INF. 3
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)
(Luxembourg, 9-11 April 2008)

SUMMARY OF ABSTRACTS FOR WORK SESSION PAPERS

Note prepared by UNECE Secretariat

TOPIC 2(i): ADVOCATING FOR STATISTICAL METADATA IN A CORPORATE CONTEXT (CMF PART A)

Strategy for the Development of a New Integrated Metadata Driven Statistical Information System in the Statistical Service of Cyprus (Costas K. Diamantides, Cyprus)

The Statistical Service of Cyprus (CYSTAT) is facing increased number of users on both national and international level. Furthermore, great challenges for CYSTAT are changing users' needs and calls for better quality of statistical information. At the same time, however, the CYSTAT is requested to increase efficiency of statistical production and to reduce burden on statistical respondents. Such development has a significant impact on the scope, contents and quality of observed statistical information and, consequently, on the methods, tools and techniques used for collection, processing and dissemination of statistical information.

In accordance with the CYSTAT strategic goals, a redesign of existing information system was launched after the accession process to the EU had been successfully finished. A global architecture of an Integrated Statistical Information System in CYSTAT (CYSTAT-ISIS) was designed.

An integral part of CYSTAT-ISIS is the Statistical Metainformation System (SMS). Effectiveness of CYSTAT-ISIS depends directly on effectiveness of its SMS. The goal of SMS is to ensure internal and external integrity of CYSTAT-ISIS from both statistical content and technology point of view. Making a business case for CYSTAT-ISIS and SMS management was indispensable precondition for SMS development and implementation.

Managing metadata in Statistics Norway (Anne Gro Hustoft and Jenny Linnerud, Norway)

Statistics Norway's metadata strategy was approved in 2005. Two of the recommendations in the strategy designed to support advocacy of metadata systems, were to document definitions of key metadata concepts, and to connect variables, classifications and file descriptions within a metadata portal. Development of the portal began late 2005 and user testing was completed in 2007. The portal displays our key metadata concepts and the contents of our master metadata systems making them more accessible and easy to use both for researchers and external metadata experts, and for internal users. Metadata managers use the portal to follow up the coverage and quality of the contents of the underlying metadata systems. The design is flexible so that the contents of other metadata systems, e.g. a questionnaire server, can be added as these become available.

Value of metainformation system for the Czech Statistical Office (Ebbo Petrokovits, Czech Republic)

The Czech Statistical Office (CZSO) has launched in 2005 a substantial redesign of its Statistical Information System (SIS). Major goals of the project are: to increase efficiency of statistical production, to improve quality of statistical information and to reduce respondents' burden.

The Redesign of SIS is based on the CZSO strategic goals. It is ranked as a first priority task for the whole office. The project development and implementation is directly supervised and managed by the top management of the CZSO.

The model of SIS is focused on statistical production, incorporating all phases of the production and dissemination process of statistical information (PDSI). Necessity to identify PDSI and its phases called for the unification of statistical processes. The major aim of this exercise was to strengthen the organization and management of statistical work.

Statistical Metainformation System (SMS) is an important basic component of a redesigned SIS. SMS tools shall ensure integration of SIS inside and outside the statistical office. Furthermore, SMS shall contribute to evaluate efficiency of statistical processes and quality of statistical data. Currently, the first part of SMS project (classifications, variables and tasks) is in the phase of a pilot testing.

Management and organization of the work, applied for the SMS development, resulted in a significantly increased cooperation and interest of subject-matter statisticians and methodologists. Lessons, learned from the project so far, clearly justified a direct involvement of top management as a necessary precondition to make SMS a success story.

The paper is presented in the following parts:

1. **Redesign of SIS** - Purposes for SIS redesign, its major goals and global architecture.
2. **Unification of statistical processes** - Standardization of statistical processes in CZSO, focus on the life cycle of statistical production.
3. **SMS goals and architecture** - SMS functions and their interpretation in the SMS model and architecture. SMS plans.
4. **SMS management and organization** - Management strategy and organization in the stages of the SMS development, implementation and operational running.
5. **Lessons learned** - Experiences from the SMS development and implementation. Important findings and lessons learned
6. **Conclusions**

TOPIC 2(ii) METADATA CONCEPTS, STANDARDS, MODELS AND REGISTRIES (PART B)

Classifications of statistical metadata (Bo Sundgren, Statistics Sweden)

Statistical metadata may be classified in a number of different dimensions, more or less orthogonal to each other. For example, statistical metadata may be classified according to

- who needs the metadata, and for which purposes (e.g. exploratory and explanatory)
- what the metadata inform about: metadata objects (attachment objects) and metadata variables
- how the metadata are structured and formalised (or not)
- where the metadata come from (source processes) and where they go (use processes)

The paper defines different classifications of statistical metadata, and it also discusses why and how it is useful to consider these classifications when statistical agencies are developing, operating, and evaluating statistical systems.

Registry facilities for supporting the exchange of statistical data and metadata (Beng-Ake Lindblad, Marco Pellegrino and Francesco Rizzo, Eurostat)

Eurostat is building an architecture for data and metadata exchange which is largely based on its SDMX registry, containing and providing over the Internet information needed to facilitate the reporting, collection, and dissemination of statistics. The paper describes the distinct registry modules and its purposes, also providing a highlight of the on-going work aimed at enabling end-users and metadata producers to access, analyse and reuse statistical metadata. The approach is based on the standardisation of concepts used, on the use SDMX protocols for the interoperability of messages, and on the availability of metadata tools and formats compliant with SDMX version 2.0 standards.

Structural and reference metadata in the European Statistical System (August Götzfried, Eurostat)

The paper describes the current production and use of structural and reference metadata at Eurostat and within the European Statistical System. The international SDMX standards - and the newly released "content-oriented guidelines" for the exchange and sharing of data and metadata - provide the opportunity for standardising structural and reference metadata between international organisations and within the European Statistical System. This should lead to a considerable improvement of structural metadata (data structures, code lists) and reference metadata (content metadata). In 2007, Eurostat launched a new standard for reference metadata called "Euro-SDMX Metadata Standard" (ESMS) which takes into account a series of quality assessment aspects. The new standard is linked to the SDMX standards for data and metadata exchange and is being tested for a full-scale implementation and will be recommended for exchange across the ESS.

General requirements for the soundness of metadata models (Tjalling Gelsema, Netherlands)

One of the challenges Statistics Netherlands is facing in the organization of its metadata repositories is maintaining coherence in vast amounts of variables, classification systems, object type descriptions, aggregate table descriptions, etc. In order to maximize the sharing of data and metadata, typical problems that need solving are avoiding duplicates in metadata repositories and identifying the basic metadata constituents of tables, especially if they are transformed (e.g., aggregated). These constituents should allow understanding of how tables relate.

This paper advocates general requirements for metadata models that are stronger than the ones usually employed, but have shown their value in computer science (and in formal semantics in particular). In short, a metadata model should not allow instances that are outside the domain of representation (no junk) and a model should not allow an instance to have more than one representation in the model (no confusion). The paper translates the ‘no junk, no confusion’ doctrine to the statistical metadata domain and, using algebraic techniques, presents a new model for classification systems as an example.

The paper will also give more general future directives on how to achieve models for statistical metadata that satisfy these requirements.

Further Developments in the Terminological Principles for Data (United States - Frank Farance (Farance Inc) & Daniel W. Gillman (Bureau of Labor Statistics))

The theory and practice of terminological methods can be used for a better understanding of the meaning of data (data semantics), and for better data exchange among statistical agencies and with users (data interoperability) including exchanging the meaning of data consistently (semantic interoperability). The key strategy is to understand a "datum" as a kind of "designation" - see the Nature of Data, WP 12 in METIS 2006. In terminology, a designation is the relationship between a concept and a sign, such as terms (whose concepts refer to more than one object) and appellations or names (whose concepts refer to exactly one object). The "datum" is differentiated from the "designation" because a datum is a designation whose concept includes a notion of equality, an essential feature of data.

By applying these terminological methods, data and the vessels that hold them (data elements and structures of

data elements) can be better understood by the precise degree of semantics intended and the interoperability afforded. Data semantics follow the basic terminology framework consisting of concepts, their characteristics, and the properties associated with those characteristics. Semantic interoperability depends on having a shared semantic framework and rules for using that framework. These issues are described in the paper.

In addition, the representation and datatype associated with a datum are needed for data interoperability. This is discussed in detail. For example, we will show how the kinds of statistical data (nominal, ordinal, interval, and ratio) are really constituents of datatypes. In addition, units of measure, often associated with quantitative data, are seen as a bridge between datatypes and semantics.

TOPIC 2(iii) METADATA AND THE STATISTICAL CYCLE (PART C); AND IMPLEMENTATION (PART D)

Metadata Management at the FSO - Introducing an End-to-End Approach to Sustain Modernisation Projects (Fabian Perrot, Switzerland)

The FSO is currently reengineering data production workflows with the definition of integrated statistical production systems. Those modernisation projects will bring a change in the way data are collected and processed at the FSO. It will mark the transition from a business model where each unit collects and processes its own data (aka stove pipe) to a model where data is collected once and reused in different statistical products. The use of administrative sources will also be favoured in order to reduce the burden imposed on respondents and increase the overall quality of source data. In order to achieve these goals, a new metadata management strategy has to be defined and introduced in the whole organisation.

Previous developments in centralised metadata management at the FSO did not envision the end-to-end approach and consequently efforts were focused almost exclusively on the analysis and dissemination phases of data processing. This stance proved to be difficult to maintain in the long term due to issues with latency problems in workflows. Also, the metadata system could not provide internal modernisation projects with information about source data and collection methods.

The presentation will focus on organisational issues linked with the introduction of a centralised metadata system, this time affecting the whole statistical production cycle in the FSO. Emphasis will be given to:

- How modernisation projects will work as enabler of harmonisation, bringing the need to harmonise, but also the incentive to do so in production units;
- How centralisation of metadata will gradually take place, as decentralised metadata production systems are aggregated in the central metadata system;
- How metadata management and harmonisation will be dealt with in the organisation, with the splitting of tasks between central, intermediate and decentralised actors, incorporating all metadata stakeholders in a matrix organisation.

Basic information about change in metadata modelling and architecture paradigm will also be briefly mentioned for potentially interested parties.

Experiences from distributed registering of metadata in MetaPlus (Klas Blomqvist, Sweden)

MetaPlus is a tool primarily developed for documenting micro data. The data model is compliant with ISO 11179, which makes it possible to develop the system further to handle all metadata aspects of the statistical life cycle. The paper focuses on the experiences so far from working with MetaPlus, in particular registering metadata. Since the subject matter departments are responsible for registering the contents of the documentations in MetaPlus the paper stresses the need for building a support organisation with the KMI (classifications, metadata and content harmonisation) group at the Process department and the documentation network. Connections to other Metadata systems, such as the archiving functionality and future plans are also mentioned.

Metadata Requirements for Archiving Structured Data (Alice Born, Canada)

The final step of data stewardship in a national statistical office is archiving its structured datafiles. In Canada, the Library and Archives of Canada Act along with the federal government's Policy on Information Management requires that government institutions fully protect and safeguard for future generations those government records deemed by the National Archivist as warranting preservation.

In the mid-1980s, Statistics Canada initiated an archival system for its final master datafiles, which has subsequently been incorporated into the Agency's corporate metadata repository, the Integrated Metadatabase (IMDB). The IMDB contains descriptive information concerning the location, format and content of the master datafiles for all of Statistics Canada's surveys as well as a set of standardized metadata related to the survey from which the data were derived. The Agency is currently revaluating the metadata requirements for archiving all structured data.

The paper presents a proposal for the "archive" process as part of the statistical life cycle; what the sub-processes are; and where they fit in the broader Common Metadata Framework. It also describes how data archiving fits into the IMDB metadata model, and the role of the IMDB in documenting and managing the metadata for these archived datafiles. This phase in the development of the IMDB will bring it closer to end-to-end support of the statistical life cycle at Statistics Canada.

Proposal for a Generic Statistical Business Process Model (Steven Vale, UNECE)

Part C of the Common Metadata Framework (CMF) – Metadata and the Statistical Cycle – refers to the phases of the statistical cycle (ie the 'statistical business process') and provides generic terms to describe them. The intention is for statistical organizations to agree on standard terminology to aid discussions on developing statistical metadata systems. In 2006, when the structure for Part C was first being developed, these phases were proposed to be: (1) survey planning and design; (2) survey preparation; (3) data collection; (4) input processing; (5) derivation, estimation, aggregation; (6) analysis; (7) dissemination; and (8) post survey evaluation.

During the workshop on Part C of the CMF, held in July 2007, several national statistical offices reported on the terms currently used to describe phases of the statistical cycle within their organization. After some discussion, it was considered that with the addition of 'Archive' and 'Evaluate' phases, the model currently used by Statistics New Zealand could provide a better basis for the generic CMF model.

Using the work of Statistics New Zealand as a basis, the METIS Steering Group has developed a proposal for a Generic Statistical Business Process Model. This work will be presented at METIS 2008 for further discussion.

This session will also include:

- Demonstrations of metadata systems by Canada, European Central Bank, Ireland and Sweden.
- Presentation of case studies from the national statistical offices of Norway, Croatia, Portugal, Australia and the United Nations Industrial Development Organization (UNIDO).
- Optional lunchtime demonstrations of metadata systems by Croatia and UNIDO.