as well as for Classifications.

5.2 Description of the team/individuals involved in development and maintenance of metainformation systems.

5.2.1. System Developers

The deliverables expected from the supplier include a Skills Transfer Plan and Strategy. The goal is that the supplier will train Stats SA system developers in how the system is designed and implemented. At the end of the contract, these Stats SA developers should be knowledgeable to maintain, upgrade and/or enhance the system. Thus, we should not be dependent on the supplier for any development beyond the expiry of the contract.

5.2.2. Data Quality Officers and Specialists

The Data Quality Officers and Specialists are trained on how to use the system. They are also trained to be trainers ("train the trainer"). Once again, the supplier's deliverables includes training Stats SA Data Quality Officers and Specialists in how to train users on how to train other users to be trainers themselves.

5.2.3. Methodology and Standards Professionals

The Methodology and Standards staff members provide support by developing Policies and Standards. They are subject matter experts in survey operations. They are also involved during the design phase in order to help explain and clarify the requirements.

5.2.4. Project Managers

The Stats SA project manager works closely with the supplier's project manager. They bridge the gaps between the two organizations and make sure that the deliverables are managed properly and on time.

5.2.5. Training approaches and knowledge management

Users are required to spend at least a day in a training session, taking them through the functionality of the system as well as how to use it.

The Training Manual is used during the training sessions. The Training Manual contains complete descriptions of the system. The users can also use this document for reference purposes.

The system is designed such that tool tips (online help) are available to the user when hovering over certain areas of the user interface. These tool tips explain the features over which the mouse may be hovering. This allows the user to have information directly at a point of need without having to go through the Training Manual.

5.2.6. Partnerships and cooperation between agencies

In Latvia, we learned that during the development of their system, their outsourced supplier took a while to understand the business of the statistical organization. It came as no surprise when we ran into similar problems with our supplier, as much as we were not happy about it.

Their Integrated Statistical Data Management System (ISDMS) uses Bo Sundgren's model of metadata system, which they used as a firm foundation for the theoretical definition of metadata. We learned the importance of having a solid foundation in the definition of metadata

In Ireland, we learned about the issues regarding communication between the customer and the supplier. Additionally, they had the same problem as in Latvia in that the development of their system also took longer than originally planned. This happened even after Ireland provided very detailed documentation on most of the major aspects of the system. Once again, when we ran into similar problems, we were not surprised, as much as we did not like it.

In Slovenia, their metadata model is also based on Bo Sundgren's model, with some modifications in areas where they believe that their components are adequate to meet Bo Sundgren's requirements for a metadata system.

Their development model is to build the system in-house and outsource when they get to maintenance phase. They continuously re-skill and train their staff as they bring in new technologies aboard.

From New Zealand, we adopted a few of their practises. For example, we brought in the Statistical Value Chain into Stats SA. This is how we view the business of statistical production processes within Stats SA. We also adopted the way they broke down metadata into five categories, namely, definitional, operational, system, dataset and procedural/methodological metadata. One of their experts helped us to evaluate the respondents to the tender for the development of the ESDMF.

In our trip to Australia, we learned that in order to have a successful data warehouse project, there is a need to develop policies and standards which will define how the system should be designed. When we returned to South Africa from that trip, we restructured the team into two groups, the Policies and Standards team and the Technology team. The Standards and Policies team developed policies and standards which were used by the Technology team in the development and implementation of the ESDMF.

Experts from Sweden occasionally came to Stats SA to advise us on various aspects of metadata and statistical production processes. For example, a few years ago, Bo Sundgren, a well known expert on metadata, came to Stats SA to advise us on how to proceed in the development of a metadata system. Recently, another expert from Stats Sweden came to conduct a workshop on SCBDOK, the Stats Sweden metadata template. He also conducted training on quality definition and quality declaration of official statistics. This gave us a better idea on how to develop a data quality template, as well as how data quality should be reported on.

Last year (2006), we met Alice Born (from Stats Canada) when we attended the METIS conference. We engaged her regarding their development efforts of their metadata system, Integrated Metadata Data Base (IMDB). We applied that knowledge during the development of our Survey Metadata Capturing Tool.

Consultants from Canada help us in other projects within Stats SA. During their tenure we engage them for advice and other consultation.

We used the Corporate Metadata Repository (CMR) model by Dan Gillman, from the US Bureau of Statistics in our understanding the metadata model, especially with regard to the ISO 11179 Specification. We also sent our metadata model to him and other metadata experts for review and critique.

5.3 Organizational Change Management

5.3.1. Climate and Culture Assessment

Preliminary Organisational Change Management (OCM) initiatives necessitated a review of the operating culture at Stats SA in order to understand the 'lie of the land' in which the system will be introduced. The information contained in the Culture & Climate Assessment was obtained through a number of OCM diagnostic interventions, targeted specifically at internal stakeholders. This was done by holding focus groups as well as running an online survey via Stats SA intranet website.

A key challenge to Stats SA is to focus the organisation on the strategic importance of the DMID project, not only in as far as it assists an individual in their immediate job function, but even more importantly how it contributes to the overall wellbeing of the South African society at large and the contribution it makes to strategic decision making at government level. DMID communication messages need to create a sense of higher purpose to help individuals with long term strategic thinking.

5.3.2. Change Readiness Assessment

A Change Readiness Assessment was conducted to determine the current capacity of Stats SA to change, and to identify areas of resistance towards DMID requiring Organisation Change Management (OCM) interventions.

The Change Readiness Assessment was conducted via a survey and series of focus groups.

The following 'change readiness dimensions' are integral to enable commitment towards DMID and formed the basis of the Change Readiness Assessment:

- Clear vision
- Effective leadership
- Positive experience with past change initiatives
- Motivation to do the project
- Effective communication
- Adequate project team resources

5.3.3. What is Change Readiness?

OCM is a critical, although often bypassed element in organisations. It focuses on the 'human response to change', helping people understand, accept and commit to a new way of working. One of the key upfront steps in the change process is the Change Readiness Assessment.

The Change Readiness Assessment is a process used to determine the levels of understanding, acceptance and commitment likely to affect the success of the planned change. Change readiness is gauged along an axis known as the Change Commitment Curve, which is depicted below:

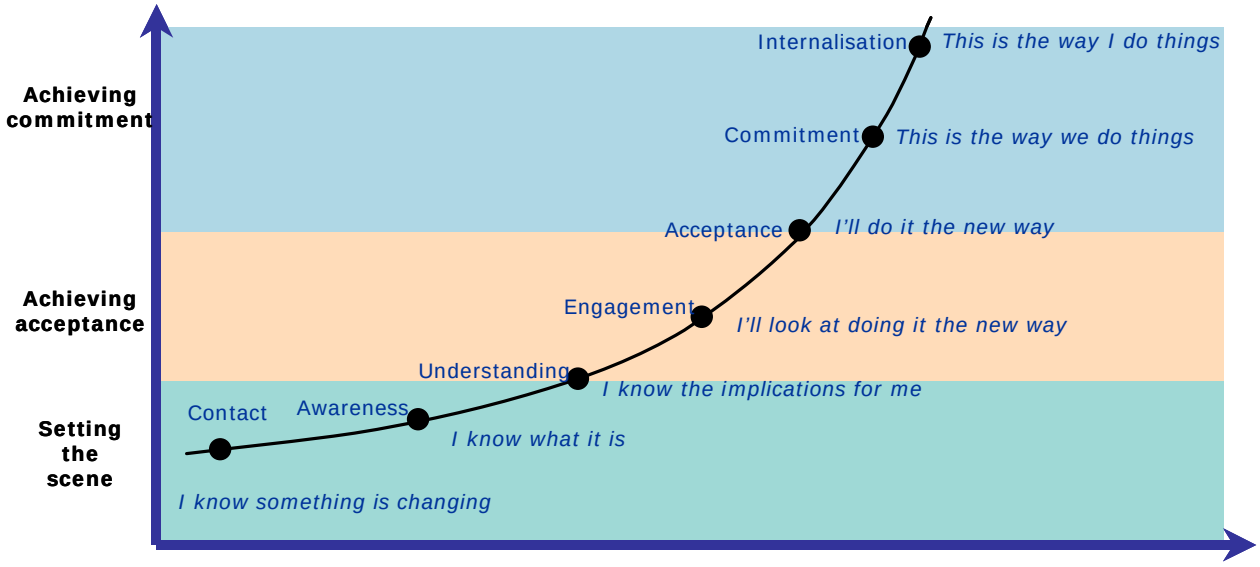


Figure 13: Change Commitment Curve

As the DMID project phases roll out, different stakeholders will need to be at specific levels of commitment. The level of commitment required will be dependent on the role they play in the DMID project and their ability to influence the program. The Change Commitment Curve will provide a framework for understanding and tracking the requisite levels of commitment that stakeholders need to be facilitated through so that OCM interventions can be developed accordingly.

A Change Readiness Assessment will become an obligatory OCM intervention prior to the rollout of a new phase on the DMID project.

5.3.4. Findings

The following were the finding from the assessments:

- Executive Management does not have the same understanding of the DMID project.
- Lack of communication between management and sub-ordinates; this makes it difficult for sub-ordinates to understand the purpose of the project and the impact it has on their working lives.
- Lack of support from Executive management will result to resistance and difficult success of the project
- If management does not communicate, does not understand, and does not promote the project, it will result in difficulty to deliver the message and get buy-in from staff in the organisation.

5.3.5. Next Steps from the Findings

The findings of the assessments resulted in identifying where some of the key staff members belonged on the Change Commitment Curve. In general, most were in the "Setting the Scene" and "Achieving Acceptance" area bounded by in time by "Contact" ("I know something is changing") and "Understanding" ("I know the implications for me"). Obviously, a lot of effort is needed in order to move from that area to "Achieving

Commitment" demonstrated by "Internalisation" wherein staff can claim that "This is the way I do things"

Another outcome of these assessments was to organize a Leadership Alignment workshop. In this workshop, the Executive Committee was given a presentation of the findings and the path forward. The path forward is to ensure that the leadership understands the goals of the project and how they line up with the vision of Stats SA. The leadership was also instructed on how to communicate the same message about the project.

6. Attachments & Links

Documents to be attached:

1. Survey Metadata Standard Template (*Survey Metadata Capturing Tool_v0.10.doc)*
2. Flow Chart *(Flowchart4(4).vsd)*
3. Web page of the Metadata Capture Tool in MHT format *(Summary of Survey Metadata Record.mht)*
4. SASQAF *(South_African_Statistical_Quality_Assessment_Framework_V05.doc)*
5. Metadata Entity Map (*Metadata entity map_V0.05_June_30.xls)*

*** END ***