E

UNITED
NATIONS

Economic and Social Council

Distr.
GENERAL

ECE/CES/2007/METIS
6 July 2007

ENGLISH ONLY

ECONOMIC COMMISSION FOR EUROPE STATISTICAL COMMISSION

CONFERENCE OF EUROPEAN STATISTICIANS

Vienna, 4-6 July 2007

REPORT OF THE WORKSHOP ON STATISTICAL METADATA (METIS)

Note by the Secretariat

1. The UNECE workshop on statistical metadata was held in Vienna, Austria, from 4 to 6 July 2007. It was attended by participants from: Austria, Belgium, Bulgaria, Canada, Croatia, Cyprus, Denmark, Estonia, Germany, Ireland, Lithuania, Mongolia, New Zealand, Norway, Portugal, Romania, Russian Federation, Slovakia, Slovenia, Spain, Sweden, Switzerland, and the former Yugoslav Republic of Macedonia. Representatives of the Statistical Office of European Communities (Eurostat), International Monetary Fund (IMF), United Nations Educational, Scientific and Cultural Organization (UNESCO) Institute for Statistics, and the United Nations Industrial Development Organization (UNIDO) also attended. Participants from the Catholic University of Leuven (Belgium) attended the workshop at the invitation of the Secretariat.

2. The agenda consisted of the following sessions focused on Part C of the Common Metadata Framework: Metadata and the Statistical Cycle.

(a) Setting the Context for Part C
(b) Sharing Experiences
- Models of the statistical cycle
- Functions and attributes of statistical metadata
- Organizational, managerial and cultural aspects of metainformation systems

(c) Case Studies

3. The workshop was co-chaired by Ms. Joza Klep (Slovenia), Mr. Matjaz Jug (New Zealand), and Ms. Alice Born (Canada), who also shared the role of session organizers

4. Dr. Peter Hackl (Austria) opened the meeting, welcoming delegates to Statistics Austria and highlighting the need for a Common Metadata Framework to guide statistical offices in their development and management of metainformation systems. Workshops such as this one are a valuable opportunity for sharing experiences and holding detailed discussions in order to progress this important work.

RECOMMENDATIONS FOR FUTURE WORK

5. The participants discussed the future work on the Common Metadata Framework (CMF), as well as the broader activities of the UNECE/Eurostat/OECD on statistical metadata (METIS). They made the following recommendations:

(i) Part C of the CMF should be further developed, in particular: revise the template for case studies and solicit case studies from other countries (See Annex, paragraph 23).

(ii) The Steering Group on Statistical Metadata (METIS) should consider issues raised in the Annex, Section IV, and make relevant proposals for discussion at the 9-11 April 2008 UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS 2008).

(iii) The participants made the following suggestions for items to be included on the agenda of METIS 2008:

- Countries would appreciate if Sweden and Canada present their projects on archiving
- Cyprus offered to present their experience in preparing a strategy for the development of a new integrated metadata driven statistical information system
- Presentation of any new case studies that have been prepared
- UNIDO volunteered to present a case study using the template – a premiere from the international perspective.

(iv) Part B of the CMF should provide more information about application of standards – which standard for what purpose.

(v) Part D of the CMF should include the case studies developed for Part C, a summary of lessons learned (to be collated from all case studies and categorised), success stories, best/good practices. More contributions to this should be solicited.

FURTHER INFORMATION

6. The conclusions reached during the discussion of the substantive items of the agenda are contained in the Annex. All background documents and presentations for the meeting are available on the website of the UNECE Statistical Division:

http://www.unece.org/stats/documents/2007.07.metis.htm.

7. The participants thanked Statistics Austria for providing excellent facilities for the workshop and for their hospitality.

ADOPTION OF THE REPORT

8. The participants adopted the present report before the workshop adjourned.

ANNEX

SUMMARY OF THE MAIN CONCLUSIONS REACHED DURING THE WORKSHOP ON STATISTICAL METADATA (METIS)

I. SETTING THE CONTEXT

Session Organizer: UNECE

2. Following the METIS work session in April 2006, the Common Metadata Framework (CMF) editorial board developed a proposal for structure of Part C: Metadata and the Statistical Cycle. This proposal was revised to include comments received during a review phase in mid 2006 and was published on the CMF website (http://www.unece.org/stats/cmf/PartC.html).

3. The following issues were raised during the discussion on this topic:
   (a) The definition and application of the term survey. Some statistical agencies consider a survey as those activities using data collected from respondents, thus excluding administrative sources. Other statistical agencies consider it as any activity that collects or acquires statistical data. Included are censuses, sample surveys, the collection of data from administrative records, and derived statistical collections. Although each office will continue to use the terminology appropriate to their environment, it would be useful for METIS participants to agree on the use of the term so there is a common understanding within the group.
   (b) Further development of a generic high-level process model of statistical cycle phases. Any generic model shall be considered only as a general schema that may not be fully applicable to all cases. The previously proposed model (C1-C8 below) was presented for review. Its further development was discussed during the 'Finalizing the Framework' session at the end of the workshop:
      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.
   (c) There were several, slightly different, terms used for the cycle, one of them being **'statistical survey life-cycle'.** The appropriateness of this term was discussed further on the final day of the workshop.

4. This session also tabled the issue of classification of statistical metadata. Part C should include a common classification for the different types of metadata (to be developed). For example, some offices refer to structural and reference metadata, others classify into categories such as definitional metadata, conceptual metadata, etc. Historically, statistical metadata has focused on particular phases on the production process, mainly dissemination, or in some cases, survey design. This scope has expanded to include process, operational, archiving and knowledge management metadata. A classification will help metadata managers in developing comprehensive metainformation systems, and to communicate with each other about the issues associated with each type of metadata.

5. Participants noted the review of statistical metadata (METIS) work by the Bureau of the Conference of European Statisticians. The Bureau has suggested the establishment of a Steering Group on Statistical Metadata (METIS) to continue the work on the CMF. They also requested that future work includes activities focusing on:
   (a) advocacy of statistical metadata; and
   (b) technical assistance.

6. These matters will be considered by the Steering Group and raised for further discussion at METIS 2008.

7. One motivation for development of the CMF is the existence of numerous statistical metadata standards and

the resulting demand for clarification by statistical offices. The following points were made in the discussion:

(a) Individual standards were created for particular purposes and are often complementary rather than contradictory.

(b) The CMF should explain situations in which certain standards can be used. The CMF also attempts to provide links between individual standards, for example between SDMX registries and ISO 11179, etc.

8. The UNECE presented results of a survey related to the content and structure of Part C of the CMF (WP.9). This survey was sent to METIS participants in early June 2007 list and 20 responses were received (approx. 34% of organizations involved in METIS).

(a) The results showed that 95% of respondents had consulted the CMF website (www.unece.org/stats/cmf) and had no suggestions for changes to the proposed structure of Part C.

(b) Respondents were asked to describe their preferred process model and map it to the existing generic model of C1-C8 (outlined at 3(b)). The results highlighted that while a generic model is useful, statistical agencies will adapt it to suit their own environment. Further discussion of the process model is summarized at paragraph 22 below.

(c) Respondents were asked to comment on their usage of existing standards. This revealed that the four most implemented standards were: Dublin Core; DDI; SDMX; and ISO 11179. A review of standards usage was featured in a survey conducted by MetaNet in 2003. The results of this survey are attached at Appendix 1.

(d) The survey also identified common organizational and cultural issues facing metadata experts.

(e) The participants expressed their satisfaction with the survey, and suggested to repeat it in the lead up to METIS 2008. Suggested enhancements to the survey include:

i. List the different standards by category, according to their function;

ii. Actively seek responses from more offices, particularly those known to be implementing metainformation systems.

9. Participants expressed their interest in the Metadata Common Vocabulary (MCV). The representative of Eurostat provided an update on the status and future plans. The discussion highlighted the following issues:

(a) The MCV project is currently a part of the SDMX standards – Content Oriented Guidelines, but the work commenced well before the SDMX initiative was created.

(b) The MCV should be useful in different frameworks. It is a complex product composed of several sub-glossaries, and it exceeds the scope of SDMX. There are plans for improving its dissemination, one of them being the use of a XML (SDMX-ML) version of the MCV.

(c) The sponsors of the project would like to increase the participation of experts from member countries. New governance rules aimed at increased participation will be announced later this year. A facility through which interested countries can contribute will be established.

(d) Translation into national languages of member countries represents a challenge for the MCV project. For the time being it exists in English.

II. SHARING EXPERIENCES

Session Organizer: Joza Klep (Slovenia)
Documentation: Contributed papers by Canada, Austria, Catholic University of Leuven, Sweden, and the International Monetary Fund (IMF).

10. The presentation of contributed papers focused on the following issues:

(a) Corporate metainformation systems are being implemented to replace the traditional stovepipe approach, and to cover the whole cycle. This introduces a need to collect process and operational metadata, a detailed definition for which is still needed (refer to paragraph 4 above).

(b) There is some agreement on what the statistical cycle is. However, when it comes to details, the models differ between statistical organizations, as well as across individual surveys (refer to paragraph 3(b)).

(c) The roles and responsibilities of subject matter statisticians, metadata managers, senior managers and other participants in the statistical cycle were raised. Issues relating to change management,

metadata quality, centralised versus decentralised metainformation systems, legal aspects, partnership and cooperation, outsourcing versus in-house development, and training and knowledge management were also considered.

11. The following issues were raised during the discussion on this topic:
    (a) Subject-matter specialists have to be involved in the early stages of development of metadata related projects. This can be achieved through introducing an appropriate governance structure. A concept of stewardship[1] was mentioned, which aims to devolve ownership of the metadata project throughout the organization.
    (b) Clear corporate guidelines and business rules should be established to ensure quality and consistency of metadata. Compliance with these rules may be monitored through internal and external audits.
    (c) Some of the international, but also national, statistical agencies attempt to introduce a standard subject matter classification base emanating from SDMX. This may improve understanding of data availability through different channels and decrease burden on respondents.
    (d) Reuse of metadata is a key method to achieve greater efficiency, consistency, and synergies across the statistical office. The possibility of using "wiki[2]" technologies was mentioned in this respect.
    (e) Metadata is needed for archiving of data files and this is emerging as an important issue for national statistical offices. Archiving and archival metadata is currently missing from the CMF model of the statistical cycle and will need to be incorporated. Based on the interest expressed by participants, Canada is prepared to present results of its current project related to metadata for archived files at METIS 2008.
    (f) Differences between data elements used at the design versus the dissemination phase were questioned. There were expectations that they would be the same, but some differences are apparent. Further research is needed in this area.
    (g) Administrative records were not originally collected for statistical purposes. Therefore, metadata on collection and processing methodologies, and quality of data from administrative sources is vital.
    (h) Several participants stressed that versioning of metadata brings new issues to be resolved.
    (i) As raised during discussion about the MCV, the translation of metadata can be problematic, particularly for international comparability. The experience of countries is diverse. Some countries are relatively inexperienced or under resourced, and others have a long tradition of translating their national statistics into English and other languages.
    (j) The granularity of quality-related metadata was discussed. As an example, many offices evaluate the overall response rate, but do not register it with every data item or data file.
    (k) Management of metadata related projects is also of interest. The following two aspects were highlighted:
        - The IT departments often lead metadata related projects, because they are positioned to understand the corporate needs of the office. It is important to establish communication and participation of subject matter departments. The perception of ownership is also important.
        - Advocacy is needed to convince management to assume start-up costs. In long run, well established metainformation systems bring costs benefits, but it may not be always easy to prove this in the initial phase of the project.

12. The following suggestions for Part C of the Common Metadata Framework were made:
    (a) From the viewpoint of national statistical offices, the ongoing international work on the CMF, SDMX and other initiatives towards harmonized standards, glossaries, etc. is very important in ensuring comparable international statistics.

---

1 See contributions by the IMF and Canada. In IMF stewards will be responsible for the content of all dimensions in the warehouse (IMF.Stat). They include countries, units, economic concepts and data source, status, time and frequency. In the Canadian case, followed also by Austria, the concept of stewardship is used for subject-matter areas.

2 Web based technology that allows users to update the content of pages quickly and easily within their internet browser. Wikipedia (www.wikipedia.org) is an example of public domain wiki, which is an online encyclopedia in many different languages, the content of which has been established and maintained by the general public.

(b) A list of success stories, with references, would be useful.

III. CASE STUDIES

Session Organizer: Matjaz Jug (New Zealand)
Documentation: Case studies by New Zealand, Canada, Sweden, Ireland, and South Africa.

13. In accordance with the recommended content for Part C of the Common Metadata Framework (CMF), countries were invited to present case studies on use of metadata through the statistical survey life cycle. The case studies raised the following issues:
    (a) Organization and governance
    (b) Project management programme/project approach
    (c) Life cycle models – there are several similar models with some differences
    (d) Metadata groups (general, statistical model based groups) – Should there be an agreement on common groupings?
    (e) Metadata phase by phase
    (f) IT architecture technical issues
    (g) Standards, models, file formats
    (h) Sourcing approaches (in-house, external/outsourced, learning and training)
    (i) Roll-out approaches (big-bang vs. step by step implementation)
    (j) Organizational and cultural issues

14. In addition to the discussion on case studies themselves, the workshop participants agreed on amendments to the template and planned for the next round of collection of case studies. Case Studies have provided information relevant not only for Part C but also for other parts of the CMF. Participants agreed that having complete information covering all parts is good so Case Studies should continue to provide similar scope in the future.

15. The participants suggested following improvements to the case study template:
    (a) Case studies were found very useful. English speaking statistical organizations have a relative advantage in preparing them with existing materials. The overall structure was satisfactory, and may be improved taking into account suggestions outlined below. Structuring the information according to a template is useful, in particular to better enable comparability between organizations.
    (b) More precise instructions would be useful. They would help to have a common understanding of the template.
    (c) Lessons learned should be moved as a separate section at the end of the template.
    (d) Part 3 of the template should take into account that not all offices currently undertake end-to-end projects.
    (e) Other interesting information would be whether the office has a “big bang” or stepwise approach to implementation.
    (f) It should be indicated whether the project phase/component is planned, or has been implemented.
    (g) In some offices there are other metadata uses, for example for management and budgeting. These should be added to the template.
    (h) Emphasis should be given to system and design issues.
    (i) The costs and benefits issues would be of interest.
    (j) The template may be clearer if using a table format.

16. There were several models of the statistical survey life cycle presented within the case studies. In general these models are quite similar with only some minor differences. Therefore, there is a chance to reach an agreement on a high-level generic model of the statistical survey cycle. It was suggested to add archiving as a new phase to the model of the cycle.

17. The following issues were raised during the discussion of the case studies:
    (a) Metadata projects require a multitude of skills combining the expertise in business analysis, in various statistical areas, statistical methodology and IT.
    (b) Some offices are considering SDMX’s data model for metadata at the macro level. Some other

offices are using the PC-Axis data model. These models are seen as a way for harmonization of metadata. There are numerous issues for further investigation with respect to the position of SDMX within the statistical information system architecture. If the SDMX standard is considered as a structural standard for metadata throughout the cycle, it would not be perceived as an additional exchange format and the data and metadata transformations would be much simpler.

(c) Generic data models provide a useful base for harmonizing data processing, within the current efforts of migrating from a stovepipe approach. However, they are not ready for direct linking with applications that require specific data structures. Service Oriented Architecture (SOA) can provide a solution. A service layer (message and data bus) provides interface between loosely coupled components (data warehouse, process management, business rules, applications, etc.) of the information system.

(d) An alternative to loosely coupled applications is a gradual migration to an Enterprise Architecture based on generalized systems. This is linked with creating a corporate culture through providing metadata up (and down) the value chain, and making the metadata modules obligatory. For example, in some offices survey managers cannot release data before completing the metadata[3].

(e) What to keep in the central metadata repository and what to keep in decentralised systems? The answer may differ in relation to the level of centralisation/decentralisation of the office itself.

(f) Numerous legacy applications with large volumes of data and metadata make it difficult, if not impossible, using a "big-bang" approach. Implementing new metadata systems with stepwise linking to legacy applications is a possible solution.

(g) Versioning is an important issue as definitions, concepts, and structures of surveys change over time (called a "time travel" in some offices). Business rules should define which objects do or don not have version control. For example, a survey object should have version control reflecting changes that affect populations, such as movement of administrative boundaries, etc. On the other hand, the survey instance object does not need version control – a survey instance can be considered as a version on its own.

(h) Statistical offices are at different stages in developing user-friendly interfaces for the electronic metadata capturing tools. Some legacy applications still require a specially trained staff for entering metadata.

(i) Survey managers should have ownership for the respective metadata. It is important that survey managers have a good understanding of metadata systems and principles (e.g. why and how variables are defined).

(j) Wiki technology is seen as a cost-effective decentralised information sharing tool. However, an example showed that it can serve well as an integration tool[4]. Wiki may facilitate better organization and discussion by replacing numerous e-mails and tracking input through editing histories. Work on definitions, business rules, processes, etc, can be carried on efficiently and transparently using a wiki.

(k) Costs are usually estimated and at the beginning and throughout the various phases of projects. How to estimate the value of benefits? The types of benefits identified during discussion include improved quality, increased knowledge and better mobility of staff, harmonisation, and avoiding duplication. However, it is difficult to value the benefits of metadata systems in monetary terms.

(l) Reuse of metadata was emphasised as a requirement throughout the workshop. It may be valuable for international organizations to reuse metadata already available in national metadata repositories. This goal would require some technical issues to be addressed. Firstly, some countries do not yet have well integrated metadata repositories. Other issues are the metadata structures, interchange formats, etc. Non technical issues include the problem of languages. National repositories primarily focus on metadata in the national language. Translation of metadata represents an additional burden for countries that do not routinely disseminate statistics in English.

18. The following suggestions for Part C of the Common Metadata Framework (CMF) were made:

(a) Update the structure for case studies on the basis of the discussion. Update the existing case

[3] For example the internal business in Canada rules do not allow survey managers to release data without appropriate metadata. The Statistics Act in Sweden requires that the information on the purpose, data quality and some other metainformation are available publicly (on the Internet) for each survey.

[4] Example: STC wiki of Statistics Canada is being used as an internal communication tool through which definitions and other metadata issues are discussed.

studies and distribute the new template to all countries.

(b) Norway and Croatia offered to submit their case studies for inclusion in the CMF. Other countries are invited to submit their case studies. Provide more time for completion of case studies in the future.

(c) As the CMF is a living product, the aim is for case studies to be regularly revised (eg annually) and updated.

(d) The CMF should provide a platform for agreeing on a metadata classification (what kinds of statistical metadata are used) and related terminology.

(e) More information on which standard is useful for what purposes would be valuable for national statistical offices. This should be further developed within Part B of the CMF.

IV. FINALIZING THE FRAMEWORK

Session Organizer: Alice Born (Canada)

19. The definition of a survey was discussed as there were diverse understandings of this term. It was proposed to adopt a common definition for use in METIS materials and meetings, which combine the original MCV definition with one used by Statistics Canada:

> *"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. Included are primary sources, such as censuses and sample surveys, and other sources, such as the collection of data from administrative records. Derived statistical activities are also included."*

This proposal was adopted by the participants at the workshop and will be forwarded to the METIS Steering Group for review.

20. There are several classifications for types of metadata (metadata groups) currently in existence. The workshop participants proposed that after further work to clearly define the terms, the following classification should be presented to the METIS Steering Group for consideration for inclusion in the CMF:

(i) Survey Metadata
(ii) Definitional Metadata
(iii) Methodological Metadata
(iv) System Metadata
(v) Operational Metadata
(vi) Quality metadata

21. Additional proposals included:

(a) To collect synonyms for types of metadata from the METIS participants, and to add explanatory notes on what is the scope of each category before the next round of the round of discussions at METIS 2008.

(b) CMF should try to reconcile various classifications (SDMX, MetaNet, etc.). This would help to help in the present proliferation of metadata classifications.

22. Feedback was collected from the participants during the workshop on the phases of the process model (as outlined in 3 (b) above). The following points were made by the participants:

(a) Need for more detailed explanation / sub-processes
(b) Support for the addition of a phase for 'archiving'.
(c) Include 'Information enrichment' (a step before dissemination) – It was later clarified that this is a part of the 'Analysis' phase.
(d) Evaluation should be a natural part of C2-C7 but it is often a final phase (feedback loops).
(e) Is use of the term 'survey' appropriate everywhere, or only C1, C2 and C8?
(f) Cross tabulate C1-C8 with censuses, sample surveys, collection of data from administrative sources and derived statistical activities, and what do you get? What is applicable and not applicable?
(g) A process concerning building the IT system could be useful (like the New Zealand case).

(h) Data reporting should be included as a special case of data collection.
(i) Processes that have to be communicated to the staff should be captured in the model. This implies that building systems is a process that should be communicated to the IT staff.

23. Considering the feedback from participants, it was proposed that the C1-C8 model be replaced with the following generic process model:
    1. Need
    2. Develop and design
    3. Build
    4. Collect
    5. Process
    6. Analyse
    7. Disseminate
    8. Archive
    9. Evaluate

24. It was noted that some surveys will not use all phases of the generic model and this needs to be made clear in the presentation of the model. For example, there will be a difference in the phases used for a survey instance, compared to those needed for establishing a new survey. To assist, Slovenia is willing to provide clarification on how this model could apply to survey instances. For further clarification, New Zealand will provide details of the sub-processes they use under stages 1-7 of this model. A final proposal for the process model will be submitted to the METIS Steering Group and discussed at METIS 2008.

25. There are several possibilities for naming this process model. The feedback received through a questionnaire distributed at the workshop showed that 20 participants agreed with the suggestion it be called **"Statistical survey life cycle"**. This title will be used in Part C of the CMF to cover the nine processing phases spelled out above.

26. Any further suggestions and contributions should be addressed to the UNECE Secretariat.

V. RECOMMENDATIONS FOR FUTURE WORK

27. The following actions were recommended for the future work on Part C of the Common Metadata Framework (CMF):
    (a) Revise and publish template for case studies on the CMF website
    (b) Update existing case studies by September 2007, providing further details in the areas of expertise identified by the session chair
    (c) New case studies will be prepared by Norway and Croatia
    (d) UNECE will aim to establish a 'wiki' for editing and possibly publication of case studies
    (e) Countries that cannot provide end-to-end case studies should be encouraged to use only a selection of the template for providing a partial case study
    (f) The material collected before and during the workshop will be published on the CMF website.
    (g) New Zealand, acting as the editor, will oversee the synthesis and organize of the information for part C.

28. The following recommendations were made in connections to other parts of the CMF:
    (a) Part B should provide more information about application of standards – which standard for what purpose.
    (b) Part D should include case studies developed for Part C, a summary of lessons learned (to be taken from case studies and categorised), success stories, best/good practices. More contributions should be solicited.

29. Possible topics for METIS 2008 were also proposed:
    (a) Countries would appreciate if Sweden and Canada present their projects on archiving
    (b) Cyprus offered to present their experience in preparing a strategy for the development of a new integrated metadata driven statistical information system

(c) Presentation of any new case studies that have been prepared
(d) UNIDO volunteered to present a case study using the template – a premiere from the international perspective.

VI. OTHER BUSINESS

30. There were two lunchtime presentations of the Swedish MetaPlus application.

31. The representative of Eurostat announced a training workshop on statistical metadata for beginners that will be held on 8-10 October 2007 at Statistics Norway.

32. The UNIDO representative presented their electronic questionnaire project. UNIDO has started working on development of the SDMX Data Structure Definition (DSD) for industrial statistics.

33. The UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS 2008) is planned to be held on 9-11 April 2008 in Luxembourg. It will be back-to-back with the UNECE/Eurostat/OECD Meeting on Management of Statistical Information Systems (MSIS 2008, 7-9 April 2008) and the OECD Expert Group on Statistical Data and Metadata Exchange (SDMX, Paris, 14-15 April 2008).

* * * *