WP.2
ENGLISH ONLY

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

PROPOSAL FOR STRUCTURE OF PART C OF THE COMMON METADATA FRAMEWORK
Part C: Metadata and the Statistical Cycle

Submitted by the Common Metadata Framework (CMF) Editorial Board*

Background

1. Based on the discussion at the 2006 METIS work session, and in the Task Force, it was agreed that the Common Metadata Framework Manual would have a high level structure of 4 parts. The following broad guideline was agreed.

Part A: Metadata in a Corporate Context - covers WHY we should have Statistical Metadata Systems.

Part B: Metadata Concepts, Models and Registries - covers the 'theoretical' underpinnings ie the container to discuss models, standards, etc.

Part C: Metadata and the Statistical Cycle - covers WHAT we need in context of the business of the organisation.

Part D: Implementation - covers HOW.

From the 2006 METIS Meeting Report

2. Attachment A to this note contains an extract from the April 2006 METIS Meeting report about the session on 'topic (iii) - Metadata and the Statistical Cycle'. A possible new structure for Part C of the manual was suggested (see paragraph 14 of the attachment) and a number of issues were raised for consideration in terms of content of the Part.

Proposal for Part C

What is the Purpose for this Part?

3. Statistical metadata is relevant to all stages of the statistical processing cycle, although many statistical agencies have had an initial focus on metadata being published with statistics during the dissemination phase. There are a growing number of examples of statistical agencies viewing statistical metadata as part of their data collection processes. [A common title is end-to-end (E2E) process, meaning from one end of a chain of processes to the other end. The METIS Framework Manual will continue to use the term 'statistical cycle', as in the title of this Part.] Some examples of the involvement of statistical metadata in an integrated set of processes might be:

a. Data elements are defined in the 'Survey Planning and Design' phase - both data elements that are obtained from providers or administrative sources, and data elements that are derived in other processes - and are then used in subsequent phases, such as 'Input Processing', 'Analysis' and 'Dissemination''.

* Prepared by Graeme Oakley (Australian Bureau of Statistics), June 2006

b. Creation and capture of information about the quality of a statistics, for example the response rate for a survey captured in 'Data Collection', edit error rates captured in the 'Input Processing' process, and RSE's captured in 'Derivation, Estimation and Aggregation' process. This information is used to generate quality metrics for use in processes such as 'Analysis', 'Dissemination' and 'Post Survey Evaluations'.

4. So this part of the manual should contain information, case studies, best practices and other material to assist the metadata developer in a national statistical office design and develop a statistical information system relevant to the business requirements of their office. To do this, it will be necessary to provide some thoughts and examples of:

- possible process models for the statistical cycle;
- what metadata is created in each stage of the statistical cycle;
- what metadata is required and used in subsequent stages of the statistical cycle (demonstrating re-use);
- issues to consider in designing metadata processes linked to the statistical cycle eg IT issues such as architecture, tools; project management issues such as risks, change management, versioning; legal issues such as confidentiality; and people issues such as roles, training.

Response to Issues raised in METIS discussion

5. One of the issues raised in the METIS discussion was the question of terminology with respect to use of terms 'survey', 'questionnaire', 'statistical collection'. (See attachment A, paragraph 12 (c)).

6. I think that we should select terminology to suit our purposes from what already exists. Given the involvement of OECD and Eurostat in compiling the Metadata Common Vocabulary (MCV), I looked there and following are some relevant definitions from which we should make our choice.

Survey
An investigation about the characteristics of a given population by means of collecting data from a sample of that population and estimating their characteristics through the systematic use of statistical methodology.
Source: Economic Commission for Europe of the United Nations (UNECE), "Terminology on Statistical Metadata", Conference of European Statisticians Statistical Standards and Studies, No. 53, Geneva, 2000.
Context: The term survey covers any activity that collects or acquires statistical data. Included are censuses, sample surveys, the collection of data from administrative records and derived statistical activities. (Statistics Canada, "Statistics Canada Quality Guidelines", 4th edition, October 2003, page 7)

Sample survey
A survey which is carried out using a sampling method, i.e. in which a portion only, and not the whole population is surveyed.

Source: The International Statistical Institute, "The Oxford Dictionary of Statistical Terms", edited by Yadolah Dodge, Oxford University Press, 2003

Data Collection
The process of gathering data.

Source: Economic Commission for Europe of the United Nations (UNECE), "Glossary of Terms on Statistical Data Editing", Conference of European Statisticians Methodological material, Geneva, 2000
Context: Data collection encompasses such concepts as: the type(s) of interview used for data collection (e.g. personal or by telephone, paper and pencil, facsimile, computer-aided personal or telephone interview (CAPI/CATI), or mailed questionnaires); the duration of the field work (specify the dates); the period used for data collection; whether a permanent survey organisation exists or personnel for each survey round are recruited, etc. Data may be observed, measured, or

collected by means of questioning, as in survey or census response.

Administrative data collection
The set of activities involved in the collection, processing, storage and dissemination of statistical data from one or more administrative sources. The equivalent of a survey but with the source of data being administrative records rather than direct contact with respondents.

Source: OECD, IMF, ILO, Interstate Statistical Committee of the Commonwealth of Independent States, "Measuring the Non-Observed Economy: A Handbook", Second Draft, Annex 2, Glossary, Paris, 2002.
Context: The administrative source is the register of units and data associated with an administrative regulation (or group of regulations) viewed as a source of statistical data.

Type of data collection
The type of data collection refers to the main process used in the collection of statistical data by the primary source of the data, those commonly used being survey data collection and administrative data collection. Each of these broad types may be further broken down on the basis of some characteristic, e.g. the nature of the data provider (enterprise / household) or exhaustiveness (sample survey, complete enumeration census).

Source: Statistical Data and Metadata Exchange (SDMX) - BIS, ECB, Eurostat, IBRD, IMF, OECD and UNSD - Metadata Common Vocabulary

Statistical processing
The processes for manipulating or classifying statistical data into various categories with the object of producing statistics.

Source: Statistical Data and Metadata Exchange (SDMX) - BIS, ECB, Eurostat, IBRD, IMF, OECD and UNSD - Metadata Common Vocabulary.
Context: In SDMX, "Statistical Processing" refers to a description of the data compilation and other statistical procedures to deal with intermediate data and statistical outputs (e.g., data adjustments and transformation, and statistical analysis). The items covered include, inter alia, weighting schemes, methods for imputing missing values or source data, statistical adjustment, and balancing/cross-checking techniques and relevant characteristics of the specific approach/approaches applied.

7. Based on the above definitions, it is **proposed** that we adopt the following terminology for the purpose of the Common Metadata Framework Manual:

a. The term Data Collection is used as the global term to cover Survey, Administrative Data Collection and the process of accounts compilation. The type of data collection concept helps to articulate this breakdown. To deal with the concept of frequency of a survey ie monthly, quarterly etc, we could refer to the individual events as a survey instance or survey cycle.] The authors of Part C might develop the discussion further to indicate the handling of sample survey and census (complete enumeration survey)

b. Statistical Processing - a group of processes applied to statistical data. The statistical cycle refers to the complete set of processes used to conduct a survey from its inception to archival.

c. For higher level terms ie higher than a survey, we could have statistical domain and statistical themes where a domain would be based on the UNECE classification and a theme is at a lower level, but could span domains. In both cases, one or more surveys may be involved in producing statistics for the domain or theme. [Consistency with the SDMX Content-Oriented Guidelines and the UNECE Classification of International Statistical Activities should be our objective.]

8. This amount of definition is probably sufficient for the purposes of Part C. Part B of the manual would be the appropriate place to go into more detail about definitions.

Structure

9. This section proposes the high level structure for Part C together with some proposals about the content of each.

1. Introduction

- Purpose of this Part of Manual
- Definitions eg data collection, statistical cycle, process
- Scope - covers direct acquisition of data from providers, use of administrative data sources and registers, special compilation activities eg financial account statistics
- Useful reference sites and references to general documents.

2. The Statistical Metadata System and the Statistical Cycle

- Model showing the statistical processing cycle, introducing the major components ie the previous structure of this Part, namely C1 - survey planning and design; C2 - survey preparation; C3 - Data collection; C4 - Input processing; C5 - Derivation, Estimation, Aggregation; C6 - Analysis; C7 - Dissemination; C8 - Post survey evaluation. [This section of the manual will provide 'context' for the detail in other sections. It will assist in organising material. However, as the editors proceed with preparation of the manual, they may find a variant of the list that works better. For example the Statistical Value Chain example from a paper presented for 2006 MSIS - see http://unece.org/stats/documents/2006.06.msis.htm]
- Model showing the Metadata store in relation to input and output data stores, workflow, data collection, dissemination etc. [Stats NZ BmTS would be a case study, along with Stats South Africa ESDMF. The developers of this Part should also investigate the Business Dimension Model of the Corporate Metadata Repository Model used by Statistics Canada, USBC and US BLS to include as a case study and possibly to inform the structure of the overall statistical cycle model.]
- Process metadata discussion. In the detail of the next chapter of this Part of the Manual, use and creation of all types of metadata eg definitional, discovery, process, quality etc metadata will be highlighted in terms of specific phases, Process and Operational metadata may not get the same explanatory treatment that definitional and descriptive context metadata will receive in Part B. Therefore, it is suggested that Part C have a good explanation of Process and Operational Metadata, by explaining what is meant, what this metadata does, examples, etc.
- In this section, the manual should also address issues related to the longer term retention of aggregate and confidentialised unit record data for researchers of the future. (This aspect could be an extension of C7, or an entirely new part C9 - Archiving and Future Access.) There is a considerable range of metadata required to fully describe a survey and its available output formats, such that a researcher can usefully access and work with data extracted from the archive.

3. Statistical Metadata in each phase of the statistical cycle

- For each of the current C1 to C8 processing phases, describe and discuss:
    - the phase in more detail eg what happens in the phase
    - metadata used (can't be an exhaustive list but a number of examples would be helpful)
    - metadata created (examples rather than exhaustive list)
    - specific issues eg quality, risks, problems related to this phase of the cycle
    - useful case studies and references.

4. Systems and Design Issues

- IT architectures eg Service Oriented Architecture (SOA), and other technology considerations
- Information about tools and techniques to link metadata repositories and statistical processing tools, such as Blaise, PC-AXIS, SAS
- standards and file formats for metadata exchange (international and defacto) could be mentioned, eg XML standards for describing questionnaires, standards for editing rules. [SDMX would get mentioned here as a standard for exchange of metadata.
- As web services are increasing as a means for facilitating access to data and also making statistical processes eg coders available for others to use, discussion of applicable standards and what metadata is required could be included in this chapter.
- Workflow systems to link together statistical processes
- Version control and regular review and updating; how to handle the changes to metadata over time
- Case studies eg metadata used for question module repositories, Stats Canada's IMDB

5. Organisational and cultural issues

- Roles and responsibilities for each participant eg subject matter statistician, senior manager, metadata manager.
- Concept and practice of 'registration' of metadata. [This matter might need to be raised to a chapter heading because of its importance and extensiveness. Statistics Canada is developing registration procedures that could be examples to guide others.]
- Issues such as change management; getting good quality metadata; centralised metadata management or decentralised - pros and cons for both
- Partnerships and cooperation between agencies
- Legal issues?
- Outsourcing development versus inhouse development
- Training and knowledge management , eg 'help' files
- Case studies

6. Links

- Case studies
- Previous METIS papers
- Other resources eg CODACMOS, METANET

Graeme Oakley
Australian Bureau of Statistics
1 June 2006

Attachment A

Extract from ECE/CES/2006/4/Add.3: Report on the April 2006 Work Session on Statistical Metadata (METIS)

11. The inputs contained in invited and supporting papers dealt with the roles that metadata plays in the statistical cycle:

(a) Central metadata repositories in support of the survey life-cycle for the whole set of statistical surveys (including administrative collections);
(b) Use of knowledge bases and document management systems;
(c) Automation in questionnaire design and more generally;
(d) Quality assurance, particularly automated editing and imputation;
(e) Machine-to machine communication;
(f) Delivery of quality related information to end users;
(g) Capturing and updating of metadata, registration and use of workflow engines in support of registration processes;
(h) Versioning: changes to mandates and characteristics of statistical surveys.

12. The following issues were raised during the discussion:

(a) Possibility to store questions and their specifications in a database, with a view to automate their inclusion in various questionnaires. There is evidence of an overlap of questions among various forms and questionnaires, and such a centralised database might optimise the design process.
(b) Some offices concentrate help files from distributed statistical systems into a single knowledge base. The example referred to at the meeting used the Lotus Notes database coupled with a documents management system.
(c) The relations between terms “survey” and “questionnaire” or “statistical observation”. One survey usually covers one statistical area, for example wholesale trade, retail trade, etc., and may use several different questionnaires that are adjusted to the type of the business that is surveyed. In this connection there is a need to clarify the terminology. The classic way makes a distinction between statistical surveys (that collect data exclusively for statistical purposes) and the use of administrative sources, while some statisticians cover both by the term “survey”. One possibility may be to use the term “statistical operation”, “statistical activity” or simply “collection”. Also need to cover operations such as accounts compilation.
(d) Centralized metadata repositories need a clear definition of responsibilities. One model is that the subject-matter units are responsible for the content, while the metadata team ensures coordination and consistency/quality checks. It is also important to synchronise the schedule of update of the metadata repository with the release calendar of statistical data.
(e) Partnerships between statistical offices, with a view to development of metadata systems, may be preferred to outsourcing development to the private sector. The disadvantage of outsourcing can be that private enterprises rarely have a sufficient knowledge and understanding of the environment to grasp the metadata issues.
(f) Linking the statistical software used in various phases of surveys with metadata repositories.
(g) Workflows are used in some offices for production as well as in management information systems. One of the applications is to set the timeliness based on the target release date and the expected workflow. Another application that might be implemented on the basis of workflow management is Customer Relationship Management (CRM).
(h) Quality information relates to both the quality of the process and metadata, and the data quality. The latter is probably more important for end-users.
(i) Discussion about costs, which were mentioned during this session, will be added to the framework as part of the review of Part A.

13. The original plan was to organize the material collected for Part C under the following headings:

C1. survey planning and design;
C2. survey preparation;

C3. data collection;
C4. input processing;
C5. derivation, estimation, aggregation;
C6. analysis;
C7. dissemination;
C8. post survey evaluation.

14. An alternate method might be to structure Part C primarily by cross-cutting issues, like versioning, linking, etc., and use the phases of statistical survey cycle as a secondary division. Both methods had their support among the participants. As a conclusion, the participants agreed to re-organize the structure of Part C. A possible new structure is:

C1. Positioning SMS in the Statistical Cycle- model of processes and metadata types relevant to each;
C2. Definitions - cover what is meant by 'survey'- what terminology to use to handle administrative data sources, account compilation etc;
C3. Understanding the cycle- using the current C1 to C8 list of processing phases, discuss the metadata used, metadata created etc;
C4. Architectures, Business and System Design, and Integrating Processing Tools;
C5. Organizational and cultural issues;
C6. Links to useful case studies, previous METIS papers, other resources.

15. The following non-exhaustive list will be considered by the Task Force on the Common Metadata Framework as potential content for Part C:

(a) Explanation of what the statistical cycle is and what metadata are needed at each stage;
(b) Organizational and cultural issues, case studies and complex methods to manage the metadata throughout the entire survey cycle;
(c) Information that might relate to architecture and good practices in transfer of metadata from central repositories to statistical tools like Blaise, PC-AXIS, etc.;
(d) Versioning, changes of metadata with time;
(e) Information on statistical operations, defining the statistical survey, use of administrative data and combination of survey data with administrative sources.