



**Economic and Social
Council**

Distr.
GENERAL

CES/2000/20
8 November 1999

ORIGINAL : ENGLISH

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Forty-eighth plenary session
(Paris, 13-15 June 2000)

**REPORT OF THE SEPTEMBER 1999 WORK SESSION ON
STATISTICAL METADATA**

Note prepared by the Secretariat

1. The meeting was held from 22-24 September 1999 in Geneva, Switzerland. It was attended by participants from Australia, Austria, Bosnia and Herzegovina, Bulgaria, Canada, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Hungary, Ireland, Italy, Latvia, Netherlands, Norway, Poland, Portugal, Russian Federation, Slovakia, Slovenia, Sweden, Switzerland, United Kingdom and United States. The European Commission was represented by Eurostat. International Organizations present were Food and Agricultural Organization (FAO), International Labour Office (ILO), International Monetary Fund (IMF), Organisation for Economic Cooperation and Development (OECD), United Nations Industrial Development Organization (UNIDO) and United Nations Statistical Division (UNSD). Representatives of Scopeland-Software GmbH (Germany) participated as observers at the invitation of the secretariat.
2. The meeting was opened by Darryl Rhoades, Officer-in-Charge of the UN/ECE Statistical Division.
3. The provisional agenda was adopted.
4. Mr. Dan Gillman (United States) was elected Chairman and Mr. Lars Rauch (Sweden) was elected Vice-Chairman.

CES/2000/20

page 2

GE.99-

ORGANIZATION OF THE SESSION

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

- (i) The role of metadata in supporting the broader use and better understanding of the content of statistical information;
- (ii) Responsibility for the management, control and nurturing of statistical metadata;
- (iii) Review of experiences with implementing the "Guidelines for statistical metadata on the Internet";
- (iv) The adaptation, evaluation and implementation of statistical metadata standards (terminology, taxonomy);
- (v) The integration of the IMF Special Data Dissemination Standards into statistical practice.

6. The following participants acted as Discussants: Mr. Michael Colledge (OECD) for topic (i); Mr. Ernie Boyko (Canada) for topic (ii); Mr. Jan Byfuglien (Norway) for topic (iii); Ms. Cathryn Diplo (United States) for topic (iv); and Mr. Robert Di Calogero (IMF) for topic (v).

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

- by the U.S. Bureau of Labor Statistics and U.S. Bureau of the Census, and OECD for topic (i);
- by the Netherlands, Norway and Slovenia for topic (ii);
- by the ECE secretariat for topic (iii);
- by Austria, Canada and U.S. Bureau of the Census for topic (iv);
- by IMF for topic (v).

Other papers contributed to the meeting were prepared by Australia, Austria, Belgium, Bulgaria, Canada, Czech Republic, Denmark, Estonia, Finland, Hungary, Italy, Israel, Latvia, Lithuania, Luxembourg, the Netherlands, Norway, Poland, Portugal, Russian Federation, Slovakia, Switzerland, Turkey, United Kingdom, United States, Eurostat, Interstate Statistical Committee of the Commonwealth of Independent States, FAO, UNIDO and UNSD.

8. The Work Session considered it important to further work on the following issues to develop methodological materials: (i) recommendations on formats relevant for the downloading of statistical data from Internet, and (ii) best practices in Website design. Austria, the Netherlands, Norway, Sweden, United Kingdom and Eurostat volunteered to cooperate on issue (i). Canada, Norway, United Kingdom, Eurostat and OECD volunteered to cooperate on issue (ii). The meeting recommended that the countries working on those issues inform the next Work Session about the progress in the development of the above-mentioned methodological materials.

9. The Work Session concluded that the material "Guidelines for statistical metadata on the Internet" (CES/1998/32) complied very well with

the needs of statistical agencies. The meeting noted that at its 1998 plenary session the Conference of European Statisticians encouraged the NSOs of the ECE member countries to experiment in using the "Guidelines" in the course of the next year or two and to inform Statistics Norway on the results achieved in using them. The meeting asked the secretariat to notify Statistics Norway and the Chairman of the Conference that it considers that the "Guidelines" should be published in the Conference's "Statistical Standards and Studies" series as soon as possible.

10. The meeting discussed the "Terminology on Statistical Metadata". This methodological material was developed in the framework of the Work Sessions on Statistical Metadata over many years and was finally completed by Dan Gillman (U.S. Bureau of the Census). The participants considered this material to be a highly efficient tool for further work and recommended to publish it in the Conference of European Statisticians' "Standards and Studies" series. Furthermore, it was recommended to maintain the content of this database in future and to keep it electronically. Participants were encouraged to contribute actively to this work. The U.S. Bureau of the Census will chair this activity.

FUTURE WORK

11. The participants recommended to organise the next Work Session on Statistical Metadata in November 2000 to consider:

- (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.

12. The U.S. Bureau of the Census and the U.S. Bureau of Labor Statistics offered to host the 2000 Work Session on Statistical Metadata.

OTHER BUSINESS

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

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

**SUMMARY OF THE MAIN CONCLUSIONS REACHED
AT THE WORK SESSION ON THE SUBSTANTIVE ITEMS OF THE AGENDA**

A. The role of metadata in supporting the broader use and better understanding of the content of statistical information

1. Statistical metadata can be discussed from diverse perspectives. The Work Session focused its attention on statistical metadata as it facilitates sharing, querying and understanding statistical data over the lifetime of the data. The discussion touched on a range of issues, including classifying the users and their uses of statistical metadata, identifying metadata requirements, modelling metadata, metadata dissemination procedures, educating metadata users and metadata producers, and the role of the Output Database Conferences. The important methodological basis for the discussion was the UN/ECE 1995 publication "Guidelines for the modelling of statistical data and metadata".

2. Metadata users and uses were analysed in detail. Two broad classes of metadata users can be distinguished: producers of statistics (designers of data collection processes, data collectors, data processors, data evaluators) and users of statistics (e.g. policy analysts, students, teachers, scientists, journalists and data analysts).

3. The importance of user orientation and of more detailed research on metadata users was highlighted. It is desirable to classify users into groups and to analyse the characteristics of these groups and the tasks for which they need metadata. Although it is not feasible to define precisely a necessary and sufficient set of metadata for any given use/user, the requirements for each use/user in general terms can be specified to provide a basis for practical implementation.

4. The meeting touched upon how the metadata dissemination procedures could better address the diversity of user requirements for metadata. Often users need metadata to see whether the data suits their purpose. The challenge is in providing the metadata that the user needs and nothing more. The dissemination systems should allow the user to choose readily the metadata he or she needs at the appropriate level of detail.

5. An important issue in this respect is metadata quality. Metadata quality can be estimated as the degree to which the metadata serves its purpose, i.e. allows the user to find and understand the data which it describes. Some attributes mentioned that could be of assistance in measuring the quality of metadata were: completeness of the set of metadata attributes, unique description of each data object by the mandatory metadata attributes, existence of and adherence to specified guidelines in forming

definitions, completeness of classification schemes, etc. Research is needed on the development of these criteria and on ways of measuring metadata quality using them.

6. The successful fulfilment of the role of metadata in supporting data dissemination necessitates educating and training data users with the aim of improving their knowledge and use of metadata for the analysis and interpretation of data. Survey designers and data collectors need to be educated regarding the use of metadata to improve the design and operation of data collection procedures, and to emphasise the importance of servicing the data users.

7. The major goal of metadata design in many statistical agencies is to implement a unified data and metadata system (e.g. corporate metadata repository) that serves both the end users and producers of statistical data. For this purpose, it is necessary to develop standard models for various types of metadata (data element definitions and formats, questionnaires, survey methods, compilation methods, etc.). Different statistical offices have developed different metadata models. It would be beneficial to identify the most useful models that could serve as a recommendation for other statistical offices.

8. Some participants questioned the feasibility of developing a standard model covering all metadata, suggesting it was more practical to narrow the scope and to develop individual models for smaller areas of metadata (e.g. classification servers, survey metadata). Other participants pointed out that this approach might lead to redundancies in metadata, lack of harmonisation between different models and diminishing of the possibility to re-use metadata from different areas. More ambitious models aimed at covering all different types of metadata can provide a more lasting solution to this problem and form a basis for strategic development of future metadata systems.

9. The major aim of the Third Output Database Conference (1-5 March 1999, Canberra and Murramarang Resort, Australia) was to share information and experiences concerning Information Warehousing. The Work Session recommended the exchange of information and close cooperation with the Output Database Conferences in the area of statistical metadata. Possible areas of joint work could be, e.g. the development and use of metadata standards, statistical metadata modelling, metadata management, and the role of metadata in output databases.

10. The need for better international harmonisation of metadata practices in national statistical offices was emphasised. As a starting point for this work, it was considered necessary to obtain an overview of different national practices used as a basis for common approaches in further

development of tools and applications for metadata. Specific proposals for immediate action after the Work Session included the collection and sharing of:

- sets of topics/themes currently used by national statistical offices and international organizations;
- metadata models currently used;
- a set of broad user/task classes into which data uses could be classified, and the metadata needs of each class;
- a checklist of metadata design principles;
- a list of components of metadata quality.

B. Responsibility for the management, control and nurturing of statistical metadata

11. The Work Session discussed strategies for the management, control and nurturing of statistical metadata through metadata collection, production, storage and dissemination processes.

12. The responsibility and most efficient procedures for collecting metadata were considered. Metadata collection is often viewed as a tedious, expensive and time-consuming task. It can not be carried out successfully without the direct involvement of the statisticians who create metadata. The process has to be viewed not as an additional task to statistical work but rather as part of doing the statistical work properly.

13. The strategies for successful metadata production and utilisation are focused on facilitating the production of metadata. The ideal situation might be to build metadata collection, maintenance and access into mainstream survey design and data collection processes. A distinction is made between 'active' metadata (driving any data collection or dissemination process, particularly an automated one) and 'passive' metadata. Some procedures to convert passive metadata into active metadata were considered.

14. The goal is to ensure that the reported metadata presented to the end-users match the metadata that drove the process and emerged during the process. Full correspondence between data and metadata can only be obtained and maintained in a practical manner if metadata are tied to data throughout all stages of the statistical process. For this, a shift in the role of metadata is needed from human-oriented and human-based metadata to machine-oriented and computerised metadata.

15. Some problems related to the collection and maintenance of metadata in register-based statistics were discussed. Often the statistical offices have no control over the quality of metadata. One proposed solution was to have interchange of personnel with the supplier of administrative data.

16. Roles and responsibilities in metadata management and maintenance were discussed. It is reasonable to distinguish between different roles in the process and to assign the responsibilities accordingly. For example, **metadata administrator** whose responsibility is to oversee the content and operational availability of metadata, **theme owner** who is responsible for a particular theme, **data custodian** - a manager to ensure that metadata are loaded and maintained, **Webmaster** responsible for the technical infrastructure, etc. Furthermore, it was emphasised that the control and nurturing of metadata requires a special management effort. There might be a need for the creation of a special unit in the office.

17. An important topic concerning metadata nurturing and dissemination is the interaction with users. Approaches to identify and meet user needs and to bring users into metadata design and development process were considered. Efficient metadata dissemination procedures ought to lead to satisfactory outcomes for both the producers and the users of data.

18. Metadata is a way of communicating with users. The need for a common language in this regard becomes especially important in an international environment (e.g. Internet). An opinion was expressed that solutions to translation problems should be developed in cooperation with NSOs and international organisations collecting statistical data.

19. The control of statistical metadata is further complicated by the proliferation of different (in-house and general-valid) standards for documenting the surveys and collecting metadata. Users are interested in consistency of metadata standards and documentation while metadata producers often ask for more flexibility in recording metadata. Due to the diversity of data collection and production practices, there does not seem to be a 'best way' of creating metadata content or a single standard for metadata content and format. Standardisation of metadata work is an issue within the statistical agencies as well as between different agencies and countries. The relations between standards used within the statistical agency and those used by the rest of their government and by international agencies are often problematic.

20. The need for cooperation between international organisations in collecting metadata from NSOs was emphasised. Coordination is required for identifying a set of common requirements for the collection of metadata from national agencies. This would help to reduce costs by minimising duplication and to avoid imposing undue burden on national agencies. International organisations need both to agree on and adopt a minimum list of metadata items, methodological description within each item as well as procedures for the management of metadata.

21. The Work Session was informed by Statistics Sweden that the software product BRIDGE 2.0 developed in the framework of the Integrated

Metainformation Management System (IMIM) project under the 4th Framework Research Programme of EU is available free-of-charge. Further development of this software product will be organised by the software company run-Software GmbH in Berlin and will be charged.

22. Eurostat presented "Business Methods", a guide to european business statistics methodology containing methods, legal texts and a database of concepts and definitions (CODED). "Business Methods" is accessible via DSIS-IRC (Information Resource Centre) at <http://forum.europa.eu.int/Public/irc/dsis/Home/main>.

C. Review of experiences with implementing the "Guidelines for statistical metadata on the Internet"

23. The "Guidelines" have contributed to raising awareness of the issues of statistical data and metadata disseminated over the Internet. Several countries have used or plan to use the "Guidelines" for developing their Website. In the countries where Internet dissemination was already in place when the "Guidelines" were issued, it has served as a check-list to evaluate the Internet dissemination strategies. However, there seems to be a lack of knowledge about the "Guidelines" and a broader distribution of the document should be ensured.

24. Several countries requested the preparation of examples and "Best practices for a statistical Web-site design". It was pointed out that metadata for international comparability of data, metadata on data quality, metadata formats and design issues could be developed in more detail. Some participants considered it also necessary to upgrade the material into a more extensive document "Guidelines for the modelling of statistical data and metadata".

25. Some issues that were proposed to be included in the "Guidelines" were, however, considered to be out of the scope of this meeting and need to be followed up under other program areas of the CES. Examples of these issues were harmonisation of subject-matter classifications, www services as part of overall data production/dissemination process, management of the www services, pricing policies of Internet dissemination and guidance on technological issues.

26. The Work Session concluded that the "Guidelines" comply with the basic requirements for the preparation of this material expressed by the Bureau of the Conference of European Statisticians in 1997. It concluded therefore that no revision of the existing version of the "Guidelines" is needed.

D. The adaptation, evaluation and implementation of statistical metadata standards (terminology, taxonomy)

27. The meeting reviewed different standards being developed for metadata. Several initiatives that could be relevant to statistical metadata were described in more detail (e.g. Dublin Core, Data Documentation Initiative (DDI), Government Information Locator (GILS), Extensible Markup Language (XML), Machine Readable Cataloging (MARC)).

28. The approaches for working with these different standards vary according to countries' needs for metainformation systems and the stage of its development. One way is to build 'crosswalks' or a mapping of one scheme to another. Another possibility could be to put the metadata content into a database and tag it using some general level metadata standard (e.g. SGML) which makes it possible to map to other standards. It can be recommended to focus on using standard interfaces rather than attempting to choose the "best" one of all existing standards for metadata.

29. It was pointed out that making statistical data holdings accessible via Internet often triggers efforts to find quick "practical" solutions based on message transmission standards. However, still missing are content standards that could host data models capable of representing statistical data sources and data holdings down to details relevant for metadata-driven data processing. Further progress in the field will depend on advances in theoretical modelling of statistical information processing, particularly in official statistics.

30. The advantages and disadvantages of using an object-oriented approach for building data metadata models were considered. The development of a standard set of objects was also discussed. In addition to the standards for data elements (ISO 11179), standards are needed for the whole datasets, surveys, questionnaires, etc.. Further work to establish the feasibility of an object-oriented approach for building corporate metadata repositories is required.

31. Standardization requirements depend on the purposes for which a standard is going to be set up (demand-oriented view). Therefore, there is a need for a family of standards rather than a single standard. Final consumption aimed mainly at human interpretation may include a significant amount of documentary metainformation. Further processing stresses the need for operative metainformation enabling automated process control.

32. In addition to capturing the "active" metadata, there is also a need for standard methods to capture and preserve historical metadata kept in data libraries and archives. Efforts to standardise these two activities have to be made in parallel, together with developing interfaces between these two approaches.

33. The adoption of high-level metadata standards would also contribute to dissemination of "best practices", and thus contribute to better data

quality. The development of metadata standards might imply re-assessment of statistical information processing by (i) evaluating existing processes, methods, resources and organisation schemes, and (ii) joining statistical theory and information management considerations. Data and metadata standards should go hand-in-hand. It is not reasonable to set up metadata standards independent or ignorant of data standards.

34. It was pointed out that there is a culture gap between the information providers/users and the IT world in their perception of metadata. Efficient standards for metadata cannot be developed without a closer cooperation and dialogue between computer scientists and the knowledge workers in agencies responsible for developing and disseminating statistical information to the general public.

35. Participants' attention was drawn to the ISO Open Forum on Metadata Registries that will be held in January 17-21, 2000 in Santa Fe, New Mexico, USA. More information on the conference can be obtained on Internet at the address: <http://www.nist.gov/openforum2000>. National Statistical Offices were encouraged to participate in this event and to join the ISO IEC JT1 SC32 WG2 (Metadata) international standardisation efforts.

36. Another initiative where national statistical offices were encouraged to contribute was the Analytical Data Management Subgroup of the Object Management Group (OMG) which develops specifications for the software industry. More information on its activities can be obtained at <http://www.omg.org>.

37. The meeting discussed the "Terminology on Statistical Metadata". This methodological material was developed in the framework of the Work Sessions on Statistical Metadata over many years and was finally completed by Dan Gillman (USA). The participants considered this material to be a highly efficient tool for further work and recommended to publish it as an official UN publication. Furthermore, it was recommended to maintain the content of this database in future and to keep it electronically. Participants were encouraged to contribute actively to this work. Dan Gillman agreed to chair this activity.

E. The integration of the IMF Special Data Dissemination Standards (SDDS) into statistical practice

38. The Work Session discussed the SDDS in relation to statistical practice. The subscription to the SDDS could provide a framework and an incentive to initiate improvements in national statistical practices. In many countries the SDDS has had a positive impact on coordination among statistics-producing agencies at the national level but there have been problems in some countries.

39. The uniform format for the presentation of metadata posted on the DSBB is now well recognised among users. As a result the format has proved useful for the organisation of metadata on publicly disseminated data beyond the SDDS categories. For example, the IMF has been collaborating with Eurostat and the Central European Bank on a project to organise the presentation of metadata on euro-indicators in SDDS format. The possibilities to develop cooperation with other organizations to utilize the SDDS framework for the dissemination of metadata on regional economic and financial aggregates were considered. Such metadata are useful in their own right and complement the metadata disseminated under the SDDS.

40. Among the concerns which have been raised is the fact that the SDDS may focus too heavily on the timeliness of data dissemination, thus sacrificing data quality. The SDDS allows some flexibility in timeliness to take account of differing national circumstances. However, it was noted that timeliness should not be an impediment to quality because untimely data, no matter of what quality, would not be used by market participants. Users would seek other unofficial sources of data to make decisions on a timely basis. A related issue concerns the impact of SDDS subscription on resource allocation, specifically whether efforts to meet SDDS requirements have absorbed resources that would have been used to compile and disseminate data more relevant to national circumstances.

41. Another issue was the level of detail of the metadata, specifically the need to strike a balance between the amount of available information and the scarce resources to provide it. Users of the DSBB need sufficient information to be able to determine whether the data are adequate for their purposes. On the other hand, statistical offices have scarce resources to produce or reformat detailed metadata.

42. Some participants expressed an opinion that the contents of the metadata could have been dealt with more pragmatically by the IMF. It was also pointed out that the IMF Transparency Reports did not take plans for future work into account, but just stated whether a country passed or failed the Standard on a particular date.

43. The ways to improve the dissemination of information on the SDDS and to encourage an interactive flow of information in future were considered. The Work Session recommended that the UN/ECE secretariat and IMF would continue cooperation on this issue in the framework of the Work Sessions on Statistical Metadata.