UNITED NATIONS **E**

Economic and Social Council

Distr.
GENERAL

CES/2001/30
12 January 2001

ORIGINAL : ENGLISH

STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Forty-ninth plenary session
(Geneva, 11-13 June 2001)

REPORT OF THE NOVEMBER 2000 WORK SESSION ON STATISTICAL METADATA

Note prepared by the Secretariat

1. The meeting was held from 28 - 30 November in Washington, D.C., United States. It was attended by participants from Albania, Austria, Canada, Croatia, Denmark, Finland, Germany, Hungary, Israel, Italy, Netherlands, Norway, Slovenia, Sweden, Switzerland, United Kingdom and United States. Other UN member country present was Australia. The European Commission was represented by Eurostat. International Organizations present were European Free Trade Association (EFTA), Food and Agricultural Organization (FAO), International Labour Office (ILO), International Monetary Fund (IMF), Organisation for Economic Cooperation and Development (OECD), United Nations Educational, Scientific and Cultural Organization (UNESCO), United Nations Industrial Development Organization (UNIDO) and United Nations Statistical Division (UNSD).

2. The meeting was opened by Cathryn Dippo, Associate Commissioner for Survey Methods Research, U.S. Bureau of Labor Statistics, and by Cynthia Z. F. Clark, Associate Director for Methodology and Standards, U.S. Census Bureau. The meeting was also addressed by Katherine Wallman, Chief Statistician, U.S. Office of Management and Budget.

3. The provisional agenda was adopted.

4. Daniel Gillman (United States) was elected Chairman and Jean-Pierre Kent (Netherlands) was elected Vice-Chairman.

ORGANIZATION OF THE SESSION

5. The following substantive topics were discussed at the meeting:

(i) Statistical metadata for dissemination;
(ii) Metadata modelling and terminology issues;
(iii) Needs and responsibilities of international organisations for metadata;
(iv) Impact of the IMF SDDS on statistical practice.

6. The following participants acted as Discussants: Ernie Boyko (Canada) for topic (i); Mark Wallace (United States) for topic (ii); Michael Colledge (OECD) for topic (iii); and Robert Di Calogero (IMF) for topic (iv).

7. The invited papers were prepared by the following countries and organizations:

- by Australia, Canada, Norway and the U.S. Bureau of Labor Statistics for topic (i);
- by Denmark, Slovenia, Sweden and U.S. Census Bureau for topic (ii);
- by OECD, Eurostat, ILO, IMF and UNIDO for topic (iii);
- by IMF for topic (iv).

Other papers contributed to the meeting were prepared by Armenia, Australia, Austria, Azerbaijan, Canada, Denmark, Hungary, Sweden, United Kingdom and Eurostat.

FUTURE WORK

8. The Work Session considered the progress with the development of the following methodological materials: (i) best practices in statistical website design, and (ii) recommendations on formats relevant for the downloading of statistical data from Internet. The participants considered these materials unique and highly efficient tools for the dissemination of statistics in both national and international statistical agencies. It was recommended to finalise them based on the recommendations made during the discussion and on feedback from national and international statistical offices. The participants recommended that after finalisation, both materials should be published in the Conference of European Statisticians "Statistical Standards and Studies" series as soon as possible.

9. The Work Session expressed the view that close cooperation on the development of metadata model of statistical classifications conducted by the Neuchâtel Group and the work of the UNSD Expert Group on International Economic and Social Classifications would be highly desirable.

10. The Work Session discussed a proposal by IMF that DSBB play the role of reference point or anchor for statistical metadata and that a common XML-based language for statistical metadata and data be developed. The need for such a content-based language was widely acknowledged and several participants at the meeting representing, among others, Statistics Canada, the U.S. Bureau of Labor Statistics, U.S. Census Bureau, Eurostat, OECD and the Neuchâtel Group, indicated a willingness to collaborate with IMF in developing such a language.

11. The participants recommended to organise the next Work Session on Statistical Metadata in March 2002 to consider:

(i) infrastructure issues for statistical metadata;
(ii) users and metadata;
(iii) metadata and quality.

The following countries and organisations plan to contribute to the above-mentioned topics: topic (i) - Australia, U.S. Bureau of the Census, Eurostat and IMF (invited papers); Canada, Netherlands, Switzerland, Neuchâtel Group and OECD (contributed papers); topic (ii) - Canada, U.S. Bureau of Labor Statistics and Eurostat (invited papers); topic (iii) - Eurostat, OECD (invited papers); Slovenia, United States, IMF and UNSD (contributed papers).

OTHER BUSINESS

12. The participants adopted the report of the meeting at its closing session.

13. The participants expressed their great appreciation and gratitude to the U.S. Bureau of Labor Statistics and the U.S. Census Bureau for hosting this meeting and for the excellent working conditions.

14. The main conclusions reached by the participants during discussion of the substantive agenda items are outlined (in English only) in the Annex to this note.

ANNEX

MINUTES FROM THE WORK SESSION ON STATISTICAL METADATA
(28-30 November 2000, Washington, D.C., United States)

A. STATISTICAL METADATA FOR DISSEMINATION

1. Metadata play a major role in the dissemination of statistics, including helping users find, understand and assess statistics in the context of their specific objectives. As standard tools and approaches to creating and managing metadata are developed, the metadata used for dissemination can also be used for survey design and statistical production activities.

2. There is a close link between data quality and metadata. Many metadata items provide information about the different dimensions of data quality, like relevance, coherence, consistency, etc.. The same dimensions can be used to assess the quality of metadata as such, to analyse its accuracy, relevance, timeliness, consistency over time and subject areas, interpretability, coherence, and accessibility.

3. It would be desirable to reach an agreement on a relevant standard framework for metadata on data quality. Several national statistical offices and international organizations are working on identification of a set of measures to assess the dimensions of data quality. However, often there is not yet a corporate view on what can be considered good practices in disseminating data quality information, and no consolidated view on users' expectations. It is not possible to rely on one single measure for data quality as user expectations and priorities concerning the quality dimensions differ. Reliable metadata description of statistical methods and concepts can often provide users with a better assessment of data quality than diverse specific measures of the variety of data quality dimensions.

4. The Work Session emphasised the need for agencies to develop jointly a set of measures that could support the data quality objective, minimum standards that should to be included in various dissemination channels, and templates for different styles of dissemination. Statistical agencies are in the position to play a leadership role by setting an example in terms of presentation of data quality attributes and encouraging other agencies to adopt similar practices. Experience with quality guidelines and "standards" in different statistical offices shows that it is not possible to prepare an ideal quality report that would be suitable for all occasions. The important role of research and academia in this work was also stressed. Trade-offs have to be made between different dimensions of quality based on the intended use of the data. Some of the metadata can even be considered confidential as its release could lead to breaches of confidentiality of data.

5. Users, who are the ultimate judges of data quality and the metadata that portray it, have differing needs. Statistical offices have to balance between fulfilling their duty to inform users about data quality and streamlining access to the data.. Often users do not fully understand the different aspects of quality and statistical offices need to raise awareness and educate users about how the information on data quality can be used. It is also important to be realistic about what it is feasible to achieve in terms of data quality, as there is no such thing as a perfect survey and absolute quality. In general, not enough analysis of users' feedback and usability testing is done in statistical offices. There is a significant gap between user expectations and the existing data dissemination practices. The needs and expectations of users with respect to metadata require more exploration.

6. The Internet has an impact on data and metadata that goes far beyond the structuring and packaging of content; it changes the models of communication and creates new methods and patterns of collaboration. An open "bazaar" type model for sharing and collaboration in developing metadata was

discussed as an alternative to a “cathedral” solution. Some participants pointed out, however, that users are often interested in getting data as simply as possible and do not want to take the time to share their experiences. Only certain kinds of users might be willing to contribute, e.g., to publicise the results of their research based on the data.

7. An approach to involve more users in the creation of metadata was presented in the Web-based data access and dissemination system NESSTAR (Networked Social Science Tools and Resources). In this tool, metadata are not limited to what is issued in the statistical production, but are seen as a network of information that is developed and enriched through the lifecycle of the dataset. The extended metadata concept includes various types of knowledge products derived from the use of data.. The knowledge of previous users of data allows to learn from past experiences and to add new approaches. This added value can then be exploited by data producers. The role of the data producer is to initiate the process by publishing the core metadata, use standards that are interoperable on the Web and provide feedback systems. An important concept is information sharing where metadata is the facilitator of the interchange. Version 1 of NESSTAR has been released and is being further developed in the context of other European projects.

8. Sharing information and metadata requires standard solutions both for the technology and the content. To enable sharing, the standards have to be open and to allow links to information at a detailed level. The Internet is a good example of the needed level of standardisation: developing and recommending the basic protocols and general languages provide the necessary level of stability and flexibility that allow the development of the domain-specific standards.

9. The design and use of thesauri for searching online resources on Internet was discussed. Thesauri can be used for providing a mapping between statistical terms and everyday language, and for links among terms in different languages. They allow the use of classifications and standard vocabularies (as a specific form of classification) to index the products to facilitate their search and access. However, the emergence of several domain-specific metadata initiatives and the adoption of domain specific controlled vocabularies in several languages can lead to interoperability problems between metadata standards. At this point, it may not yet be feasible to develop a cross-national statistical thesaurus but it seems to be a good strategic target for the international statistical community in the near future. The role of research in this domain was highlighted.

10. Use of metadata as an information tool is a key process within the e-Government initiatives undertaken in several countries. A drive towards metadata standards and greater interaction of national information resources obliges statisticians to link their metadata solutions to solutions used in other governmental authorities (e.g., development of e-government, online data collection, data retrieval from administrative registers). Integrating government services forces statistical information, as part of this system, to become more user and citizen-oriented.

11. Statistical offices have long experience in providing metadata in paper publications. The challenge is how to handle data quality and metadata in electronic publications, and to ensure that metadata are linked to the data. Often the preparation of metadata in electronic format can be more efficient and flexible than in paper publications. This requires careful management and a policy of constant maintenance and updating which can require a cultural change in the management.

12. The draft methodological material “Best practices in statistical
The meeting considered the draft to be highly needed in statistical offices, especially in those that do not have long experience with the development of Internet sites. The draft will be updated to take into account the recommendations of the meeting and published as a methodological material under the CES “Standards and Studies” series.

13. The statistical offices that are more advanced in the development of the Websites were encouraged to contribute and provide concrete examples of good practices in Website design and management. Australia, Canada, Sweden, USA and UNSD expressed their willingness to contribute to this exercise. All contributions should be sent to Sweden and the ECE secretariat as soon as possible but no later than in the middle of January 2001 to enable to finalise this methodological material quickly.

B. METADATA MODELLING AND TERMINOLOGY ISSUES

14. The Work Session concentrated among other issues on metadata requirements for statistics based on administrative sources. The increasing use of administrative data for statistical purposes raises new problems in specifying relevant metadata. The synergy in the use of administrative data and statistical data requires the solution of problems related to the level of matching administrative and statistical concepts, the quality of data, reference date, inefficient redundancy of data and avoiding redundancy of metadata. Metadata needs to be collected continuously as administrative systems, the related legislation and concepts of variables change over time. This requires cultivating the requisite subject matter expertise within statistical organisations as well as minimising the burden of providing metadata imposed upon agencies responsible for the administrative data.

15. A question was raised about the influence of the use of diverse data sources (administrative registers, enterprises' information systems) on the definition of statistical concepts. The aim in the data collection is to get to as close as possible to the data source but to produce data according to statistical concepts. The statistical concepts have to reflect the changing reality, but the comparability of data over time should be carefully considered.

16. Metadata quality requirements are part of the fundamental principles of official statistics. To facilitate a correct interpretation of data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of statistics. It is essential for the users of statistics to have as complete set of metadata as possible. Therefore, statistical agencies should ensure that descriptions of a complete methodology for all their collections are documented and up-to-date.

17. The introduction of a successful database approach for statistical production requires well-organised metadata. It is a prerequisite to rationalise the whole production process. For a statistical office, it can be recommended to develop and implement an integrated system of metadata resources around a centralised metadata repository. It is often necessary and more efficient to create special metadata architectures for different purposes. The central repository that is able to feed other local metadata systems allows to avoid redundant collection of metadata.

18. International organisations play the most significant role in building common concepts in global statistical community and are therefore crucial in promoting worldwide common understanding. Common templates, based on the minimum metadata requirements for assessing international comparability of statistical data, and promoting and using standards will achieve both goals: foster common understanding and give the worldwide users the opportunity to become acquainted with the meaning of the concepts applied in official statistics on the one hand and to give the NSIs the opportunity to deploy sound and comprehensive metainformation systems on the other hand. More integration of statistical data and metadata models on national and international levels is needed.

19. The work of the Neuchâtel group was presented. The aim of the group is to agree upon a common terminology for statistical classifications and related metadata concepts. It defines the key concepts for how to structure and present classification metadata to different kinds of users. It also aims to bridge the gap in experience and understanding between classification experts and IT specialists. The work of the Neuchâtel Group can be seen as complementary to the UN glossary of statistical terms which

is under development. More close cooperation of all involved groups on the issue of modelling of metadata on statistical classifications was strongly recommended.

20. There is a high number of metadata models available. However, these models have different aims and it will not be possible to develop one single model encompassing all these. We should rather aim for interoperability and consistency of models, and standards for communicating data and metadata. A good tool for such kind of interchange is the XML language.

21. The Work Session considered highly desirable to concentrate in future on the ongoing experiences of existing metadata repositories and possibly to prepare some recommendations for statistical offices concerning the design and implementation of metadata repositories. Metadata structures are more complex than statistical data structures. For the statistical office, it can be recommended to develop and implement an integrated system of metadata sources around a centralised (physical or logical) repository. Building metadata repositories is a complicated task. Statistical offices need to look at the practical (possibly less complex) approaches to achieve a feasible solution.

22. The central repository need not necessarily be just one system. A set of related models can present a workable solution to the problem of metadata sharing. It may consist of a number of subsystems that are linked together to a central repository. The Federation of Unique but Related Metadata Repositories approach is a model that is viable at interagency as well as intra-agency levels. A logical, rather than physical, metadata base approach may be most feasible. This approach can enable agencies to manage metadata at the intra-agency as well as interagency levels if we can agree on a standard core set of metadata tags for XML interchange.

23. A metadata model needs to be a structure complex enough to include the necessary categories of metadata, yet simple enough to be practically implemented and maintained. The role of metadata to support applications using integrated statistical information has become increasingly important. Some people have criticised the ISO 11179 model as being overly complex and, as such, difficult to implement in practice. However, given that the need for a common approach is important, especially as we are moving increasingly toward the release of integrated statistics, steps must be taken to ease the burden of creating and maintaining metamodels like ISO 11179.

24. The development of the statistical and spatial metadata infrastructure support system to drive the integration of data and metadata from multiple agencies needs to be based on international standards, such as the multi-part ISO/IEC 11179. 11179 compliant applications can input metadata from and output metadata to other metamodels, such as the DDI. But, though this is possible from a technological standpoint, we need to agree on standards for metadata components and to work on common definitions of data elements to make this possibility a reality.

25. The Work Session considered the prepared draft for the methodological material "Recommendations for formats relevant for downloading statistical data from Internet". It was decided that more information is needed from countries concerning their current practices and future requirements for formats used to upload and download statistical data on the Internet. Furthermore, international organisations have an important role in identifying the needs for tools and formats that would facilitate data transfer through Internet. The paper should be updated to focus more on the requirements that are specific for statistical data and the integration of the already existing data transfer standards (e.g., EDIFACT, XML). FAO, Eurostat, OECD and UNSD agreed to contribute to the material.

C. NEEDS AND RESPONSIBILITIES OF INTERNATIONAL ORGANISATIONS FOR METADATA

26. The metadata requirements, collection and exchange has to be coordinated between international organisations to facilitate data and metadata exchange and not to overburden countries with duplicate requests from different agencies. The international organisations are obliged to minimise the reporting burden of countries for supplying both data and metadata. It should also be ensured that the data disseminated are accompanied by appropriate metadata. Furthermore, IOs should provide countries with tools for the dissemination of metadata to users. The discussion identified areas where international cooperation could give results in the near future, and areas where further discussion is required for enhanced cooperation.

27. In order to minimise for countries the burden of providing metadata, more integration of data and metadata between international organisations is needed. Coordination of metadata collection would require that international agencies define clearly the purpose for the collection of metadata from countries. An agreement should be reached among international agencies on the collection of metadata from national agencies. International organisations should use as much as possible the metadata collected by focus organisations (e.g., ILO for labour force statistics, FAO for agricultural statistics, etc.). Practical cooperation steps can already been taken in this direction. It requires also to develop a process for coordinated updating of metadata by international agencies and adhering to the principle of free and open access to metadata.

28. It can be identified what metadata could be shared already now. A lot of metadata is available on the Websites of international organisations and national statistical offices. Links could be inserted from the metadata on the Websites of international organisations to the more detailed metadata on national Websites. Coordination of access could be achieved through a single gateway for data and metadata, e.g. through a portal site. A good basis for developing such a portal could be the IMF Dissemination Standards Bulletin Board (DSBB) that could be used as an anchor or reference point. The current status of the DSBB ensures that the metadata is continuously updated and there is always an exact correspondence between the metadata and the data actually disseminated by subscribers on their internet sites (so called National Summary Data Pages).

29. It was agreed that there is a need of a common standard for presentation of metadata. The use of a common metadata template could allow to harmonise the requests for metadata by international organisations, simplify comparison of national methodological practices and facilitate electronic search. The international agencies could agree on the use of an existing template (e.g., IMF SDDS) or to develop a new one. The starting point can be the IMF SDDS which needs strengthening and development to make it applicable to other areas of statistics not covered by the data categories used in SDDS. Agencies requiring a more detailed template could consider developing models that would be consistent with the IMF template or with other generally accepted standards in this area. Eurostat and the European Central Bank are already using the IMF model for Euro- indicators and for monetary aggregates for dissemination through their Websites. It was also suggested that metadata should not be sent to international organisations but should be compiled "in situ" and made accessible via Internet.

30. The meeting also discussed the advantages of using a common XML language for both data and metadata. The need for developing such a content-based language was widely acknowledged and several participants at the conference representing the U.S. Bureau of Labor Statistics, U.S. Census Bureau, OECD, Eurostat, Statistics Canada, and the Neuchâtel group among others indicated a willingness to collaborate with the IMF in developing such a language.

31. The participants highlighted the need for commonly understood terminology as a starting point for standards. In statistical practice a variety of different terms are used for the same concepts. The practical issue is how to identify, or construct, and promote an common international terminology. Many glossaries exist but they are located in different places and the definitions provided are not always consistent. It would be very useful if international agencies could coordinate their activities and further develop a common reference site to bring together the terminologies already available on the Web. An international standard glossary developed in English should be appropriately translated into other languages.

32. There are a number of general international metadata standards available. Also, international organisations are developing several statistical recommendations and guidelines. Better overview of these is needed. The list of methodological publications that is under preparation in the UNSD could be a good basis for that. Dan Gillman (US Bureau of Labor Statistics) promised to put together an inventory of existing metadata standards and to make it available on Internet.

D. IMPACT OF THE IMF SDDS ON STATISTICAL PRACTICE

33. IMF gave an overview of the developments and future plans with the Special Data Dissemination Standard (SDDS). It has been designed as a best practice standard, aiming to bring together dissemination of data for the assessment of macroeconomic policies. The introduction of the standard has required technical assistance from the IMF to improve data dissemination practices and/or for the preparation of metadata suitable for dissemination on the Internet. Lately the emphasis has been put on the development of guidelines and manuals, and organising seminars aimed at training in the new methodologies. The subscription to the SDDS has helped to raise the profile of official statistics among policy makers in many countries.

34. The development of SDDS has been parallel to the growth in the use of Internet for the dissemination of statistical data. For the bulk of the data categories data is available on Internet, either on National Summary Data Pages (NSDP), or on national Websites. SDDS reinforces the importance of Internet as the primary medium for the dissemination of statistical data and creates a tool for the users to better understand the data through the linking of data and metadata.

35. The same approach to metadata in the standard SDDS format has given rise to the dissemination of metadata also by other organisations. The development of the SDDS has contributed to the widespread use of the Internet in the dissemination of both data and metadata through the requirement of a NSDP and the adoption of the SDDS format by other organisations in the dissemination of their metadata.

F. FUTURE WORK

36. The participants recommended to organise the next Work Session on Statistical Metadata in March 2002 to consider:

(iv) infrastructure issues for statistical metadata;
- Repository design,
- Technical solutions for integrating multiple (stovepipe) macro-level sources,
- Use of XML, XML schema, etc. (Who is doing what? How?),
- Organizing and managing the maintenance and updating of metadata (metadata life-cycle),
- Building and maintaining thesaurus and similar types of thematic or topical classification systems,
- Search engines and their implementation within a statistics website,

- Practical tools for documenting data,
- Role of metadata in integrating statistical information Websites with broader government portals,

(v) users and metadata, statistical information portals;

- Use of metadata for explaining statistics, improving statistical literacy, etc.,
- Statistical "info-tainment", role of multimedia in integrating metadata with data,
- Use cases: how does the use of metadata vary by type of user, user task, etc.?
- Aids for information seeking, including site design of portals,
- "wide-area-data-web",
- Integrating statistical agency metamodels with end-user access systems,
- Information retrieval aspects, query languages for users, terminology,
- User studies focused on the use of metadata,

(vi) metadata and quality;

- What measures of data quality should be provided and how?
- How to use the internet for integrating data quality metadata with the data,
- Standards and priorities for data quality metadata,
- Assessing the quality of metadata, the users' perspective,
- How to develop quality measures within administrative records system, i.e., automatic metadata creation,
- How to assess metadata quality across countries.