Innovative technologies for statistical production: experiences of Statistics South Africa

Matile Malimabe[1] and Ashwell Jenneker[2]

1. Introduction

Statistics South Africa's development of the metadata management system has its origins in the organisation's need to develop a data warehouse. This idea came about because the organisation wanted to improve the quality of the statistics it 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 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. To be successful, a data warehouse also needs to operate in this environment.

1.1 Metadata strategy

A formal process for standardisation was developed through consultation with standards experts. A standards development and implementation life cycle was devised to monitor the standardisation process. The following is the standards development life cycle (see Appendix A).

The next step for us was to investigate the strength of our standards and policy foundation. In the course of 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 system had to form a good mix with all the other ingredients identified as 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 its name implies, the ESDMF will consist of tools and applications to support the whole statistical production process. Within this facility there is a metadata subsystem 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 are used and generated at every stage of the statistical production process.

---

[1] Statistics South Africa, Data Management and Information Delivery, 170 Andries Street, Pretoria 0001, South Africa. E-mail: matilem@statssa.gov.za.

[2] Statistics South Africa, Statistical Support and Informatics, 170 Andries Street, Pretoria 0001, South Africa. E-mail: ashwellj@statssa.gov.za.

As a data factory, a statistical organisation needs to organise and package data in ways that make it useful for the end user. Produced data must also meet certain minimum quality standards. To satisfy both these requirements, the use of metadata is indicated. In packaging its data and statistical products, a statistical organisation must ensure that they are accompanied by 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 include descriptions of concepts used in the organisation, classifications of these concepts, methodologies and business rules.

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 are part of how statistical data are presented to the end user. Such presentation engenders trust in the users.
- *Facilitation of correct interpretation of statistical data*: metadata accompany datasets and other statistical products.
- *Quality of statistics*: standard metadata contribute to the improvement of a number of quality dimensions. Standardisation of concepts and their definitions and classifications are essential ingredients of standardised metadata.

This project is aimed at supporting the strategic theme “Enhancing the quality of products and services”. Within the Data Management and Information Delivery (DMID) project, the metadata management system, more than any of its components, addresses this strategic theme.

1.2 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 deliverable 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, are the first of these metadata deliverables. The survey metadata component is fashioned along the lines of Statistics Canada’s Integrated Metadata Data Base (IMDB) Metastat.

The survey metadata component is followed by 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.

2. Statistical metadata

The essence of Stats SA’s meta-information system is captured by how the organisation uses the metadata. Metadata are used within the organisation to enable statistical production processes. This means that metadata are used during various stages of statistical production as essential inputs to production processes. However, the production processes, in turn, produce metadata. These metadata are 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 data interpretation.

2.1 Metadata classification (categories of metadata)

Because of the diversity of uses to which metadata are put, it was decided that the contents of the meta-information system should be aligned with these uses. The logical consequence 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.

2.1.1 Survey metadata

Often referred to as dataset metadata, survey metadata are 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 to the dataset.

2.1.2 Definitional metadata

These are 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 are classifications drawn from different study domains.

2.1.3 Methodological metadata

These metadata relate to the procedures by which data are collected and processed. These may include sampling, collection methods, editing processes, etc.

2.1.4 System metadata

"System metadata" refers to active metadata used to drive automated operations. Some 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
- Methods of access to databases
- Coordinates as kept in the metadata store
- Table and column definitions, schema and mappings of data.

2.1.5 Operational metadata

These are 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.

3. Metadata management tools

The developed metadata management application allows Stats SA staff members to perform a number of tasks in the metadata management process. Currently, the ESDMF consists of the following ***tools***:

3.1 Access control tool

Provides a central point for creating and managing access to all ESDMF tools.

3.2 Metadata registration tool

Provides a central registry for the registration, revision and relating of administered items.

3.3 Metadata browser

Enables users to browse and search for administered items in the central registry.

3.4 Quality tool

Enables users to set up the quality framework that is based on the South African Quality Assessment Framework (SASQAF) document, and forms the foundation for the assignment of quality indicators to surveys, the capturing of quality metadata for a survey and conducting a quality assessment of a survey.

3.5 ESDMF reporting tool

Allows users to generate static and ad hoc reports based on the data contained in the various ESDMF tools.

3.6 Survey metadata capture tool

Manages the metadata for a survey on series and instance level.

3.7 Access control tool

Provides a central point for creating and managing access to all ESDMF tools. Tools are linked to resources that can be linked to roles, and roles can be linked to users. This association establishes highly customisable user access privileges that are encapsulated in a single sign-on process. This allows a user to sign on once and then access any tool to which the user has privileges.

4. Standards and formats

4.1 Metadata registration tool

The metadata registration tool is based on the International Organization for Standardization (ISO) 11179 Part 6 standard, with some customisation for Stats SA purposes.

4.2 Quality tool

The quality tool is based on the SASQAF standard, which in turn is based on the IMF Data Quality Assessment Framework.

4.3 Metadata revisions and version control

Metadata are expected to change owing to revisions of concepts and their definitions, changes in 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 the 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.

5. Organisational and cultural issues

5.1 Roles in metadata/statistical life cycle management

In order to understand 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 organisation. 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 UAT, any identified defects were logged on the CA Unicentre system, which is used for IT help desk support. With the aid 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 development of the ESDMF was not done in isolation from 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 were seconded to the ESDMF project. Their role was to develop policies, procedures and standards for the system. In our development process, policies are first developed and approved. Thereafter, 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 1 was focused on capturing metadata (metadata policy) in order to ensure the 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 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 organisation. It came as no surprise when we ran into similar problems with our supplier, much as we were not happy about it.

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

In Ireland, we learned about the cultural 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.

In Slovenia, the 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.

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

We adopted a few practices from New Zealand. For example, we brought 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 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.

In 2006, we met Alice Born (from Stats Canada) when we attended the METIS conference. We engaged her regarding her agency's efforts to develop its metadata system, the 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 of 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.

6. Organisational change management

6.1 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 and climate assessment was obtained through a number of OCM diagnostic interventions, targeted specifically to internal stakeholders. This was done by holding focus groups as well as running an online survey via the Stats SA intranet.

A key challenge to Stats SA is to focus the organisation on the strategic importance of the DMID project, not only insofar as it assists individuals in their immediate job function, but even more importantly, how it contributes to the overall well-being of South African society at large and the contribution it makes to strategic decision-making at the governmental level.

6.2 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 OCM interventions.

The following change readiness dimensions are integral to enabling commitment to the DMID and formed the basis of the change readiness assessment:

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

6.3 What is change readiness?

OCM is a critical, although often bypassed, element in organisations. It focuses on the human response to change, helping people to understand, accept and commit to a new way of working. One of the key first 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 (see Appendix B).

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 programme. The change commitment curve will provide a framework for understanding and tracking the requisite levels of commitment that stakeholders need to be assisted in attaining 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 of the DMID project.

6.4 Findings

The following were the findings from the assessments:

- Executive management does not have the same understanding of the DMID project.
- Lack of communication between management and subordinates makes it difficult for subordinates to understand the purpose of the project and the impact it has on their working lives.
- Lack of support from executive management will result in resistance to the project and make success difficult.
- If management does not communicate, does not understand, and does not promote the project, it will have difficulty in delivering the message and obtaining “buy-in” from staff in the organisation.

6.5 Next steps from the findings

The findings of the assessments made it possible to identify 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 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”, as demonstrated by “internalisation”, wherein staff can claim that “This is the way I do things”.

Another outcome of these assessments was to organise 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.

7. 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 recognise 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.

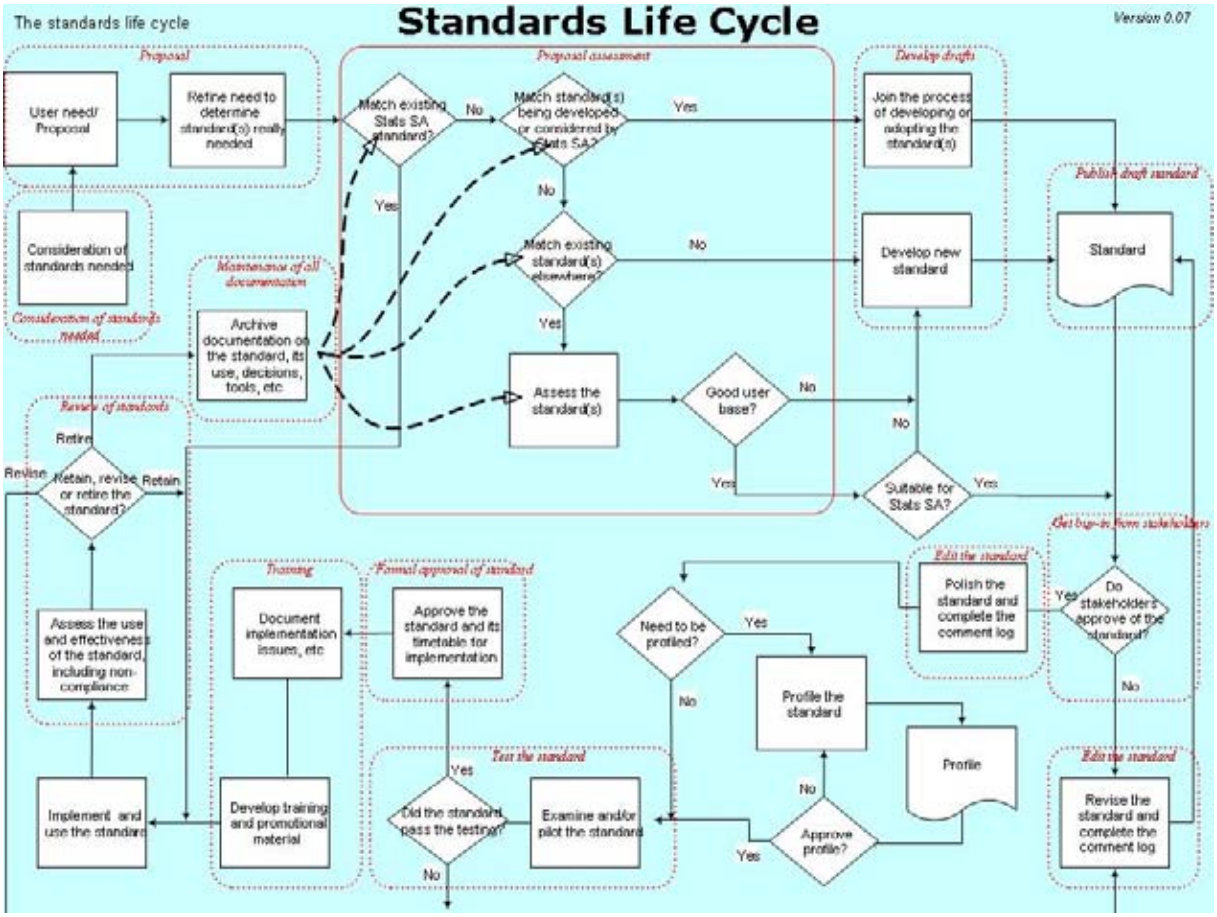
For a project of this magnitude (three years) and complexity, we decided to break down the deliverables into 12 phases. Each phase was planned to be three months long. Also, each phase was envisaged as a complete deliverable in its own right, even though the next phase was designed to build on the previous 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.

Phase 1 took a very long time, in view of the time allocated for the project, owing to challenges uncovered in the capabilities of the service provider, as this was a new undertaking for it, and in capacity as a result of staff turnover (for both Stats SA and the service provider).

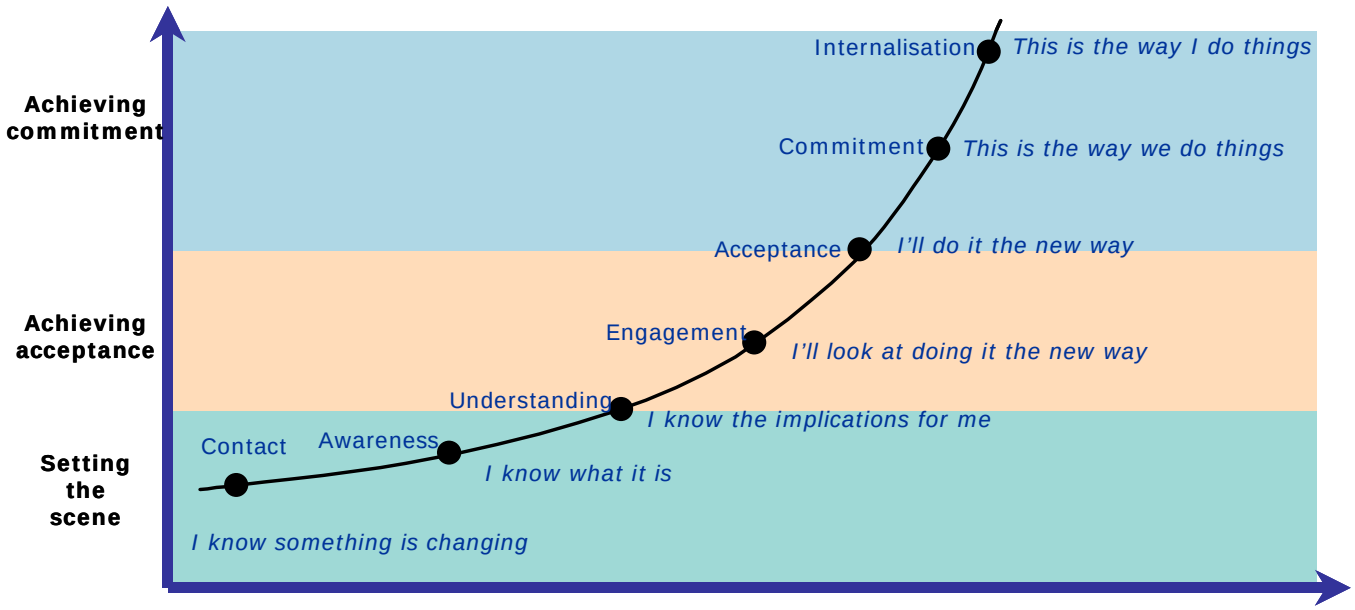
It became evident that the approach was not effective, as the project was only in phase 2 out of 12 phases after two thirds of the time allocated had elapsed. An alternative approach was then put in place for progress to be based on tool delivery versus the phase approach.

It was during this process that the service provider instituted some claims against Stats SA, which has led to a disassociation pending a legal conclusion, thus endangering the DMID project. At this stage only seven out of 52 tools have been delivered.

Appendix A:
Standards development life cycle



Appendix B:
Change commitment curve



8. References

SASQAF (South_African_Statistical_Quality_Assessment_Framework_V05.doc).

Survey metadata standard template (Survey metadata capture tool_v0.10.doc).

Web page of the survey metadata capture tool in MHT format (Summary of survey metadata record.mht).