NATIONS UNIES | ОБЪЕДИНЕННЫЕ НАЦИИ | UNITED NATIONS

COMMISSION ECONOMIQUE POUR L'EUROPE | ЭКОНОМИЧЕСКАЯ КОМИССИЯ ДЛЯ ЕВРОПЫ | ECONOMIC COMMISSION FOR EUROPE

SEMINAIRE | СЕМИНАР | SEMINAR

STATISTICAL COMMISSION AND
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN
STATISTICIANS

Seminar on integrated statistical information systems and related matters (ISIS 2000)
(Riga, Latvia, 29-31 May 2000)

Distr.
GENERAL

CES/SEM.43/2/Add.1
4 July 2000

ENGLISH ONLY

MINUTES
from the Seminar on Integrated Statistical Information Systems and Related Matters (ISIS 2000)
held in Riga, Latvia, 29-31 May 2000

Topic 1: Improving data dissemination strategies

1. Nowadays, when a growing number of statistical offices are more user-oriented, data dissemination could play an important role in budgeting and planning. The dissemination strategy should be seen in a broader context and should include the issues of communication and training of data producers and users.

2. The dissemination of data, especially via Internet, should comply with users needs to obtain data efficiently and quickly, as well as to meet more complex and broader user requirements. Dissemination strategies should benefit from the new technological developments which permit improved building of statistical databases for easier access, search and retrieval, and for integrating data-metadata and improved data presentation (visualisation, interactivity, maps, integration of geography and statistics, etc.).

3. When considering innovations in data dissemination, three major aspects should be taken into consideration: (i) identification of user requirements and the purpose of dissemination; (ii) conditions of technical implementation; and (iii) its consequences on the organisation and cost of maintenance. Some participants pointed out that in the implementation process, most of the effort is concentrated on the technical implementation and not enough consideration is given to the two other aspects.

4. The discussion also highlighted the fact that the quality of the disseminated data depends heavily on the quality of the data collection and production processes. New technologies enable the collection of statistical data to be organized directly from different sources (e.g. enterprises), and this has a significant impact on data analysis and dissemination. Because of the increased visibility of statistical offices, using Internet places increased pressure on harmonisation and integration of different concepts, classifications and methodologies at national and international levels.

5. The need to carry out market analysis and to learn more about users and their requirements was highlighted. Often statistical offices do not have enough capacity and expertise for this; closer cooperation with research institutes and private companies could be a solution. Several countries stressed the need for improved training of users. Research institutes could have a more important role in the organisation of training for diverse user groups.

6. A balance has to be found between distributing standard output tables and specific information required by a limited group of users. Statistical offices have traditionally distributed large tabular data sets containing much more data than many users want. As the focus shifts to providing information rather than only data, the user will prefer small focused transactions, fulfilling their requirements for privacy and confidentiality constraints. It is often difficult to draw a strict line between the facts and the official explanation of what the data means (e.g. in press releases). Users are not asking for data or metadata but for information. In this connection, the use of mapping technologies for the presentation of statistical outputs was underlined.

7. As more data from multiple sources and time periods become available, users are left with the additional burden of integrating data sets without the necessary tools or the knowledge of whether the data sets can be integrated or if the results are meaningful. The statistical community must work together to produce a public good that would enhance decision-making processes, help minimise data user burden and data uncertainty, and maximise data quality and usefulness. Existing data dissemination tools are not able to provide a problem and solution-oriented view of the data. Participants were informed about some interesting approaches to encouraging the use of common processes for product conception, development and delivery, including integration of customer feedback, re-use of existing data sets, and easy electronic access to all data and metadata sources.

8. The need to conserve historical data and to preserve today’s data for future use was stressed. Two issues need to be addressed in this connection: methodological continuity and formats in which data are stored. Some offices are looking for software independent methods of storing statistical data, such as a statistical table format including metadata, possibly using XML for this purpose.

9. It was pointed out that change of administrative boundaries over time could be significant bottlenecks in integrated data presentation. In this context, the important role of the geographic dimension of statistical data as a key integrator was stressed.

10. The telecommunication connection between the user and statistical offices has now become easy and is almost “standardised” by the Internet. Possible future developments of Internet and its implications for statistical offices were discussed. Electronic dissemination/publishing is, in several countries, becoming the predominant and default mode while paper-based publishing will become a value-added service for most products.

11. Internet is often seen as a cost efficient, standardised communication system. There is a trend in most countries to disseminate data over Internet free-of-charge; it might even be difficult to get acceptance for prices on Internet. Data disseminated free-of-charge could contribute to building a positive image of the statistical office. Internet also offers a good means to implement a pricing system for data dissemination which can be used as a measure for the relevance of statistical output.

12. Unlike the case of paper publications, when additional readers require higher print runs, there are only marginal costs involved in informing additional clients through Internet. However, it should be noted that the costs involved in setting up and maintaining the site can be significant as the content grows. The database publishing method permits the creation and/or updating of Web pages in a dynamic and automated form, using an organised set of information as the source.

13. Data dissemination through Internet requires very thorough explanations and descriptions of the data to avoid misunderstandings and to make the data suitable for professional use. Metadata for finding and accessing the data is as important as that for interpreting the data. It would be desirable if the explanations and

descriptions could come from a centralised metadata system. Canada, the Netherlands, Sweden and the United States are investigating the creation of data and metadata repositories for a more strategic use of their statistical data assets including integrated data products. Eurostat has long been involved in integration and harmonisation policies and strategies for national and regional comparative analysis. In the academic community, massive digital libraries are being developed to allow easy access to multimedia information from a diversity of sources.

Topic 2: Data warehousing and the development and use of statistical databases in a network environment

14. A growing number of national statistical offices are considering output database and data warehouse approaches as a basis for the future development of the data management environment. There are growing demands to link statistics from different subject matter areas for evaluation purposes. A data warehouse approach that is able to join data from different sources would provide the necessary technical support and would play an important role as an information management tool.

15. However, there is still no generally accepted definition of what should be considered as a statistical data warehouse. To be consistent with developments outside the statistical offices, it is important to be careful when speaking about statistical data warehouses so as not to generate confusion. On the other hand, there is common agreement that it is important to make a clear distinction between a statistical data warehouse as a data storage place and application tools that are used to access and analyse stored data.

16. Within the context of a statistical office, a statistical data warehouse can be defined as a single, complete and corporate repository of data and metadata which have been acquired from different sources, assembled, combined to form one structure, documented in a standard format, and stored in a structure that allows users to view, query, combine and download data for analysis at different levels. A data warehouse embraces the entire statistical life cycle by connecting the source systems to the output systems.

17. Concerning the relation between a data warehouse and an output database, different approaches can be seen in statistical offices. A data warehouse typically consists of three main parts: input, data management and dissemination facility. Some offices establish the facility fulfilling the dissemination function as a separate (output) database and the end-users have access to the data warehouse only in exceptional cases; others consider the dissemination facility as an integral part of the warehouse. A distinction between separate databases for internal and external users can also be made because of practical security considerations.

18. Different data warehouse architectures were discussed. Several countries (e.g. Austria, Finland, the Netherlands) use a dimensional model consisting of data cubes and data marts. Thematically-linked data cubes form data marts which are linked with each other. Data marts offer statistical offices a solution for the step-by-step transfer from existing methods of statistics production, where more and more data comprise an integrated data warehouse. Significant gains are expected in terms of checking, correction and analysis of data. However, the complex multi-layer architecture requires the development of a common data model for the whole office which, in practice, can put into question the implementation of the whole data warehouse project.

19. The use of modelling tools (e.g. CASE tools) for the development of database and data warehouse systems was discussed. At present there is no broad use of such tools in the NSOs. But several NSOs have started to test and use powerful CASE tools. In this context, Unified Modelling Language (UML) has also been taken into consideration. UML may be a powerful tool for modelling not only databases but also the complete application.

20. Different tools are available to perform the required functions of a warehouse. At present there are a number of so-called OnLine Analytical Processing (OLAP) tools

that can be used with data warehouse approaches on the market. OLAP tools are provided by all large vendors of database systems and additional tools have been developed by a number of smaller companies. The costs for the different packages may differ significantly and it is always worth making cost/efficiency analyses before any decisions are taken. It has to be recognised that new software packages with new and improved functions will frequently appear. OLAP tools should be regarded as end-user tools and, due to their often relatively short lifetime, it is important that training on their use does not require a lot of resources. They should also be easy to use for inexperienced users as well.

21. While disseminating data from a warehouse via Internet, the application tools to external users must also be provided. The Internet interface should be connected in a very open way to the storage system of statistical data available for public use. In particular, the end-user need not be aware about the technical solution behind the application tools. End-user oriented OLAP tools are often very powerful in their presentation of statistical information using tabulation features, graphical and geographical presentations. Geographical presentation and analysis tools are of growing importance for the advanced use of statistical data sources.

22. The publication of statistical databases via Internet using flexible data access and downloading functions raises the question of data security. Different nets for internal and external use can solve this problem but it is then necessary to ensure that the content of the external database is consistent with the databases in the internal part. This consistency has to be kept dynamically.

23. Data warehouses have a tendency to explode with regard to storage requirements. Today this is not considered as a real hindrance because the storage media are very cheap. On the other hand, the Australian software package SuperCross is a good example of software that solves the storage problem by reducing the necessary amount of storage through advanced data compression and fast processing of microdata. When choosing application tools it is important to carefully specify the users and their demands of the tool. Some users require only powerful and easy cross-tabulation facilities, whereas others need more sophisticated analytical facilities.

24. One of the crucial considerations of building the data warehouse is the organisation of metadata. Metadata must be either a part of the data warehouse or be very closely linked to it. Ideally, there should be a central metadata base used by all programs belonging to the warehouse. The approach to develop a corporate metadata repository in a central position for all metadata in a NSO was presented. Sharing metadata across the whole office can lead to lower costs for maintenance and administration of metadata and can also stimulate a broader use of metadata. In particular, in connection with Internet it is necessary that appropriate metadata be always available. Metadata should be linked to the data and could be made visible, for instance, by using hyperlinks. This would avoid overloading data with metadata but would still ensure that metadata is available when needed.

25. Different solutions have been demonstrated concerning the implementation process of a statistical data warehouse. Very few statistical offices can opt for a "big bang" approach because of the wide scale and high cost of building a warehouse system. Therefore, many offices have chosen an incremental approach based on data marts (or small warehouses) which house smaller, more summary datasets. Some offices have chosen an approach to implement a limited ‘pilot’ exercise, involving one or two data businesses, to analyse the feasibility, practicability, and implications of implementing a comprehensive data warehouse, since the business and cost implications of data warehouse implementation will not become apparent for some time to come.

26. A data warehouse implementation often means introducing an office-wide infrastructure change, which is not possible without the support of high-level management, as well as the staff of the statistical office. For this purpose, in-

office communications to overcome resistance to change, training a critical mass of in-office users, and constant person-to-person support are important.

Topic 3: Innovations in data collection and exchange

27. The growth of electronically distributed services and new methods of communication have created new opportunities for the collection of statistical data. The discussion under this topic considered the innovations in data collection and exchange from the viewpoint of the state-of-the-art technology and concerning the usability of these technologies in countries with different levels of development. The discussion touched upon the innovation in different phases of data collection and exchange: obtaining information from the respondent and transforming it into a standardised format, transmitting the formatted data to the statistical office, and delivering data within the statistical office to the production process.

28. The following new trends were mentioned: use of more powerful “standardised” formats (GESMES, XML), better tools to generate electronic questionnaires, “automated” access to the respondents’ information, transmission via Internet, and combining Internet transmission with electronic questionnaires. The driving forces for innovation are the need to lower the burden of respondents, the need to obtain better and clean data in a more timely manner, and the need to lower the cost of data typing and editing within statistical offices.

29. Most statistical offices have the sophisticated, up-to-date tools that are needed to exchange data with international organisations. However, these tools cannot be applied to collect data from respondents throughout the country, especially in less developed countries. The same is often true concerning data collection from small- and medium-sized enterprises (SMEs) in developed countries. Therefore, statistical offices often need to maintain and link both the "classical" data collection and electronic data interchange facilities. It was recommended to minimise the number of different tools used in one statistical office.

30. New methods of collecting statistical data are most obvious in enterprises. **Electronic questionnaires** (EQ) can improve the process of data collection and reduce the response burden. Wherever possible, these questionnaires should also be combined with functions that retrieve data from the respondent’s information system. An EU project called TELER (TELematics for Enterprise Reporting) has proved this concept to be viable. There are, however, a number of problems that remain to be solved. One of them is the need for harmonised metadata at the data collector’s end and in the enterprises. An efficient promotion of EQ is needed. It was also pointed out that proper training in using EQ is needed not only for IT staff in statistical offices but also for respondents.

31. Lack of standards on the clients’ side makes it difficult to develop EQ applications that work with all common browsers. In some cases, Java is used as a standard language but it requires new skills in the statistical office and its maintenance cost is quite high. Another possible alternative is to use standard packages such as Word, Excel or special tools like Blaise to create advanced questionnaires.

32. Countries also reported their experiences returning EQs via Internet. While in some countries this technology was considered positively, an opinion was expressed that some respondents were not very satisfied for security reasons. The danger imposed by computer viruses was also mentioned in this connection.

33. The increasing use of the UN/ECE **EDIFACT** standard for the exchange of statistical tables and time-series was demonstrated in several countries. The EDIFACT GESMES message is the only format for statistical data exchange between the Central Banks of the European Union countries. It allows easy automation and integration and is comparatively simple to implement. The automation of the corresponding regular data transfers has resulted in a huge efficiency gain. This

demonstrates that GESMES/CB provides the flexibility and efficiency essential for rapidly defining and describing data and metadata structures when new requirements arise. Another success story was reported by Hungary.

34. To assist statisticians at Eurostat and in national statistical offices in their data transmission tasks, several research projects have been launched in order to solve specific problems of the message standardisation (**GESMES**), the data collection monitoring (**STADIUM**) and its inventory (**EDIFLOW**), and the transmission modules hiding the telecommunication layer (**STATEL**), **IDEP/IRIS**. Eurostat also informed that the exchange of experience on best practices concerning electronic data collection is available on http://forum.europa.eu.int/public/irc/dsis/edicom.

35. Extended Markup Language (XML) is regarded as a potential development to enable the use of electronic data exchange (EDI) between small- and medium-sized enterprises (SME-s) and statistical offices, since its implementation does not require large investments. However, the EDI messages represent a unique knowledge of business processes reflecting many years of development by competent people all over the world. Therefore, it is not likely that the XML will completely replace the EDIFACT standard in the near future. In order to increase the acceptance of GESMES, its XML and OO representations are under development. The world-wide standardisation of these representations is planned under the umbrella of XML-EDI. Another development which was mentioned is an Intelligent Questionnaire Markup Language (IQML) including functions of data entry and validation, data extraction from databases, link to data warehouses, etc.

36. The general issue in data collection and exchange is standardisation. Standards are needed both for data and metadata. Standardised tools could significantly facilitate metadata collection, which is often seen as a tedious and unproductive task by the subject-matter statisticians. Standards are also required for the exchange of metadata.

37. Reporting institutions often express concern that they report similar or almost similar data to different organisations using sometimes quite different means (e.g. different classifications and concepts, formats, reporting specifications, etc.). The technical means to achieve interoperability can also facilitate and encourage the harmonisation of statistical data. The benefits of using statistical EDI messages are especially justified in an open community framework.

38. Another important consideration is security of data transfer, including questions of confidentiality, authentication and integrity. It is important that respondents feel confident that organisations such as National Statistical Institutes (NSIs) treat the data with necessary care. Security issues will therefore become increasingly important. Some of the possible solutions mentioned were Secure Sockets Layer (SSL) security that uses passwords as authentication for the respondents, Public Key Infrastructure (PKI), etc.. The level of security for data exchange can vary depending on the sensitivity of data.

39. The costs for managing the exchange of information between public authorities can be reduced drastically when the process can be standardised and carried out via the Internet. However, a statistical office is dependent on other government authorities and national infrastructure (e.g., tax and social insurance agencies) acting as the driving forces in development. These authorities provide the services that will motivate enterprises, organisations and individual persons to acquire the necessary components, such as certificates and EID cards.

Topic 4: Planning and management of statistical projects

40. Planning and management of IT investments becomes increasingly important for statistical offices as the role of IT in statistical production continues to grow. The Seminar discussed several issues that have to be taken into account when

implementing statistical projects, such as the importance of strategic planning, the balance between insourcing and outsourcing, standardisation of tools and methods, setting up control and support for projects, looking at projects in a wider framework of user needs and future maintenance costs, advantages/disadvantages of commercial tools and in-house developed tools, and the importance of project management skills.

41. The major issues that should be considered in project planning are to plan strategically, create the required management structure organisation-wide, use analytical and project management tools to minimise risk and use prototyping and thorough testing. It can be recommended to break down large projects into discrete phases that permit the evaluation of progress before the project continues on to the next phase. It was also pointed out that the project management can sometimes become too big and too bureaucratic. A certain tension can exist between innovation and project management as new concepts and trends unknown during the planning phase must be managed.

42. It is important to create strategic plans that reflect organisational priorities and align IT investments with those priorities. The use of management processes and tools is necessary in order to help evaluate the cost, benefits and risks of IT projects. The organisation should thoroughly review its operational practices before automating them "as is". The use of Business Process Reengineering methods, business case analysis and high level organisational review were discussed in this connection.

43. Strategic planning requires the review of all IT investments from a corporate-wide perspective. Defining the life-cycle costs and benefits of an IT project can be a major challenge. The life-cycle costs can be a combination of internal efficiencies, reduced data collection burden, and increased value created through new products. The managers should continue to measure and capture the costs and savings when the project is realised.

44. Often metrics need to be established to permit testing and validations of a project and its components. These should allow verification and validation to assure that the project development phases are delivering measurable functionality and benefits. Some participants pointed out that such metrics are not well established within the management culture of statistical organisations and that there is no consistent set of measures. For example, human resources are often measured in person-days which is a measure of input rather than a measure of output or results. Consistent *size* and *complexity* measures for the systems developed are also required. These output measures should allow comparison of projects and assessment of the relative effectiveness of project processes and development practices. The participants were informed that the U.S. Bureau of the Census developed methodological material on this issue.

45. The consideration of an organisation from a management perspective can focus on the processes by which organisations learn and mature. How to use the Capability Maturity Model (CMM) for this purpose was demonstrated. The model can be used to rank the maturity of a statistical office and the tools used. Most statistical offices seem to be at the stage where projects are dependent on a few key individuals, with high risk of failure and cost/time overruns, and management is often a case of crisis and intervention. Improved planning, sharing practices and learning from experience would make the performance repeatable. The next stage to be achieved can be characterised by standards defined for deliverables, performance and interactions between players, use of common tools to promote sharing of practices and make it feasible to introduce metrics to manage on a quantitative basis. The ultimate goal for offices is to develop the tools and the infrastructure that have the capacity to learn and to feed this knowledge back to the staff. To be a *learning* organisation means that the agency as a whole learns from its experiences.

46. The possibilities to outsource statistical IT projects or to implement them with statistical offices' IT staff was considered. Specialised IT companies can employ many high quality IT specialists because of their permanent current workload;

this is not possible for statistical offices. Good experiences with using selective sourcing, i.e. the combination of outsourcing and insourcing, were considered. For example, outsourcing can be used for development activities and the statistical office can employ regular IT personnel for the implementation and system maintenance.

47. Statistical offices often encounter difficulties in retaining high-quality IT staff within their employ because the salaries are not competitive with the private sector. One solution could be to employ external experts and outsourcing projects. It was pointed out that special competencies are needed to work with external consultants, to write tenders, to identify the best offer, etc.. Fixed rules of behaviour could facilitate this. In tenders, it might be helpful to specify the detailed results, not the exact process of achieving them.

48. Any project depends on the support of internal and external customers. Therefore, it is important to develop communication strategies for dealing with principal stakeholders, such as customer surveys, sharing information on the systems and processes among executive staff, program managers and customers. Also, analysis of customers' needs is required before going into technical details of the project.

49. Organisation-wide changes meet resistance and take time to implement. To use project management tools efficiently, some offices reported good experiences with training their employees in professional project management, and using multidisciplinary teams to plan, schedule and implement projects. This requires commitment from agency top management, as well as financial investment for training.

50. Strong project management is required to ensure that projects are completed on time and within budget. Another key to success is to regularly implement cost control, risk management and clear communications. Prototyping, pilot projects and extensive testing are important components of a successful IT project.

51. With decentralised management of projects, it is difficult to promote common practices and discipline across projects. Project managers are often drawn from survey management or subject-matter staff and the project management practices vary widely. This increases the risk of inadequate planning and control over projects. To some extent, the use of central methodology and systems staff, and rotation of staff can mitigate the lack of common project practices and provide an opportunity to spread best practices. Statistical offices should capitalise on the eagerness of knowledge workers to use tools and employ this media (in conjunction with training) to spread common practices throughout the office. A "tools, not rules" approach is better suited to organisations that have an interdisciplinary and decentralised nature.

Future work

52. The participants had the following thoughts on the topics which they had recommended for the agenda for ISIS 2002:

(i) Application of Web technology to integrate statistics

This topic should focus on the implications of networks on integrated statistical information systems. The contributions should deal with the following issues: Internet and Intranet applications, internal communication within the office, knowledge-based management, ISIS data banks, use of groupware, data collection through Web, electronic questionnaires, and others.

(ii) Secure communications and data confidentiality

This topic should deal with the problems of security and confidentiality of statistical data under the conditions of networking and increasing user requests for more detailed data. Different technologies can be considered for guaranteeing the

security of data transfer (tools for authentication, data integrity, cryptography, etc.) and for disclosure control.

(iii) Object oriented technologies, component architecture

This topic could deal more broadly with the impact of new IT concepts, methods and technologies of statistical information systems, such as object-oriented analysis and design, unified modelling language (UML), object-oriented databases, component architectures (COM/DCOM, CORBA, Enterprise Java Beans), XML and other new or proposed standards connected with XML and the Web. It could also consider how to cope with the problems caused by the increasingly shorter innovation cycles and the rapid changes in tools and technologies, and how to acquire the necessary skills. Part of this topic could deal with experiences with lesser known software products (Case tools, data modelling tools, object-oriented modelling tools, tools for the management of web sites, data warehousing tools) and with new hardware which might affect the collection and dissemination of statistical data (for example mobile phones with WAP, UMTS phones, handheld PCs).

(iv) Ways of making statistical information systems more responsive to users

This topic should deal with further development in statistical information systems, bearing in mind the needs of both external and internal users, including problems in data warehousing, communication with customers, and any new developments relevant to information society technologies.