STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CES/BUR.2005/29
25 January 2005

CONFERENCE OF EUROPEAN STATISTICIANS

<u>Second meeting of the 2004/2005 Bureau</u>
Geneva, 10-11 February 2005

<u>Item 7: Progress reports on</u>
<u>ECE Steering Groups</u>

DRAFT OUTLINE OF THE COMMON METADATA FRAMEWORK
(Progress report of the Task Force)

Note prepared by the ECE secretariat

BACKGROUND

1.   The Task Force on a common metadata framework was set up in 2004[1]. The objective of the work is to organise available information about statistical metadata into a framework that can be used by national statistical organizations in their development of metadata systems. It is expected that international statistical organizations would also benefit from this work.

2.   When approving the Terms of Reference of the Task Force (CES/BUR.2004/44), the Bureau expressed its wish to be consulted after the completion of the first phase of the project: when the draft outline of the common metadata framework is developed. A draft structure of the framework and an annotated outline is presented below for consideration by the Bureau.

CONCLUSIONS REACHED BY THE TASK FORCE IN THE FIRST PHASE

3.   The structure of the draft framework is based on a proposal made by the Australian Bureau of Statistics. The Task Force agreed with the framework and considered that the focus should be on processes rather than activities. When populating the framework, it was decided to concentrate initially on those areas that can be completed easily, and to develop other areas later on. Publishing what is available may incite agencies to provide input on areas where less information is available.

4.   Initially, the Framework had been planned as a publication. However, publications often become quickly outdated. Therefore, the Task Force suggested issuing the framework on Internet as a living document that will be continuously developed in future. This would facilitate the updating and use of the framework as a tool to interconnect the various metadata elements. It would also permit its linkage with SDMX.

5.   Based on the assumption of a web-based product, the Task Force also suggested having a small, self-sufficient core text included in the framework, and to link it to other material for further reading. This additional material can be included either via a link to its original location or can be mirrored to the UNECE website.

---

[1] The membership of the task Force includes: Graeme Oakley and Alistair Hamilton (Australian Bureau of Statistics), Miroslava Brchanova (Czech Statistical Office), Søren Netterstrøm (Statistics Denmark), Bo Sundgren (Statistics Sweden), Max Booleman (Statistics Netherlands), Daniel Gillman (US Bureau of Labor Statistics), Juraj Riecan (UNECE), Jana Meliskova (expert invited by UNECE), Denis Ward (OECD), Marco Pellegrino (Eurostat).

STRUCTURE OF THE FRAMEWORK

6.   The agreed structure of the framework follows. More detailed notes on the structure of the framework and its expected contents are outlined in the Annex to this document.

A.   Corporate Context

i)   Executive overview
ii)   Corporate value proposition from metadata management
       - benefits for internal purposes
       - benefits for respondents
         We have in mind the Metadata Common Vocabulary definition of "respondents" here, which
         seems to include (maintainers and) providers of administrative by product, including register,
         based data
       - benefits for end users of statistical output
       - benefits for national statistical systems
       - benefits for inter-operation with other national and international agencies
iii)   Establishing core principles for metadata management
iv)   Metadata management strategies and policy frameworks
v)   Corporate governance models for metadata management

B.   Metadata concepts, standards, models and registries

i)   Terminology
ii)   Definitions, purposes and taxonomies
iii)   Metadata standards
iv)   Common metadata constructs
       - populations, units and object classes
       - classifications and classification schemes
       - variables and data elements
       - data sets
       - documentation of data sources (e.g. documentation of survey processes)
       - discovery metadata
       - time
       - space
       - quality measures
v)   Metadata models
vi)   Metadata registries and registration processes
vii)   Interchange of data and metadata

C.   Metadata and the statistical cycle

i)   Survey planning and design
ii)   Survey preparation
iii)   Data collection
iv)   Input processing
v)   Derivation, estimation, aggregation
vi)   Analysis
vii)   Dissemination
viii)   Post survey evaluation

D.   Implementation

i)   Implementation, planning and management
ii)   Identifying leading practice
iii)   Usability considerations
iv)   Infrastructure development options
        - build
        - buy
        - sharing and collaboration between NSOs
        - open source software
v)   Updating and improving statistical processes
vi)   Change management
        - Influencing corporate culture (includes communication plans)
        - Transition planning and management

DIVISION OF RESPONSIBILITIES IN PHASE 2

7.   The Task Force agreed to split the responsibilities for loading the above structure with the initial contents among its members in the following way:

Part A: Graeme Oakley (Australia), Juraj Riecan (UNECE), Miroslava Brchanova (Czech Republic), Jana Meliskova
Part B: Issues related to standardisation and terminology: Denis Ward (OECD) and Marco Pellegrino (Eurostat)
Issues related to models and objects: Paul Johanis (Canada) and Daniel Gillman (United States)
Part C: [to be covered at a later stage]
Part D: Max Booleman (Netherlands)

ANNEX

Draft annotated outline of the individual parts of the Common Metadata Framework

A.i) Executive overview: this would probably be written after the other elements are compiled. It provides an opportunity to introduce and encapsulate issues that will be covered in much more detail in Part B and Part C (e.g. basic definition of "metadata" and overview of metadata management in the context of the statistical cycle).

A.ii) Corporate value proposition from metadata management: it is proposed to address this key issue from a broad perspective and in a structured manner. It starts with "benefits for internal purposes" for pragmatic reason, addressing "what's in it for us?" before looking at other benefits which might be perceived as more "altruistic". There are many compelling subpoints that can be made under this item. "Benefits for national statistical systems" provides a good opportunity to include the perspectives of international statistical organizations, SDMX etc.

A.iii) Establishing core principles for metadata management: starting with high-level principles and working down to more practical details is probably a good approach for readers who are not already convinced of the need for metadata management, and who think it is a complex, difficult and technical field. (Those who are experts in metadata management can go directly to the elements of the framework document, which are of interest to them.) The Czech Republic and ABS have material that would be highly relevant in this area.

A.iv) Metadata management strategies and policy frameworks: no NSO has yet achieved perfection across all aspects of metadata management. The question of how an organization might advance from the current situation to an improved future situation is therefore very important.

A.v) Corporate governance models for metadata management: this was a key issue discussed since the beginning, and it is essential to organizations making real progress with their metadata strategies. We could emphasise "whole of organization" governance, and buy-in, reinforcing that metadata management is not an IT driven activity (although it may be IT enabled).

B.i) Terminology: this part will give a description of the importance of a well-defined, standard terminology. (The definition of a few specific terms such as "metadata" would be explored under the next heading.) The Task Force members have already had a discussion about "Which comes first, terminology or a metadata model?" The SDMX Metadata Common Vocabulary (MCV) should receive good coverage under this heading.

B.ii) Definitions, purposes and taxonomies: definition(s)(?) of "metadata" itself and key concepts around it. This includes taxonomies (e.g. ABS currently uses "definitional", "procedural", "operational", "system" and "dataset"). The concept of "active" vs. "passive" (but maybe using different words) is also important. Statistics Netherlands has a powerful concept of metadata either describing "states" or the process for moving between states. There are also various typologies of metadata "purposes"/"uses".

B.iii) Metadata standards: various metadata standards will be referenced throughout the document. At some point, however, it is necessary to talk about standards in general. Content might include broad and general coverage of:
- what types of standards exist;
- how do they get created and maintained (and how can NSOs influence the directions taken);
- how well do standards typically fit together;
- likely benefits from applying standards;
- implementation and compliance strategies;
- sources of further information

B.iv) Common metadata constructs: a concrete approach to this area is still to be developed. For example, the Metadata Common Vocabulary doesn't make a call between 11179-3 based "data element" constructs and Oslo Group (and others) "variable" constructs. There may finally be more entries or less entries on the list. The original intention was to focus on key objects that are relevant throughout the cycle. Concepts such as "frame" and "sample", for example, might be better documented under Cii) (Survey preparation).

"Time" is generally treated as a special consideration (e.g. in the work by Professor Sundgren and within the ABS). It may be moved up to be the first on the list of third level headings (or even a second level heading in its own right).

"Space" (geography) tends to be the second most "special" consideration in the ABS. There are also very well developed standards related to spatial metadata. While focusing on statistical metadata in general, there should be a cross reference to metadata for spatial analysis.

"Quality Measures" could also be a second level heading in its own right. Quality, in more general terms, will also be a theme for Part A.

B.v) Updating and improving statistical processes: to cover a discussion of roles, and application, of metadata models, as well as referencing examples. It would be good to include very extensive models (e.g. the Metanet Reference Model) as well as models that are "narrower but deep" in their application (e.g. the 11179 metamodel). It is not clear whether specific models for specific purposes would be recommended.

B.vi) Change management: there will be a need to describe the concept of metadata registries and how registration adds value. 11179-6 provides a paradigm - but not necessarily the only one. This element could include an indicative list of metadata objects that may warrant registration. It could also illustrate how registries can support both "collection" and "use" of metadata throughout the statistical cycle.

B.vii) Interchange of data and metadata: an important consideration. The absence of any "universal" metadata registry means that some level of interchange will be required for the foreseeable future. SDMX has models and standards related to this, but some members of the Task Force questioned whether these standards cover all statistical needs.

C. Metadata and the statistical cycle: the Task Force agreed to talk about the "Statistical Cycle" rather than "Statistical Activities" in order to emphasise that C8 (Post survey evaluation) feeds into C1 (Survey planning and design). The Task Force members would also be comfortable, however, with "Statistical Activities", in order to emphasise the relevance of metadata for activities other than a typical survey cycle.

It is possible to break the cycle down to more categories or fewer categories. A very simple three-part model of "collection", "processing" and "output" has been used for a number of purposes. The proposed breakdown of the statistical cycle in part C is similar (but not identical) to the division used in the paper on metadata systems in the statistical production process prepared by Professor Sundgren for the 2004 METIS meeting. This division is also similar (but not identical) to the division used within the ABS.

It is typically during Ci)(Survey planning and design) that surveys have an opportunity to fundamentally re-engineer and improve their metadata management practices. Ideally, surveys should, at least at a high level, be able to identify most of the metadata they will need during subsequent steps. This includes reuse of metadata (where it is fit for purpose) that has been defined previously by this survey or by others.

D. Implementation: the Task Force members suggested focus also on the implementation issues, and this became part D. The material for part D should be drawn on practical experiences of national

statistical offices recently having implemented or re-engineered their statistical metainformation systems. Therefore, the initial contents, which will be further developed and generalised, will be drawn on the experience of Statistics Denmark.

Annexes: The material after part D of the Framework would include:
- an index;
  - a glossary of key terms [hopefully drawing its content from a standard source such as the Metadata Common Vocabulary];
- a bibliography;
- a list of contacts;
- Frequently Asked Questions (FAQs);
- illustrative examples on various issues;
- other appendices may be added.

* * * * *