NATIONS UNIES **E**

Conseil économique et social

Distr.
GÉNÉRALE

CES/2002/23
15 avril 2002

FRANÇAIS
Original: ANGLAIS

COMMISSION DE STATISTIQUE et
COMMISSION ÉCONOMIQUE POUR L'EUROPE

CONFÉRENCE DES STATISTICIENS EUROPÉENS

Cinquantième réunion plénière
(Paris, 10-12 juin 2002)

RAPPORT DE LA RÉUNION DE TRAVAIL CEE-ONU/EUROSTAT SUR LES MÉTADONNÉES STATISTIQUES TENUE EN MARS 2002

Note du Secrétariat

1. La réunion a eu lieu du 6 au 8 mars 2002 à Luxembourg. Y ont participé les représentants des pays suivants: Autriche, Bulgarie, Canada, Chypre, Danemark, Espagne, Estonie, États-Unis, Fédération de Russie, Finlande, Hongrie, Irlande, Israël, Italie, Lettonie, Lituanie, Luxembourg, Pays-Bas, Pologne, Portugal, République tchèque, Roumanie, Royaume-Uni, Slovaquie, Slovénie, Suède et Suisse. L'Australie a participé à la réunion en application de l'article 11 du mandat de la CEE. La Commission européenne était représentée par Eurostat. Étaient également présentes les organisations internationales suivantes: Association européenne de libre-échange (AELE), Fonds monétaire international (FMI), Organisation des Nations Unies pour l'alimentation et l'agriculture (FAO) et Organisation de coopération et de développement économiques (OCDE). La Banque centrale européenne et la Banque nationale de Belgique ont participé à la réunion en tant qu'observateurs à l'invitation d'Eurostat.

2. La réunion a été déclarée ouverte par Jean Heller, chef d'unité (Eurostat).

3. L'ordre du jour provisoire a été adopté.

4. La réunion de travail a été présidée par M. Daniel Gillman (États-Unis). M. Marco Pellegrino (Eurostat) a rempli les fonctions de coprésident.

ORGANISATION DE LA RÉUNION

5. La réunion a examiné les questions suivantes:

i) Questions d'infrastructure pour les métadonnées statistiques, y compris le projet de recherche MetaNet de l'Union européenne;

ii) Utilisateurs et métadonnées, portails d'informations statistiques;

iii) Qualité des métadonnées.

6. Les participants ci-après ont fait fonction d'animateurs: Paul Johanis (Canada) et Michel Colledge (OCDE) pour le thème i); Lars Rauch (Suède) pour le thème ii) et Cathryn Dippo (États-Unis) et Jozica Klep (Slovénie) pour le thème iii).

7. Des contributions ont été préparées par les pays et organisations suivants:

- Eurostat, FMI, Bureau of the Census des États-Unis, Australie et projet MetaNet sur le thème i);
- Canada et Bureau of Labor Statistics des États-Unis sur le thème ii);
- Canada, Eurostat, FMI et OCDE sur le thème iii).

D'autres contributions ont été préparées par l'Arménie, l'Azerbaïdjan, la Bulgarie, l'Irlande, le Kirghizistan, la Lituanie, la Norvège, la République tchèque, le Royaume-Uni, la Slovénie, la Suède, le secrétariat de la CEE-ONU, Eurostat, l'OCDE et l'Équipe spéciale sur la norme commune SDMX.

8. Les participants ont adopté le rapport de la réunion lors de la séance de clôture. Les principales conclusions auxquelles ils sont parvenus à l'issue des débats sur les questions de fond inscrites à l'ordre du jour sont brièvement décrites (en anglais seulement) à l'annexe de la présente note.

TRAVAUX FUTURS

9. La réunion du travail a recommandé qu'une nouvelle réunion soit consacrée aux métadonnées statistiques en 2003/2004. Elle a par conséquent proposé d'incorporer le texte suivant dans la présentation intégrée du programme de travail de la Conférence des statisticiens européens pour 2003/2004 au titre de l'activité 2.3 − Diffusion et échange d'informations statistiques:

> Réunion de travail CEE/Eurostat sur les métadonnées statistiques en 2003/2004 consacrées aux questions suivantes:
>
> i) Utilisation des métadonnées tout au long de l'enquête;
>
> ii) Utilisation de la norme ML et du Web dans les systèmes de métadonnées;
>
> iii) Développement et affinement des modèles de métadonnées;
>
> iv) Utilisation des métadonnées pour la recherche de données statistiques sur les sites Web et les portails Internet.

10. Les participants ont chaleureusement remercié Eurostat d'avoir accueilli la réunion ainsi que pour les excellentes conditions de travail.

ANNEX

SUMMARY OF THE MAIN CONCLUSIONS REACHED AT THE MARCH 2002 JOINT UNECE/EUROSTAT WORK SESSION ON STATISTICAL METADATA

I. Infrastructure issues for statistical metadata

1. The infrastructure issues for statistical metadata were considered from the viewpoint of the architecture, design and implementation of statistical metainformation systems. The focus was on the exchange of practical experiences, primarily from the perspective of data and metadata producers. It was suggested that many national and international metadata systems are approaching a consolidation period, during which the experience gained can be shared and new directions established. Design of metadata models has become more focused on specific aspects of metadata systems. Attention must be paid to emerging technologies as well as to methodological approaches.

2. Eurostat presented its thesaurus, *Theseus,* developed with a focus on its reference database *NewCronos* and with the objective of assisting users in searching the data. In this context, a thesaurus was defined as a structured list of the expressions used for describing the content of documents and searching these documents. *Theseus* is a multilingual tool, presently in English, French and German. The indexation procedure which links the major metadata of NewCronos to the contents of the thesaurus is an essential component in the global metadata architecture. A side benefit of indexation has been an improvement in the quality of metadata. The realization of *Theseus* involved the development of the information management system (Oracle RDBMS, client-server applications) and a web-interface for consultation. The content required specialised expertise from a group of *documentalists* having the appropriate linguistic and content knowledge. *Theseus* must be constantly maintained to follow the evolution of the *NewCronos* database. The meeting recommended that the data structure and content of *Theseus* be made available to national statistical offices, e.g. through Internet, and this was promised before summer 2002.

3. There was discussion of the requirements and features of metadata search infrastructure. The meeting considered how metadata repositories could be used to facilitate searching, finding and interpreting statistical data and how thesauri, indexing, meta-tagging, database searching and web page searching could be organized and coordinated. It was stressed that the users' needs are a driving force behind any strategy.

4. It was noted that enhancement of data and metadata exchange facilities requires development on two fronts: content and structure. Given the diverse requirements of different national and international statistical organizations, reaching an agreement on a detailed common model for metadata collection and dissemination has proven impossible. It is more realistic to develop or adopt open standards that everyone can follow according to their own requirements. Cooperation among international organizations should be focused on the identification of discrete, univocal metadata elements and related terminology providing sufficient information to enable the numerous metadata contributors to be mutually consistent.

5. Eurostat elaborated a harmonized template for explanatory texts within the *NewCronos* reference database. The template, which will soon be available upon request, reflects a detailed

metadata typology developed in collaboration with data producers. Such a typology will allow to store detailed metadata in the reference base, serving any dissemination format needed, such as the templates used in the IMF Dissemination Standards Bulletin Board (DSBB).

6. The need for ongoing work on metadata concept models or frameworks was recognized. Many such models have been developed by national agencies and international organizations. The discussion stressed the importance of developing a common understanding of the metadata atomic elements, concepts and terminology incorporated in these models. In order to utilise fully the work already completed by international organizations in this area, it was recommended that existing definitions be incorporated in a glossary that is readily accessible to national agencies, international organizations and project groups such as SDMX and MetaNet. OECD and Eurostat are already working on a common glossary of statistical data and metadata that could also provide some semantic underpinning. Other organizations and research institutions could help in this activity.

7. An emerging technology that clearly holds huge potential for metadata systems is XML. It is not dependent on any platform or software and may be used to export and import metadata from different software. It is used for developing elements for storing, transmitting and displaying statistical data and metadata. These elements are formally described using XML schema. Several international organizations are working on the use of XML for accessing, disseminating and sharing statistical data and metadata. A task force, initiated by BIS, ECB, Eurostat, IMF, OECD and UNSD is aiming at an XML based standard for Statistical Data and Metadata Exchange (SDMX). A report on SDMX developments was submitted to the ongoing UN Statistical Commission.

8. While XML has the potential to displace many technologies currently used in the transmission, storage and display of statistical data, it is not likely to do so in the short term. The limitations of XML with respect to many relations was also noted. The meeting considered what specific strategic uses of XML the international statistical community should be aiming to develop, share and disseminate. It seems that enough agencies are pursuing XML solutions to constitute a critical mass for discussing the issues and developing standard solutions. It would be useful to have a focal point providing information about the activities of statistical offices concerning XML, e.g. using the MetaNet project website or www.xml.org.

9. It was recommended that common XML schemas be developed for specific, restricted areas as the first step in the direction of standardization in this area. A possible example would be the creation of an XML schema for a classification building, using the experience of the Neuchatel group. Canada, Denmark and Australia expressed interest in working on this. XML is a good tool for the classification schema as it is designed to handle trees and hierarchies.

10. The IMF described how it plans to upgrade the functionality and capabilities of the DSBB through the introduction of a relational database management system combined with the rendering of the SDDS metadata model in XML vocabularies and schemas. The main aims are to enhance the metadata content management, develop XML capabilities for disseminating metadata, and to improve the interactivity and data query facilities. Thus, the system will provide more sophisticated and reliable data management functions, intelligent searching, dynamic querying and information discovery functions, also content aggregation technology to automate SDDS observance monitoring.

11. The SDDS is hyperlinked to the subscribers National Summary Data Pages, and to websites of other organizations such as Eurostat's Euro-indicators. The similarity in presentation format for statistical metadata on these sites is a first step in providing users worldwide with access to information on multiple sites in a readily recognizable and comparable form.

12. Metadata systems are an integral part of data management strategies. Several offices are developing warehouses including corporate metadata repositories. There is a transition from the use of metadata primarily for dissemination purposes towards their use throughout the entire statistical production process. Crucial elements of implementation activities include user acceptance testing, training staff and establishing the need for quality metadata. Effort has to be made to improve and publicize the available tools, to write user guides, and to provide courses in documentation. There is great need for education on metadata regarding the available tools and templates, how the parts of the system are connected and about the areas of use for the completed documentation.

13. The engagement of top management is a vital success factor. Data and metadata management projects have had a tendency to be under resourced in subject area departments. Metadata are now getting higher priority. It was also pointed out that making the metadata publicly available helps to improve quality as it highlights problems, thereby encouraging the metadata producers to solve them.

MetaNet project

14. The different metadata models and systems have come to a critical mass where analysis, comparison, harmonization and linking of different systems become an important issue. An illustration of what can be accomplished in this regard is the work being done in the MetaNet project funded by the European Union Fifth Framework Research and Development Programme.

15. The MetaNet is a network of excellence to share experiences between individual research and development and other expert projects. The network is bringing together experts and users from NSIs, users of official statistics, researchers and developers. The aim is to harmonize metadata and the methodology, definitions and models to describe statistical processing systems. There is a lot of experience with these kinds of projects in the European Union's Dosis project and national statistical offices but the coordination between these projects and with other related activities has been lacking. MetaNet is an open access network where it is possible to participate at different levels, some of these are still open for interested participants. More information can be found at http://www.epros.ed.ac.uk/MetaNet.

16. The network has four main objectives: to develop proposals for standards in the methodology for describing statistical metadata and information systems; to develop proposals for metadata objects in a common conceptual model, to disseminate the proposed standards and to interact with the relevant 5th framework programme projects.

17. Four Working Groups are formed to (1) establish a methodology and tools for communication, (2) establish and unify current practice, (3) establish best practices for adopting the outcomes of the work, and (4) to recommend how the results can be put in practice. WG1's report, which is expected to be available in April, covers the dimensions of statistical metadata, tools and metadata models. WG2 is developing a conceptual framework for describing statistical

metadata, identifying 5 canonical dimensions (structure, stage, role, form and function). The work is building on all standards developed for this purpose up to now, like METIS, ISO 11179, XML, Dublin Core, DDI, SDDS/GDDS, OECD, GESMES/CB and many other models.

18. It was decided that the identified gaps in existing models should be covered by identifying 5 tracks: development of metadata repositories in machine-processable form, comparison of existing (small) models, analysis of existing process models inside statistical offices, development of a structural metadata model for active use of metadata in statistical processing, and analysis of usage scenarios.

19. There was general agreement that the issue of statistical metadata should be understood in a broader context crossing the border of traditional statistical information systems.

20. It was proposed to organize, within the framework of MetaNet, a survey in the national statistical offices and relevant research organizations to identify the metadata models and systems in use as well as major gaps and problems in this field. Some participants stressed that to define metadata frameworks/models is a priority task. Metadata terminology should be integrated into predefined models, as these allow the identification of the objects that can be formalized. Metadata models should be linked with the glossary that provides the description of semantics.

II. Users and metadata, statistical information portals

21. It is still an issue to categorize the users in relevant groups with homogeneous needs for metadata support. Often the main distinction is made between internal and external users. However, other relevant breakdowns are needed. Distinction can also be made according to the function for which users need metadata. A good approach can be to identify user communities who share their interests and have similar tasks and similar problems. Classifying users will help to classify their metadata requirements.

22. We do not know enough about needs and behaviour of users. It is often difficult to support users with appropriate metadata, both from the content and the presentation point of view. The Work Session considered different methods to get more information about users. A shift in focus can be observed towards investigating users' feedback and letting users to express their requirements instead of trying to guess what these requirements are. Usability testing, user studies, etc. have been increasingly implemented to help to shape the metadata systems intended for data dissemination.

23. An important aspect in identifying users' needs is to study the context of use. Efforts to build metadata repositories will not be successful unless they work for real people in real settings. U.S. Bureau of Labor Statistics studied what aspects of the respondents influenced their use and creation of metadata and how users make judgments on the relevance of information units. The study considered the survey methodologists as a special group of metadata users. It demonstrates that user-centered approaches provide rich and useful input. Since metadata are intended to be useful to people engaged in tasks in a socio-technical context, an understanding of how real people interact with them provide signals to designers. It allows also to identify where metadata systems fail for a user or group of users. An important piece of information often lacking is rationale information – why something has been done and what other options were

considered. The study demonstrates the overlap of metadata management and knowledge management.

24. An interesting approach to identify the different layers of metadata relevant to different user groups was presented from the quality oriented perspective on metadata. The different layers are based on the different functions of metadata. Further developing of that model might help to identify the metadata needs for different user groups.

25. It seems to be very difficult to harmonise statistical metadata for different contexts of use. A possible solution developed by several international organizations is to concentrate harmonisation efforts on a more "atomic level" of metadata and to keep the user and context related adaptation to a higher level. On the lowest level the metadata would be more abstract and on a higher level the metadata will be available in catalogues, dictionaries, metadata repositories, etc. A future harmonisation of metadata to a certain degree would be very necessary to make it easier to compare data from different sources, etc. That is not only an internal question of a statistical office, but probably even more important for global considerations on national and international levels.

26. Statistical Websites are evolving into the most important dissemination channel for statistics. Most of the statistical offices have their Websites available on Internet but improving their quality is a continuous task. There was a general agreement that it would be desirable to identify a minimum set of key features that should be available on NSO Websites. Also, it would be desirable to elaborate on a "standard" structure of a statistical Website to facilitate users locating information. Recommendations (guidelines) should focus on identifying and promoting best practices in this area. Statistical offices would then be free to choose which one of the recommendations and how to implement taking into account other limitations that they might have because of corporate dissemination standards, requirements from e-government initiatives, etc.

27. Statistical offices need to balance the resource requirements for maintaining and publishing in conventional media with the development and operation of electronic dissemination through the Internet. There are significant costs involved in operating an Internet site (developing and updating content for it, etc.). Client expectations increase and statistical offices' websites need to constantly evolve to meet them. There is a need to develop a strategy to monitor the expectations of the visitors, e.g. provide more guidance and information searching capabilities, add links to other government institutions.

28. It was discussed how to improve the accessibility of statistical information on the Web. Possible solutions to improve the visibility of statistical information on Internet can be registering with Web search engines, or using alternative mechanisms (e.g. statistical clearing houses) to provide links to statistical data. The responsibility to guarantee being discovered on Internet can be delegated to experts within the statistical office or can be outsourced. The possibility was also mentioned to link to the initiatives within the digital library context where very often the research is going on about search engines and statistical information. There is a considerable overlap with knowledge management initiatives and statistical offices should make use of the experience gathered in this area outside the statistical system. There was general agreement that further development of metadata for search should remain on the agenda of statistical offices.

29. It was pointed out that statistical offices should also consider metadata needed for the dissemination of microdata. The request to access microdata for research purposes is increasing. It has a specific target audience with specific kind of metadata needs. As microdata dissemination is restricted because of the confidentiality concerns, the Websites could provide the metadata about microdata so that user can assess whether it would be suitable for his/her purposes and then to make the data request in accordance with the confidentiality regulations.

30. There are already initiatives to develop a network of national statistical databases, so called reference databases of NSO's. The possibility of establishing international statistical portals was considered. Some initiatives have already started. IMF is preparing such portal limited to their area of interest. U.S. has the Fedstats portal to the 70 agencies producing statistical data in United States. A similar portal is available also in United Kingdom.

31. One way would be to develop subject matter oriented portals that could be managed by different international organizations. However, international statistical portals will probably be developed in the future as this requires a considerable level of cooperation and coordination of the content, structure and format of the individual websites that are linked to the portal. Many problems with harmonizing statistical offices websites need to be solved before the development of international portals can become feasible. In a number of countries, the production of statistics is not concentrated in one office, but there are several offices and other organizations that are producing statistics of common interest. There is a need to develop national statistical portals, too. This could be the first step in the direction of a harmonized statistical Internet world.

III. Metadata and quality

32. The topic was considered from two different perspectives: quality of metadata, and metadata about quality of data. In order to be efficient, the metadata quality discussion should focus clearly on a specific aspect of metadata use. The role of metadata in knowledge management provides a good basis for considering metadata quality.

33. Metadata is the main tool for providing information on the quality of data and this topic is therefore very relevant in statistical offices. Some international and national organizations (e.g. Eurostat, IMF, OECD and Statistics Canada) have developed their own data quality frameworks. Although the ways of specifying quality criteria may differ, the basic principles of those frameworks are converging. It is possible to map these frameworks to each other comparatively easily. Statistical offices highlighted the need to have one quality declaration that would suit the requirements of all international organizations and would, in an ideal case, emanate from the production documentation within the NSI.

34. An essential aspect with regard to metadata about quality is the user's ability to understand and to use metadata. Metadata provide an essential service to a wide range of users, which can be classified according to specific profiles (experts, non-experts, data producers, mass-media, etc.). Participants agreed that more attention should be paid to quality dimensions that would be understandable from non-experienced users, such as readability, succinctness, simplicity of use, organizational clarity. Educating users in the use of metadata is important in this respect. Users often need a short summary statement of data quality and do not want to be overloaded with extensive information. Anyway, there will always be a need to provide information that is not included within any frameworks, as it is not possible to take into account all potential needs:

therefore, it is important to be in direct contact with users to address their concerns and questions about data. In United States, attention is also being paid to the access to data and metadata for disabled persons.

35. Feedback from the users community is needed to see how they understood metadata about data quality. This feedback can be used to shape the different quality frameworks. Standard techniques from information sciences could be used to analyze how metadata meet the task of determining data quality. By looking at the basic tasks that users are engaged in, we can identify which basic functions metadata should fulfill and from that derive the requirements for its quality.

36. The meeting considered how to harmonize and improve the quality of metadata. The International Monetary Fund has taken several steps toward this goal. By subscribing to SDDS, countries commit themselves to provide a certain level of quality of the metadata available on the Dissemination Standards Bulletin Board. The possibility was considered to expand the SDDS metadata framework/model to allow national statistical agencies to disseminate metadata for other data categories on their websites in an internationally recognized format. It was pointed out that national statistical offices should be more involved in the development of these kinds of standards to make them applicable to all data dissemination areas. The Data Quality Assessment Framework (DQAF) developed by the IMF uses metadata − whether on the DSBB or other website − to assess the quality of data against international best practices. When many countries have their metadata available on DSBB, market pressure will be an additional factor forcing agencies to guarantee an acceptable level of metadata quality. Moreover, the use of an XML based open exchange system for the dissemination of metadata on the Internet could provide the infrastructure for the automation of data quality assessments using assessment tools such as the DQAF.

37. Different metadata are required in the case of statistical surveys, register-based statistics and a combination of these two. The development of metadata quality frameworks should take into account the different requirements of these collection methods. At present, there is a proliferation of metadata standards developed for specific purposes. In developing and choosing the standard, the different levels of standardization of metadata should be addressed: definitions (units, populations, variables, concepts, etc.) and which elements a standard should contain (concepts, variables, statistical units, classification). It can be recommended to start with relatively simple standards for specific well-identified tasks, and to integrate these standards over time into more comprehensive frameworks. It is also important to define which dimensions a standard should have, in order to set up a framework for standards.

38. The operational and organizational issues are an important aspect in guaranteeing and improving the quality of metadata. Infrastructure, i.e. standards and databases alone, do not yield good quality metadata. In addition to the regular quality dimensions, cultural and organizational aspects of the metadata quality are also extremely important. The application of Total Quality Management principles could help to improve the metadata quality: it would be necessary to link this principle to measures of quality. There are currently no quantitative measures of metadata quality. Furthermore, the cost of achieving good quality metadata should not be neglected; and a balance has to be found between the cost and the ultimate usefulness of metadata quality refinements.

IV. Future Work

39. The participants recommended that the following items be discussed at the next Work Session on Statistical Metadata to be held in 2003/2004:

(i) Metadata uses over the survey life-cycle:

- Examples of uses of metadata in active or passive roles in support of the different steps in the survey life-cycle. (MetaNet stages: definition, production, transformation, dissemination, exchange; or according to other life-cycle typologies);
- Uses for data editing, transformation, estimation;
- Uses for documenting table definition, public-use microdata files for dissemination;

(ii) Uses of XML schema and Web services in metadata systems:

- Progress on SDMX;
- XML for statistical classifications;
- Other uses of XML and web services;

(iii) Extensions and refinements of metadata models:

- Standard high-level categories of metadata;
- Harmonization of terminology for metadata and relationship to metadata models;
- Metadata for analytical studies and derived statistical collections such as the System of National Accounts;

(iv) Using metadata for searching and finding statistical data in websites and portals:

- How should metadata repositories be used to facilitate searching, finding and interpreting statistical data in the collections of national and international statistical offices?
- How should indexing, meta-tagging, database searching, web page searching, use of thesaurus be organized and coordinated in an effective metadata driven search infrastructure?
- Standard themes and topics;
- Metrics about outcomes not just outputs;
- Usability issues.
