WP.19
ENGLISH ONLY

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

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Geneva, 3-5 April 2006)

Topic (iii): Metadata and the Statistical Cycle

RE-ENGINEERING PROJECTS FOCUSSING ON METADATA AND THE STATISTICAL CYCLE
Invited Paper

Submitted by Statistics South Africa, South Africa[1]

I. PURPOSE OF THE PAPER

1. This paper briefly introduces the current work on metadata management being undertaken in Statistics South Africa (Stats SA). It provides the reader with an understanding of the anticipated scope, implementation issues and identified benefits of metadata management through the entire statistical value chain. It also presents the work done thus far in developing the requirements for implementing a metadata store as part of the End-to-end Statistical Data Management Facility (ESDMF).

II. BACKGROUND

2. Stats SA, as South Africa's official statistics agency, is responsible for informing users of the concepts, classifications and methodologies used in collecting, processing and analysing data, the quality of that data, and all other relevant features describing the data.

3. The vision of Stats SA is to be *the preferred supplier of quality statistics*. The achievement of this will promote a culture of evidence-based planning and decision-making in pursuit of socio-economic development and good governance. The development and application of world-class standards, classifications, methods and procedures are central to the drive for quality statistics.

[1] Prepared by T.J.Lukhwareni, josephl@statssa.gov.za, A.C. Jenneker, ashwellj@statssa.gov.za and E. Gavin lizg@statssa.gov.za

4. Stats SA has long recognised that the centralised storage of data is a key component for improving the quality of data. The prospect of developing a data warehouse has been debated since at least 1997. Data warehousing options were evaluated in consultation with other statistical agencies. In 2004, Stats SA adopted the business case for the development of such a facility. The requirement for the data warehousing infrastructure was to consolidate the management and processing of data and metadata. Two key deliverables were identified:

- a central data storage facility in which statistical data is maintained in a standard manner, together with a set of tools for retrieval, analysis and report-generating;
- a central metadata store where data and information needed to interpret the data is stored according to standard, uniform and agreed fields and formats. This would include a store for classifications, concordances, code files and sample frames.

5. A business requirements gathering exercise indicated that, while a data warehouse may be a necessary element in addressing quality issues throughout the statistical value chain, it is not sufficient to solve these issues. Proceeding from this premise, a consultant retained by Stats SA subsequently proposed the high-level conceptual scope shown in figure 1. Statistical production units within Stats SA also gave input on the high-level conceptual scope.

6. The End-to-end Statistical Data Management Facility (ESDMF) proposed consists of a number of functionally integrated but modular conceptual components that will address specific aspects of the quality improvement through the statistical value chain. The high-level scope diagram outlines six main components in the ESDMF: the Collections Environment, Central Data Store (CDS), Dissemination Area, Workflow Control, Metadata Store and Access Control. (Respondent and client management are out of the scope of this facility).

7. The aim of the Workflow component is to control the definition and execution of formalised processes in a consistent, repeatable and standardised manner that can support delivery of improved statistical products. The processes defined for execution in the Workflow component will support the complete statistical value chain.

8. This component will act as the link between the actions performed within a specific project and the metadata stored in the Metadata Store. Metadata and other data such as control measures will be used to validate completed tasks, and to govern tasks to be performed. The Workflow component will record, in an audit trail, all actions performed on the data from collection to dissemination.

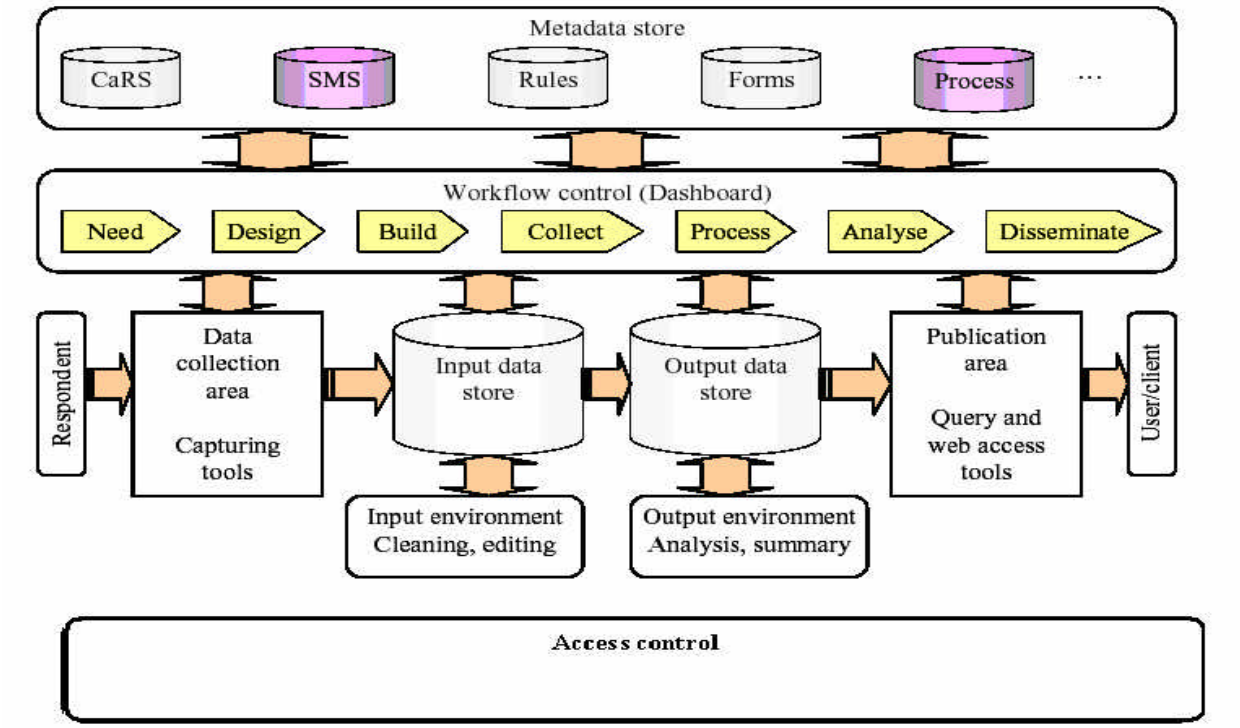

Figure1: High level conception of the End-to-end Statistical Data Management Facility (ESDMF)
CaRS: Classifications and Related Standards
SMS: Survey Management System

III. METADATA MANAGEMENT

9. "*Metadata management refers to the content, structure, and designs necessary to manage the vocabulary and other metadata that describes statistical data, designs and processes. …(It) includes the development of metadata models, building metadata registries to organise the metadata, development statistical terminologies which define and organise terms*" (Bargmeyer and Gillman, METIS 2000).

10. The management of metadata includes the following elements:

- understanding what metadata is;
- understanding the scope of metadata;
- establishing metadata owners, maintainers and regulators;
- ensuring that data users have easy access to relevant data and metadata in a timely manner;
- determining appropriate technical storage environments i.e. the metadata store;
- registration of metadata items.

A. The definition of metadata

11. Strictly speaking, there is no such thing as a perfect or absolutely accurate dataset. One of the functions of metadata is to document data structure, assumptions and limitations so that data are not used inappropriately, and so that users can easily understand limitations on use. A standard format for metadata is required to facilitate appropriate use and understanding of the data. This can be achieved by producing metadata that fits the content of data and the needs of authors, information providers and end-users.

12. Stats SA adopted the definition of metadata proposed by Statistics New Zealand (SNZ) [5]. Metadata is "*data about data, and refers to the definitions, descriptions of procedures, system parameters, and operational results which characterise and summarises statistical programs. Metadata may be passive (descriptive), i.e. the*

form of documentation, which is used by agency staff, or may be active (prescriptive), i.e. determining the actions of automated surveys processes".

13. Metadata can be further categorised into various categories. Stats SA distinguishes five separate categorisations:

- Definitional metadata: metadata that describes the concepts in the data, e.g. classifications, question and question modules for collection instruments and statistical terms;
- Operational metadata: metadata arising from and summarising the results of implementing the procedures, e.g. response rates, edit failure rates, costs and other quality and performance indicators;
- System metadata: active metadata used to drive automated operations, e.g. publication or dataset identifiers, date of last update, file size, access methods to databases and mapping between logical names and physical names of files;
- Procedural/methodological metadata: metadata relating to the procedures by which data are collected and processed, e.g. sampling, collection methods and editing processes;.
- Dataset metadata: metadata used to describe, access and update the dataset and data structures, e.g. textual description, data cell annotations, dataset title, keyword and distribution information.

14. These metadata categories may be interrelated. In particular, the metadata used to describe fully a statistical dataset could include elements of the first four categories listed above. Such information could be necessary for the appropriate use of the dataset. An analogy is that of the evolution of the documentation of geographical information, to include the explicit recording of processing steps (the *lineage* of the data) and quality indicators.

15. Metadata usually includes information about the metadata itself. This describes important features about metadata documentation such as its author, technical form, publisher, etc. This kind of information may have predefined descriptions of the information to be included and how this information is to be represented. In this context, metadata can be described as "metadata about metadata". Metadata about metadata is needed, for example, to facilitate the author's work by documenting important details, and to keep a record of important facts about the metadata updates and related issues.

B. **The Metadata Store**

16. Stats SA's Metadata Store will manage the definition and updating of metadata needed to maintain standards and to record and assist in controlling the processes throughout the statistical value chain. It must be centrally available and provide a single source of approved and reusable metadata for metadata creators and users. The Metadata Store must allow new metadata to be defined to support growth in this central repository of metadata. It will include systems such as the Classifications and Related Standards (CaRS), acquired from SNZ, which provides a central repository for classifications, concordances, concepts and definitions, and codefiles for use by all components within the organisation.

17. Metadata occupies an important role in every phase of the statistical value chain and, if developed appropriately, can enhance the quality of data and the statistical production process. Stats SA's proposed Metadata Store serves several important functions for development and storage of systematic and coherent information and data management. Besides its annotative purpose, the metadata store has an important role in maintaining and improving data quality through the following functions:

- Providing a centralised store to enable and encourage collaborative sharing and re-use of definitions and standards throughout the organisation;
- Ensuring a consistent interpretation of standards and definitions across all aspects of surveys and over time;
- Improving data quality by enforcing compliance to workflow controls and procedures;
- Minimising human error by regulating and automating downstream statistical processes.

C. **Metadata in the statistical value chain**

18. Stats SA is in the process of developing a metadata entity map that highlights the grouping of metadata elements within the statistical value chain (see Annex 1). This map also provides a link to the sources of metadata and a brief description of each element.

19. The statistical value chain can be defined by the following phases:

- Need: the need phase is entered when a new project is defined. In this phase the request is evaluated, the feasibility for a new project is investigated, the project is planned and a budget is developed.
- Design: in the design phase the project is planned in detail, the survey methodology is developed, the questionnaire is designed, the sample defined, and the data capturing tool is designed.
- Build: in the build phase the data capturing tools are developed, tested and implemented.
- Collect: in the collect phase data is collected and captured using the data capturing tools.
- Process: in the processing phase the captured data is processed. This phase includes validation, correcting, editing and imputation. The result of this phase is a clean dataset.
- Analyse: in the analysis phase all analysis on the data is performed. The result of this phase is a publishable dataset.
- Dissemination: In the dissemination phase publications are created from the dataset produced by the analysis phase, and disseminated in various forms.

D. **Metadata Standard**

20. A metadata standard outlines the characteristic properties to be recorded, as well as the values the properties could or should have. Standardisation of metadata documentation facilitates wider information sharing and usage. The use of metadata standards enables producers to describe datasets fully and coherently. The adoption of a standard also facilitates data discovery, retrieval and use. If a standard is used, finding a specific piece of information in a metadata record is far easier than if no standard is used. Standards also enable automated search and retrieval functionality.

21. Developing or adopting a metadata standard is a complex process often involving different and not always complementary interest groups. There is often significant overlap – and possibly even inconsistency - as one standard is extended into the territory of another. A number of bodies with which Stats SA interacts, both international and local, are involved in generating standards. These include the International Standards Organisation (ISO) the International Monetary Fund (IMF) and the national standard setting body, namely the South African Bureau of Standards (SABS).

22. Adoption or adaptation of an existing metadata standard holds the dual advantages of minimising development effort and ensuring a common understanding of the metadata by the existing community of users of that particular standard. A further consideration for Stats SA is to limit the burden of compliance by alignment to existing metadata standards as far as is possible. For example, many of the datasets disseminated by Stats SA have an explicit geographic component. In terms of national legislation, metadata on these datasets must be made available in accordance with a prescribed standard for documenting all geographic information datasets developed using public funding.

23. Stats SA is involving the data production components from economic, social and population statistics components as well as supporting functions, such as the ICT and geographic divisions, in the development of metadata standards to meet the organisations needs without overburdening the originating components.

24. Stats SA’s Data Management and Information Delivery (DMID) project is evaluating the following metadata standards and systems:

- **ISO/IEC 11179**
  ISO/IEC 11179 is an international standard for the specification, and standardisation and management of data elements through metadata registries. ISO/IEC 11179 gives the description of a registry of metadata describing any data.

  Part 4 of this standard provides guidance on how to develop unambiguous data definitions. A precise, well-formed definition is one of the most critical requirements for shared understanding of an administered item. Well-formulated definitions are imperative for the exchange of information.

  Part 6 of the standard provides instruction on how a registration applicant may register a data item with a central registration authority, and how to allocate unique identifiers for each data item. Maintenance of administered items already registered is also specified in this part of the standard.

- **SANS 1878**
  The South African Spatial Metadata Standard SANS 1878 is a profile of the international standard ISO 19115: Geographic Information/Geomatics - Metadata. This standard defines almost 300 metadata elements, with most being listed as 'optional'. The metadata within the standard is an aggregate of the following elements: identification, constraints, data quality, maintenance information, spatial representation, reference system, content information, portrayal catalogue reference, distribution, metadata extension information and application schema information.

- **SCBDOK**
  Statistics Sweden's metadata system consists of a number of tools and templates. SCBDOK [10] is a documentation model that specifies the basis for the different statistical production methods. The SCBDOK template is the cornerstone of Stats Sweden's metadata system. Metadok, a software tool, provides a system for creating formalised metadata for the purposes of describing final observation registers in SCBDOK.

E. Principles and benefits of metadata management

25. Benefits of investment in a metadata management programme and associated metadata store include increased re-use of corporate assets, improved impact analysis associated with these assets, increased quality for decision-making, reduced development and maintenance time, greater success in deployment of new organisation capabilities, improved user access and usage, and better understanding of corporate assets. Metadata serves as the link between business processes and data, applications and the technical infrastructure.

26. The principles guiding Stats SA's design of a metadata store include the standardisation of metadata across the organisation in so far as possible, the reuse of metadata and the alignment of metadata standards with existing standards used by key stakeholders where possible. The aim is to limit the compliance burden in order to engender organisational support and rapid uptake of new metadata standards.

IV. RECOMMENDATION AND CONCLUSIONS

27. Management of metadata is not just a technical issue for the data producers. It provides end-users of statistical products and data with adequate information for proper use. Conforming to a metadata standard is important to ensure that users can find, understand and share data.

28. Continuous updating at every stage of the statistical value chain is necessary for an efficient metadata management system. It is not enough to update and capture metadata once; it has to be checked and, if necessary, modified at every stage of the statistics value chain. Technical solutions ensuring that metadata can easily be registered and updated together with the data, at one place, is a precondition for an efficient metadata system. This is a great challenge for statistical agencies working with disparate metadata systems.

29. Management of metadata at Stats SA has moved from being an afterthought, to becoming the core of the improvement of quality of the data and production process.

V. REFERENCES

[1] Gillman, D. (2004) *Metada Registries, Data Harmonization and Maximizing Use of Warehouse,* US Bureau of Labor Statistics.

[2] ISO/ IEC Standard 11179 (1997). *Specification and standardization of Data elements. International Standard Organization*, Geneva, Switzerland. ISO/IEC 11179 standards.

[3] Jenneker, A. (2004), *The Business Case for a Data Warehouse in Statistics in South Africa*, Draft version 0.04, Statistics South Africa , Internal electronic document: DMID_DW_BC_V0-04.doc.

[4] Johanis, P 'et al' (2003). *Paper for the Open Forum 2003 on Metadata Registries* 20-24 January 2003, Santa Fe, New Mexico, USA.

[5] Merrington, R (2004) *Creating a New Business Model for a National Statistical Office of the 21st" Century*, Statistics New Zealand.

[6] Oakley, G. (2004) *Report of visit to Statistics South Africa to advice about the Data Warehousing Project* 31 May to 4 June 2004.

[7] Oakley, G. (2004) *Revised Paper from Nov 03 IRMC discussion of 'Strategy for End-to-End Management of ABS Metadat'*.

[8] Lukhwareni, T.J. 'et al' (2005). *Management of Metadata in National Statistical Agency*. Commonwealth conference paper.

[9] Statistics Norway (1998). *Guidelines for statistical metadata on the Internet. Statistical Journal of the United Nations ECE* 15 (1998) 169-176

[10] Sundgren, B. (2000) The Swedish Statistical metadata system. Eurostat conference, 2000.

[11] Lukhwareni, T.J. 'et al' (2006). Data *Quality Policy, Statistics South Africa*.

[12] Madonsela, S.F. 'et al' (2006). Metadata *Entity Map, Statistics South Africa*, draft.

[13] Lindblom, H. and Sundgren, B. (2004) *The metadata system at Statistics Sweden in an international perspective*.

ANNEX 1: METADATA WITHIN THE STATISTICAL VALUE CHAIN

1. Need 1.1 Determine information requirement • Evaluate the request • Investigate the feasibility of the project • Identify stakeholders • Consult stakeholders • Acquire feedback from key stakeholders • Create proposal and approval, etc 1.2 Develop detailed project plan • Create detailed task plan • Acquire feedback from key stakeholders • Revise plan • Approve plan 1.3 Develop Budget Plan • Prepare initial budget and plan • Outline and quantify costs and benefits • Obtain Approval	**2. Design** 2.1 Develop survey methodology • Determine detailed objectives • Determine survey options • Perform design analysis • Produce design recommendations • Sign off survey methodology • Produce survey methodology specification 2.2 Design operational requirements • Produce collection requirements • Produce processing requirements • Produce output requirements 2.3 Design and testing questionnaire • Develop and test questionnaire • Deliver questionnaire recommendation 2.4 Sampling • Perform reselection • Perform sample maintenance • Confirm sample
3. Build 3.1 Printing questionnaire 3.2 Build technology solution • Analyse or refine technical specifications • Create system solutions • Detailed module design • Code application 3.3 Test technology solution • Programmer testing • Peer review • Client acceptance testing • Client sign off 3.4 Implement technology solution-training and piloting • Controlled release • Support and training for users • Documentation	**4. Collect** 4.1 Field work • Receive questionnaires • Maintain contacts details • Receive administrative data • Follow ups 4.2 Manage respondents • Review sample • Pre-contact respondents 4.3 Close off collection • Monitor response rate • Sign off collection
5. Process 5.1 Capture data into electronic form • Batch questionnaire • Input data • Validate data 5.2 Perform macro editing • Compare data with data from	**6.Analysis** 6.1 Examine source data • Ensure data structures are correct • Assess whether observed changes reflect expectation • Compare data with data from previous periods • Compare data with other data source

previous periods • Compare data with other data sources • Investigate outliers 5.3 Run imputations or estimations • Tax modelling • Identify unlinked and special treatment • Run imputations • Monitor imputations 5.4 Produce clean datasets • Prepare handover report • Peer review • Release for analysis	• Investigate outliers 6.2 Produce statistical results • Create estimates for total survey population • Produce derived data • Apply seasonal adjustments • Apply experts knowledge adjustments • Customise analytic applications(where required) • Derive sampling errors 6.3 Validate results • Assess whether estimates reflect expectations • Compare derived data with derived data from previous period • Compare derived data with other source data • Assess quality measures • Check output have been calculated correctly 6.4 Interpret results • Determine explanation 6.5 Prepare content for dissemination • Confidentialise data • Indicate quality of data • Produce content • Approve content for release 6.6 Conduct quality control • Document main findings for process improvement • Review Methodology • Get client feedback
7. Dissemination 7.1 Receive and validate draft content • Receive dataset/content • Perform quality assurance of data content • Perform editorial changes • Editorial and content sign off 7.2 Manage and load dissemination repositories • Load repositories 7.3 Prepare pre-release for publishing • Prepare and populate tables • Apply corporate formatting standards	

• Prepare electronic distribution • Prepare hard copy outputs 7.4 Manage first release • Manage releases 7.5 Handle customer enquiries • Receive enquiry • Categories enquiry • Respond • Customer follow up	