Distr.
GENERAL

CES/AC.71/2004/22/Add.1
12 March 2004

ENGLISH ONLY

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

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Joint ECE/Eurostat/OECD Meeting on the Management of Statistical Information Systems (MSIS)
(Geneva, 17-19 May 2004)

Topic (v): Review and follow up to the activities of the Conference:

REPORT OF THE FEBRUARY 2004 JOINT UNECE/EUROSTAT/OECD WORK SESSION ON STATISTICAL METADATA

Note prepared by the ECE Secretariat

1. The meeting was held in Geneva, Switzerland from 9 to 11 February 2004. It was attended by participants from: Australia, Austria, Bulgaria, Canada, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Hungary, Ireland, Italy, Lithuania, Netherlands, Norway, Poland, Portugal, Romania, Russian Federation, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, United Kingdom and United States of America. The European Commission was represented by Eurostat. International organizations present were: the United Nations Educational, Scientific and Cultural Organization (UNESCO), the United Nations Industrial Development Organization (UNIDO), the International Labour Organization (ILO), the Food and Agriculture Organization of the United Nations (FAO), the World Health Organization (WHO), the International Monetary Fund (IMF), the Organisation for Economic Cooperation and Development (OECD), the European Central Bank (ECB) and the Bank for International Settlements (BIS).

ORGANIZATION OF THE SESSION

2. The participants adopted the agenda comprising the following substantive topics:
   (i) Functions of metadata in statistical production;
   (ii) Metadata interchange;
   (iii) Metadata models and terminology;
   (iv) Using metadata for searching and finding statistical data in websites and portals.

3. The Work Session was chaired by Mr. Daniel Gillman (United States of America). The following participants acted as Discussants: Mr. Jan Byfuglien (Norway) for topic (i); Mr. Marco Pellegrino (Eurostat) for topic (ii); Mr. Daniel Gillman (United States of America) for topic (iii); and Mr. Denis Ward (OECD) for topic (iv).

4. The invited papers were prepared by the following countries and organizations:
   - Topic (i): Ireland, Sweden and United Kingdom;
   - Topic (ii): Canada, Finland, IMF/OECD/Eurostat and SDMX Initiative;

- Topic (iii): Australia, Italy, and MetaNet project;
- Topic (iv): Canada, Norway and OECD.

Other papers contributed to the meeting were prepared by Australia, Canada, Czech Republic, France, Hungary, Italy, Netherlands, Norway, Romania, Slovenia, United Kingdom, United States, Eurostat, ILO, and FAO.

RECOMMENDATIONS FOR FUTURE WORK

5. The Work Session considered the proposals for future work put forward by the Task Force composed of Mr. Dan Gillman (United States), Ms. Miroslava Brchanova (Czech Republic), Mr. Max Booleman (Netherlands), Mr. Bo Sundgren (Sweden), Mr. Marco Pellegrino (Eurostat), Mr. Denis Ward (OECD) and Mr. Juraj Riecan (UNECE).

6. The Work Session considered having a common framework (models, concepts, definitions, terminology, etc.) as a long-term ultimate output of METIS activities. The representatives of national statistical offices expressed their particular interest in harmonizing frameworks related to macrodata.

7. An output of the METIS activities over the next 18 to 24 months would be a draft framework providing links and context to previous and current international metadata standards initiatives together with comparisons of selected examples of current practices at the national and international levels. This framework and links to recommended practices would be published. Such a framework could be organized along a number of key themes such as:

- processes for metadata collection;
- terminology;
- metadata and data interchange entailing identification of common models between international organizations to which national agencies could map;
- migration strategy from existing fragmented metadata environments;
- metadata dissemination and its relation to dissemination of statistics;
- metadata governance and corporate management issues;
- incorporation of usability concerns in metadata management.

It was suggested to create a Task Force, comprising representatives of national statistical offices, Eurostat, OECD and UNECE, to identify priorities and coordinate the work. Interested participants were invited to contact the UNECE secretariat.

8. The Work Session recommended to hold another meeting to have an in-depth discussion and complete some of the above-specified studies in about 18 months, after which the results would be published. Work on the remaining issues would be undertaken (and published) in subsequent years.

9. In order to highlight the national perspective on metadata issues, it was also recommended to organize the next Work Session at the headquarters of a national statistical office. It was, therefore, agreed that UNECE, in coordination with Eurostat, will approach the heads of possible candidate national statistical offices in this respect.

FURTHER INFORMATION

10. Presentations, detailed summaries prepared by the discussants on individual topics and all background documents are available on the website of the UNECE Statistical Division (http://www.unece.org/stats/documents/2004.02.metis.htm).

ADOPTION OF THE REPORT

11. This report was adopted by the participants at the Work Session before it adjourned.

SUMMARY OF THE DISCUSSION AND MAIN CONCLUSIONS REACHED

A. Functions of metadata in statistical production

Documentation: invited papers by Ireland, Sweden and United Kingdom; supporting papers by France, Hungary, Netherlands, Norway, Romania, Slovenia and Eurostat.

12. The session was organized by Jan Byfuglien (Norway). It aimed at clarifying the different functions of metadata in the statistical production process. The presenters provided examples of how metadata can be integrated into the production cycle, and exchanged experiences. The participants discussed the progress achieved in the area of common typology of metadata.

13. Various functions of metadata were discussed such as explicit, direct descriptions of contents, quality, and storage of data and estimated statistical characteristics, e.g. quality declarations; description of processes behind data, description of resources used by the processes (e.g. data and metadata, forms and rules, methods and tools, equipment, and staff). Finally, metadata may also include general knowledge and experiences. The different uses of metadata were discussed from the viewpoints of different external and internal users. Users of statistics benefit from easier access and interpretation of data, respondents from motivation to participate in surveys and facilitated responses to questionnaires. The focus of this topic was on internal users – designers, operators and managers. Their benefits from improved metadata solutions can be summarized as lowered risk of misinterpretations and misuse, better corporate standardization and less dependency on specific persons. On the other hand, efforts have to be made in preparing documentation and sharing the knowledge. Other sacrifices of internal users and owners (governments, policy makers, etc.) include time and costs to be put into metadata work.

14. The meeting also considered examples of practical applications. Experiences from using XML to store descriptive metadata within the statistical agency were shared. This project, which was first reported at the 2002 METIS work session, is at the testing stage. Another project that was reviewed, aims at automated migration for multiple metadata systems migration to a new single metadata system. This was a part of broader re-engineering and modernization efforts in the statistical agency, which among other components, includes improved Internet dissemination.

15. Several projects and applications related to statistical databases were presented at the meeting by national as well as international statistical agencies. The use of metadata was documented throughout the data life cycle. Federating a variety of statistical systems, heterogeneous on a conceptual as well as technological basis, was demonstrated as an application of reference systems. While the term “reference systems” is sometimes used, some participants stressed that it may need to be more concretely defined; is this a system referencing more substantial information, or is this a system containing verified, documented information ('official statistics')? More precisely, it is also necessary to define internal reference systems and external reference systems, taking into account that some of that data which are used internally within the statistical agency cannot be made available to external users as they contain sensitive or confidential information.

16. Two similar approaches used for presenting metadata were compared. One is to construct the data cubes and present the metadata at the time of dissemination together with the cubes. Another is the metadata driven approach, when metadata are used at the time when cubes are being constructed. A related question was raised in this connection: “What is the difference between statistical metadata and dissemination metadata?”

17. There was not a general agreement on the issue of how and whether to classify metadata. Some delegates pointed out that it might be difficult to categorize metadata, while others thought that it would be useful to agree on a typology of metadata, or at least to try to define some main concepts/terms such as 'statistical metadata' or 'dissemination metadata' in a similar way. It was agreed to study this issue again in the future. The issues related to terminology and models were also considered within the topic on metadata models and terminology.

18. While the classification databases appeared to be quite well developed and partly harmonized, participants felt that other elements were needed to achieve a full metadata solution. It cannot be envisaged

to achieve an ideal global architecture soon, and rather pragmatic step-by-step approaches should be used. The sequence of steps has to be planned in such a way that they provide short-term benefits, but that they also will help to achieve the strategic goal of building a complete metadata driven system. One participant estimated that it may take about 5 years or so to implement a fully consistent metadata system. In connection with classifications, it was pointed out that the CLASET model is ideal for the interchange not only for classifications, but also for any tree-structured metadata.

19. During the discussion, the necessity to motivate all subject-matter statisticians in supporting the centralized metadata systems was mentioned. Many participants warned that the efforts and resources invested in the development of any metadata system are efficient only if this system is broadly used in a statistical agency in all subject-matter areas. A related problem mentioned in the discussion was the ownership problem - that is to decide which unit in the statistical agency is responsible for updating what variable(s), as there may be some overlaps.

20. The need to continue work on harmonizing terminology was discussed. However, there was no clear idea on how this work could be brought forward. Some stressed the need to agree on the definition of some main concepts and to provide mapping between different terms and concepts. For others the difference between 'concept' and 'term' did not appear very clear. Reference was made to the example from the Metanet project trying to reach some agreement on 'statistical unit', 'statistical variable' and 'statistical value'.

21. Based on concrete examples, it was also pointed out that some more harmonization of the thematic structure of the publication programmes and especially the websites of national statistical institutes would be helpful for those seeking information across countries.

B. Metadata interchange

Documentation: invited papers by Canada, Finland, SDMX initiative and joint paper by IMF-OECD-Eurostat; supporting papers by Australia, France, Eurostat and ILO.

22. The session was organized by Marco Pellegrino (Eurostat). The aim of the session was to clarify users' needs, prerequisites and rules for the exchange of data/metadata, to provide examples of how metadata are exchanged, to show on-going progress towards improvement and to discuss how national and international organizations can cooperate to achieve a common goal of improving the exchange of data and metadata.

23. Metadata are a fundamental component of any flow of statistical data between organizations at different geographical levels (national, regional, international). Coordination is therefore required to prevent duplication, to reduce the burden on statistical offices and to assist users in the interpretation of data from different sources. There is a clear need to understand better the nature of what has to be exchanged in terms of the metadata necessary for identifying and describing statistics. Metadata is a means to an end. Developing efficient processes for the exchange and re-use of data and metadata requires developments on two fronts: on "content" issues (providing descriptions on the precise nature of the statistics being exchanged), and on the technical means for managing and transferring the information, i.e. the IT tools and related models of information flows.

24. There is a proliferation of standards and products for metadata in many areas. The general opinion was that the standards address different purposes and are complementary, not competing. Therefore, the different groups working on metadata standards should cooperate and coordinate their work. Experience shows that successful metadata projects in statistical offices have used, to a great extent, the standards that have been successfully implemented in other areas. Networking activities and mapping different metadata standards and approaches has proven to be very useful in this respect (e.g. the Metanet project is a good example). Health Canada presented the Data Documentation Initiative (DDI) standard and tools that have been developing in university data archives and libraries over time, and adapted for official statistics use by Health Canada. Its relationships with ISO 11179 and the Corporate Metadata Repository statistical extension, as well as possible linkages with the SDMX initiative, were also highlighted. DDI XML DTD has

proved to be a workable standard that is now available for improved pay off and delivery from national and international statistical systems.

25. Several presentations dealt with the use of XML-based tools for the exchange of metadata. It was stressed that XML has its limitations and it cannot be seen as a "miracle cure" to metadata exchange problems. XML cannot express semantics, only structure; semantics can be added as comments or annotations, but these have to be read and implemented by the user, not enforced directly. However, the statistical community cannot wait until the tools become perfect. It has to move forward with the available tools to make progress at all.

26. The session considered projects of international cooperation leading to the better exchange of statistical information and advancement in interoperability of statistical systems. An overview of the progress on the SDMX initiative was given: this initiative was launched in 2001 by BIS, ECB, Eurostat, IMF, OECD, World Bank and UN Statistical Division to define more efficient processes for the exchange and sharing of data and metadata, in particular by exploring common e-standards and other standardization activities. A joint IMF-Eurostat-OECD presentation was made on the development and implementation of a metadata model based on the SDDS (Metadata Repositories Project - sponsored by the IMF under the SDMX initiative) and its integration with the Metadata Common Vocabulary. The IMF also reported on progress made towards the implementation of new advanced features for metadata search and re-use within the existing Dissemination Standards Bulletin Board (DSBB). The first target of the SDMX initiative - over the next biennium - will be the release of the version 1.0 standards for SDMX Common Information Model, Data Formats and Core Metadata. More information about the project can be found on its website at http://www.sdmx.org. Some delegates recommended that models and standards be developed after tests have been carried out over a variety of socio-economic statistical domains. It was considered important to focus on developing generalized rather than domain specific models and standards.

27. It was stressed that the exchange of metadata can only be meaningful if it is based on a common terminology. There was discussion on whether it is feasible to reach an agreement on common concepts and terminology. Glossaries and terminologies have been developed over time by different organizations: the METIS group prepared a "Terminology of Statistical Metadata" in 2000; under the umbrella of SDMX, Eurostat and OECD developed a Metadata Common Vocabulary (MCV) aimed at identifying commonly used terms (and related standard definitions) for describing the different types of metadata. Work on these issues is certainly continuing.

28. To benefit fully from the new ICT developments for metadata exchange, statistical organizations must also address governance issues. There is often a lack of coordination between the metadata systems within statistical offices, not to mention coordination with outside partners; in many cases, it is quite difficult to convince statistical offices to participate in the metadata work. SDMX may provide a possibility to statistical communities to reconsider the current pattern of data and metadata exchange and change the existing division of labour. The consortium approach used by SDMX has been proven successful; later on, the consortium can be expanded to other constituencies, including statistical offices. Several participants asked for more involvement of statistical offices (especially metadata experts, not only top management) in the SDMX work, as it has direct implications on the data and metadata systems being built in statistical offices.

29. Other initiatives of exchanging data and metadata across statistical organizations are being conducted in the area of national accounts (the NAWWE project, National Accounts World Wide Exchange), international trade, BIS Data Bank and OECD statistics. The delegation of France presented an inventory of administrative metadata flows between INSEE (France) and Eurostat for the EDIFLOW database. Eurostat presented the CLASET message and related tools for exchanging classifications. ILO demonstrated its experience with gathering and disseminating metadata on household income and expenditure statistics through a posted questionnaire that might in future be filled in through Internet; this permits data entry into a database and produces output in html format of the textual descriptions of each source.

30. Statistics Finland demonstrated an application programming interface built on top of the existing metadata systems. It was pointed out that the legacy systems in statistical offices cannot be discontinued and therefore interfaces have to be built that allow using metadata in a flexible way regardless of the underlying data models and systems. A good interface would allow 'old' and 'new' systems to work together.

31. All institutions have faced difficulties in explaining the need for change and its technological implications to their management. The main challenges are that the exchange of statistical data is still complex, resource intensive and expensive; various international organizations have individual approaches for their constituencies and there are uncertainties about how to proceed with new formats and technologies (XML, web services, etc.). To be successful in building metadata systems, it is recommended to view reality as it is (and not as it should be) and design metadata systems accordingly. This could start a constructive feedback loop that will be more efficient in creating desirable change.

C. Metadata models and terminology

Documentation: invited papers by Australia and Italy; supporting papers by Czech Republic, Netherlands, United States and Eurostat.

32. The session was organized by Daniel Gillman (United States of America). The issues of models and terminology are the key issues, as proved by the fact that the discussion on the preceding topics could not avoid a certain overlap. National statistical agencies recognize the importance of metadata management systems. In this respect, the relationship between metadata management and terminology management was discussed.

33. Many metadata models may be identified, created by different statistical agencies over time for different purposes. Each of the known models is focused on specific use, and none of them seem to be sufficiently general to cover all situations. Several such models were presented at the meeting. There was a general feeling that more complex approaches to metadata models should be envisaged in the near future, however, it would require a concentrated effort. The question of precedence of terminology and models was raised. Several participants expressed their opinion that these are two components of the same development. A related remark was made, that there was no unique understanding of the term "model", and that further clarification is necessary.

34. The importance of metadata terminology was recognized with respect to increasing Internet publishing, and consequent sharing and interchange of statistics. It is important to use a clear and consistent language in communicating with users. There was a discussion on how precise and detailed the terminology should be, and whether it was possible to have a very detailed one. Several delegates suggested that it is important to agree on core fundamental concepts and definitions representing the basis for efficient communication.

35. As part of the future work on METIS, the participants proposed to look for synergies among metadata concepts created within various ongoing and accomplished metadata initiatives. It was suggested that such a review should be undertaken from the viewpoint of national statistical offices and aim at facilitating the convergence of metadata models and terminology. Ultimately this would lead to the publication of recommendations for national statistical offices. These recommendations would also include best practices and common examples.

36. In parallel with the development of a metadata methodology, important efforts are being made, mainly by international organizations, into developing standards. These cover various needs such as the interchange, dissemination and collection of statistics (e.g. SDMX, UN/EDIFACT, ISO, etc.). The discussion stressed that it is important to facilitate not only the legislative process, but also the practical implementation of such standards.

37. The delegates discussed the standard ISO/IEC 111791, called metadata registries (MDR), which is a family of standards for describing data as data elements and value domains. The use of this standard by a statistical agency for metadata management was presented, in particular, links and relationships between the Input Data Warehouse (IDW) and Corporate Metadata Repository (CMR) were discussed. ISO/IEC 11179 is written in 6 Parts: framework, classification, metamodel and basic attributes, definitions, naming and identification, and registration. However, some delegates stressed that ISO/IEC 11179 is a general standard, so some more thought may be needed to link to existing statistical standards and to link to metadata objects beyond boundaries of ISO/IEC 11179. The benefits of the broad use of ISO/IEC 11179 by an increasing number of national statistical agencies are related to the use of the same framework for describing their data and to the ability to share and understand each other's data. An opinion was expressed that ISO/IEC 11179 might enable the statistical offices to describe data from any subject matter area with extensive semantics, to manage the descriptions of data constructs in a uniform manner, to link data constructs with similar semantics, and to build extensive concept systems from semantic descriptions of data constructs.

38. The outcome of Eurostat's MetaNet project was of interest to the participants when discussing metadata models and terminology. The project was active in the period 2000-2003 and covered five areas comprising methodology and tools; harmonization of metadata – structure and definitions; best practices for migration; adoption issues; and terminology. Two models were created within the MetaNet project: the Terminology Model and the Unified Metadata Architecture for Statistics (UMAS). The findings of MetaNet show that the currently used terminology in statistical information systems can be traced to different roots, such as survey processing, informatics, mathematics, library science, economics, social science, etc. It was the opinion of experts within MetaNet that it was not realistic to standardize all these terms under the umbrella of statistics, but rather it was more feasible to try to bring some order and clarity into the variety of terms. The usefulness of definitions collected within the Metadata Common Vocabulary (MCV, SDMX) and within MetaNet was discussed, and it was stressed that while some of them are ready to be used in the operational context, others would need improvement. This is also due to the fact, that most of the definitions are taken from existing standards, and these standards were produced in different contexts.

39. In concluding its discussion on this topic, the Work Session recommended continuing the cooperation between national and international statistical agencies on development of concepts, terminology and models, as well as the coordinated work on development and implementation of metadata related standards, as part of the future work on statistical metadata.

D. Using metadata for searching and finding statistical data in websites and portals

Documentation: invited papers by Canada, Norway and OECD; supporting papers by Canada and FAO.

40. The session was organized by Denis Ward (OECD). It discussed the need to develop a set of minimum requirements on website design and metadata structure that would facilitate searches across databases on Internet, and presented examples of existing best practice at the national and international levels in this area.

41. The huge increase of statistics available on Internet has put more emphasis on metadata for search and navigation. The volume and complexity on the web requires easy access and a good and intuitive metadata structure. Ideally, conditions should be established that would enable users to undertake searches in databases maintained by a number of different national or international agencies. On the other hand, the Internet provides several advantages for metadata dissemination: it allows making metadata easily accessible and modifiable in real time so that users can always have the latest version.

42. The meeting looked at the experience with the implementation of the Guidelines for statistical metadata on the Internet, prepared by Statistics Norway in cooperation with the Work Session on Statistical

1 For more information about the ISO/IEC family of standards see http://www.jtc1.org, and search for "11179".

Metadata in 1998. The basic Guidelines are still valid but the technological advances have made it possible to further develop the requirements for search and navigation. Two trends can be observed concerning data provision on Internet: small and simple tables and graphics to cover the needs of media and the general public, and database access for more advanced users. These developments provide opportunities for new and more advanced metadata solutions, e.g., concerning recommended metadata about concepts and definitions, description of the production of statistics and quality information.

43. In general, statistical offices comply with the minimum Guidelines but there is potential for further improvement concerning search and navigation. Most statistical offices have a hierarchical subject matter classification and a local search engine but only about half of the countries have a list of key words for navigation purposes. The survey showed that often it is not easy to find the metadata on country websites. The list of symbols used is often missing. The more ambitious recommendation in the guidelines to have structured information on statistic production and quality has not yet been followed by all offices. In particular, the links between data and metadata require improvement. Important for future development is the possibility to provide metadata in several layers with different levels of detail for different user groups.

44. The quality of metadata is a crucial issue. Even if structured documentation about the production of statistics exists, the descriptions vary in length, correctness and how easy it is to understand them. Some examples of "best practice" in formulating the metadata would be useful for different types of statistics. It is a great challenge for national statistical institutes to ensure that metadata can be maintained together with the data. An efficient metadata system requires linking existing systems so that metadata can be provided (and updated) in one place. The quality cannot be sufficient before there is a system of automatic maintenance, not only on the web but also within the statistical office.

45. OECD presented an initial proposal for developing content standards for statistical metadata on the Internet drawing from both existing metadata initiatives at the international level and those currently being developed (e.g. under the auspices of SDMX). In the international context, there is a great variation in metadata practices in national agencies and international organizations. The evolution of metadata content standards has not kept pace with infrastructure developments, especially web-based technologies. OECD envisages a single document or framework incorporating existing international metadata content standards - bringing together guidelines, recommendations derived from existing international metadata content standards, plus any new standards that need to be developed. It would also provide comparisons of current practice at the national and international levels. The document would link to and include other international metadata initiatives. Several topics were mentioned that should be covered in this manual, such as pricing of metadata, relations between different standards, metadata governance, principles for maintaining metadata, etc.

46. There was general support for this initiative, though there was a preference towards the preparation of a framework linking and providing a context for existing international metadata guidelines and standards rather than a single all encompassing document. The role of the METIS group in the process of adopting such a framework was discussed. It was stressed that the framework should make references to existing standards and activities instead of trying to cover all topics. It is important to obtain contributions and to delegate responsibilities to update the material in future. Broader consultation is required outside this group among a wide range of countries and organizations.

47. Examples of good practice in organizing the search and metadata on the web were presented. Statistics Canada demonstrated its central metadata repository (the Integrated Metadatabase - IMDB). As a result of the improvements of the search engine on Internet, the search results are much more efficient, bringing users closer to the data they are seeking. Nevertheless, users still have to scroll through many links before finding what they want. A useful tool is the Common Object Registry (COR). It maintains the links between data in all different holdings of Statistics Canada and presents to the user links to related data in all its data holdings. COR makes it possible to create new resource discovery tools, e.g. the Search by Subject (planned to be launched in spring) providing fully automated access to all the objects in COR by subject. The tool makes metadata very visible and an essential component of the data holdings. As a result, the authors are more willing to provide the metadata and are much more concerned about its quality.

48. FAO considered the statistical metadata in websites and portals from an international organization perspective. The ensuing discussion raised the issue of harmonizing the metadata models for international statistical organizations. Current international metadata standards focus on country level metadata. It is apparent that more attention needs to be paid to agreement on the basic metadata models employed at the international level. Agreement at this level across a number of specific domains would then enable national agencies to map their corporate metadata models to those at the international level and facilitate efficient exchange of metadata. While some attempts are made to harmonize metadata models on national level, there is no such activity yet on international level.

49. The importance of usability testing for metadata on Internet was pointed out. A prerequisite for a user-friendly metadata interface is that development should be undertaken by designers who know web usability principles and in consultation with people who know the users and their requirements. Understanding who are the audiences is very important in this respect: users are mostly not statisticians and they might not understand the jargon used in metadata.

* * * * *