UNITED NATIONS

E

Economic and Social Council

Distr.
GENERAL

CES/2001/6
21 March 2001

ORIGINAL: ENGLISH

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

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

Expansion of Internet: a drive to make statistics more widely available[1]

Prepared by the secretariat

I. INTRODUCTION

1. To develop efficient strategies and policies for the dissemination of statistical information under the conditions of rapid development of information and communication technologies (ICT), especially the expansion of Internet services, is a complex task. It remains a challenge for statistical agencies. A primary objective is to find new ways to ensure that the growing number of users has access to and use of the required statistical information. To render official statistics more responsive to the customers of statistical agencies, both respondents and users, requires the re-engineering of many phases of statistical production.

2. Like any other organisation these days, national and international statistical agencies must decide how to adapt to, and when to take advantage of, the rapidly growing role of Internet. Although the Internet will primarily affect the collection and dissemination activities of statistical agencies, it will also impact the management of statistical production and its whole organisation.

3. The aim of this paper is to point out some of the important impacts the expansion of Internet services has on statistical production. Section II focuses on major challenges for statistical agencies; Section III explores online methods and techniques for data collection. Section IV considers users' aspect of data dissemination via Internet, the role of data warehouses and integrating role of metadata. Section V comprises some conclusions.

[1] Delegates who would like to consult the list of references are asked to contact the secretariat.

4. The views expressed in this paper reflect many of the issues discussed by the UN/ECE member countries and international organisations at the meetings on statistical information technology organised by the secretariat under the auspices of the joint programme of work of the Conference of European Statisticians over the last two years.

II. MAJOR CHALLENGES

5. Internet as a challenge to statistical agencies. As the Internet becomes the primary means of interaction between a statistical agency and its publics, demands will change in quite fundamental ways because the expectations of the public regarding the Internet are quite different than traditional sources of government information:

- **Accessibility:** The public will expect much greater access to statistical information and at finer levels of detail. All analysis will be expected to be available at a low/nil cost. There will be pressure to provide public use files in machine-readable form that can be readily input to further analysis.
- **Timeliness:** Internet users expect immediacy of access and usability of results. Traditional compilation and processing cycles for statistical data will become increasingly inadequate.
- **Coherence:** Non-expert users will routinely compare Information from multiple sources. This requires harmonisation of data concepts and common classifications that are broadly adhered to. Inadequacies will be rapidly found by many.
- **Relevance:** Information will be valued based on its relevance to the client's issues. Is it available for the precise population and area under study? Is the metadata readily available? Information will be integrated in various ways and with non-statistical information. Publication by the statistical agency will be a sign of authenticity of content in the face of many new and previously unknown publishers.

6. **Users direct involvement in data production strategies.** The fact that clients of official statistics are communicating via Internet with statistical agencies in a more interactive mode results in the increasing direct involvement of potential users in the development of data collection and dissemination strategies. Furthermore, strengthening the role of users provides new opportunities for statistical agencies to analyse user feedback, to monitor user behaviour and, through this process, to maintain an up-to-date data dissemination strategy. Therefore, a greater effort should be made in order to understand who the Internet users are, since their requirements for information may vary substantially. The dissemination strategy should be seen in a broader context and should include the communication and training issues not only of data producers but also of users.

7. **Resource management.** The use of Internet has a direct impact on staffing plans of statistical agencies. This is especially true for the collection phase of statistical production where the use of Internet clearly causes a shift of data editing and coding workloads from the statistical agency to the respondents. The transfer of resources from clerical mail-handling key-entry activities, post-collection editing and coding to quality assurance, methodological research, analysis, dissemination, system design and maintenance allows the statistical agency to focus scarce staff resources on value-added activities for most cost-effective processes.

8. **Budgeting and planning.** The increasing user-orientation in statistical agencies could play an important role in budgeting and planning. This trend could provide the management of statistical agencies with an efficient means to implement a pricing system for data dissemination, which in turn could be used to measure the relevance of statistical outputs.

9. **Integration of business and statistical information systems.** The use of Internet for online communication with respondents of statistical agencies, especially for online data collection from business enterprises, may lead to changes in the whole architecture of statistical information systems. A trend away from the survey-oriented approach towards the subject-oriented concepts could be expected.

10. **Security and confidentiality.** The need for efficient methods of privacy protection and data confidentiality will become increasingly important. The privacy concerns of individuals and businesses will increase. This will place greater emphasis on communication with respondents at the time of collection and legislated limits on data linkage. Users will require participation in public negotiations about privacy and confidentiality. Statistical agencies will be obliged to inform respondents about what personal data they maintain and for what purposes these data will be used.

11. **Statistical data and metadata concepts.** Increased networking at national and international levels will require more a concentrated effort in the development of standards and other integration tools. Standardised concepts of statistical data and metadata, as well as classifications, will became an important pre-requisite if client expectations for coherence and comparability of statistics are to be fulfilled. Integration will be determined by the adoption of administrative and legislative identities such as common business numbers, standard geographies, etc.

12. **Standardisation of production operations.** In statistical agencies, the Internet is visibly reinforcing a shift towards unification and consolidation in many data collection and dissemination operations. This consolidation may extend to subject-matter domains in terms of harmonisation of concepts, common tools and common statistical processes. Clients will require user-friendly "window" services and will expect that available information is consistent and comparable, even to the extent of using common frames and classification systems

13. **Measurement of data quality.** The definition and measurement of statistical data quality will become more explicit. Multidimensional networking with clients of statistical agencies highlights the importance of developing a generally accepted data quality framework, in order to better assess a comparability of disseminated statistics at both national and international levels.

14. Some countries have already developed a data quality framework where coherence, interpretability, timeliness, accessibility, accuracy and relevance represent objective criteria. There may be other quality attributes, however, such as measures of sampling variability, measures of coverage, indicators of editing quality, impact of confidentiality, etc. Furthermore, other key features of data quality are defined by users and therefore can vary significantly. Readily available metadata describing statistical data quality are thus required. It would be most advantageous to develop internationally accepted models and recommendations for consideration of data quality.

15. **Diverse ways to use Internet.** Although the Internet will become a primary channel for communication, the existing methods of collection, dissemination and interaction with statistical agencies will not disappear. Therefore, statistical agencies should expect to have to deal with mixed modes and should ensure the compatibility of these multiple channels. At least in the short term, this may increase costs and complexity of statistical production.

16. **Impact on statistical information systems at the national level (vertical, horizontal integration).** In general, the Internet will cause the integration of national government services to be client-centric. This means there will be less distinction between departments and agencies from a public perspective. While such a development will be driven by the demand for improved services, it will also raise concerns for privacy and data usage. A positive outcome of this dilemma could be the segregation of statistical activity from the administrative side of government. This would favour the consolidation of statistical agencies at the national level and would likely be accompanied by strict legislation on privacy and confidentiality. In other words, statistical agencies would become the sole legal integrators of information for statistical purposes, while other departments would be restricted to the minimum data required to provide services.

17. **Impact on integration when disseminating statistics internationally.** The Internet significantly accelerates the process of integration of statistical data flow at the international level. Individuals, businesses and other government agencies (national and international) will demand greater

consistency when interacting with statistical agencies. International organisations (IOs) communicating with national statistics play an important role in this process.

18. **Integration tools, need to re-design and re-engineer phases of statistical production.** The tools developed around Internet technologies will have a broad impact on all phases of statistical production. Technologies such as XML will play an important role in defining these interfaces and automating the transfer of information between processing nodes. Integration, in the traditional sense of monolithic consolidation, will be avoided in favour of smaller processing components that can interact and cooperate using common transfer syntax and consistent interface definitions. Advanced Internet tools have often proven to be a driving force for changes in the organisation of statistical services, obliging management of statistical agencies to redesign their organisational models from isolated subject-matter statistical units towards integrated production.

III. INTERNET AND DATA COLLECTION

19. The Internet/Web-based collection of data can bring most radical change in data collection of all computer-assisted data collection methods (like computer-assisted telephone interviewing, computer-assisted personal interviewing, touchtone data entry, voice recognition). User-friendly interfaces developed around the Internet permit data entry, data editing, as well as other facilities such as easy update, directly by respondents. They also offer cost-reducing features, including reducing the number of interviewers in the field.

20. **The Internet will not replace other data collection methods.** It should be underlined, however, that it would not be correct and/or realistic to expect electronic data reporting based on Internet to replace other collection methods in the near future. Therefore, the objective for statistical agencies should be to develop the best set of data collection tools possible rather than to replace all methods by Web-based collection. More generally, statistical agencies should be responsible for the whole process of data collection methods and techniques and their further facilitation.

21. **The Web mode of statistical survey** comprises a computer-assisted self-administered survey without the presence of interviewers, where an electronic questionnaire based on HTML is presented in a standard Web browser, and the responses are transferred to the server through the Internet. New skills will be required in Internet technology, information presentation, questionnaire design and security management.

22. Web surveys are often praised as the cheapest data collection mode, but are also criticised for lowering data quality and providing non-valid results. The cost error issues in these surveys could thus become a more important problem than in other collection modes. It should be mentioned that certain features of Web surveys contribute to increasing data quality. There are no routing errors (e.g. errors in question order, skipping and branching), data can be checked immediately (e.g. range checks, consistency checks), there is a random order of questions and/or answers, there is no separate data entry phase, and records on the interview process are available (for example, time and duration of interview), etc.

23. Low costs of Web surveys, on the other hand, enable large sample sizes, which could decrease a sample variance. This is, however, very often not enough to ensure data quality. The most frequent limitation of Web surveys is the frame deficiency, since not all target units have access to the Web. In household surveys, for example, such a deficiency could considerably impact the entire background of the survey, i.e. class/stratification, education, gender as well as some attitudinal variables. In establishment surveys this problem seems to be less severe.

24. **Non-response rate.** Another important issue is the potential growth of non-response. There are different factors that could influence a response rate in Web-mode surveys: characteristics of the survey sponsor, mandatory status of the data request, size of the firms, respondent selection, follow-up activities, salience of the topic, availability of requested data, survey design, length of data collection

period and many others. One feature should be generally stressed and that is the fact that Web surveys are self-administered. The absence of interviewers and less intensive contact with respondents could result, indeed, in a lowering of the response rates. On the other hand, some research studies on respondents via Internet have shown that adequate weighting adjustments can render Web respondents similar to telephone respondents regarding demographic variables but not in answering attitudinal questions and answers. Clearly, more research in this area is needed.

25. **Public cooperation.** Experiences in many countries demonstrated that an essential strategy for achieving effective results in electronic data collection using Internet is voluntary public cooperation. Even for censuses and surveys identified as mandatory, refusals are possible and can often result in loss of data. The Internet provides an opportunity to enhance public cooperation in data collection by providing an additional response mode, which some user respondents may find more convenient. These include Computer-Assisted Telephone Interviewing, fax and electronic data interchange.

26. **Computer Self-Administered Questionnaires.** A newer, promising technology is Computer Self-Administered Questionnaires (CSAQ). Web-based CSAQs are still in their infancy, very often in the trial phase. The ideal software remains somewhat elusive. Some experts are recommending off-the-shelf software to develop CSAQs and other computer-assisted interviewing technology, but it is often a case of not finding any available which fit the statistical requirements. The use of HTML/Java Script seems to cope with the requirements. It provides benefits of real- time editing and a screen-based design. Sometimes, however, this software combined with the necessary security measures can eliminate many potential respondents.

27. **Implementing CSAQ on the Web.** Early experience would suggest that CSAQ is most suited to short (one page) questionnaires that are repeated periodically. For complex and one-time interactions requiring an explanation of concepts and significant data entry, response rate falls off sharply. The available technology (common browsers) has limited a capability for editing, controlling multiple sessions and managing secure interaction. We need to be careful not to exceed these capabilities in the early adoption of online entry. Confidentiality should be a part of the questionnaire design. To promote trust in the confidentiality of respondent data, it must be clear to the respondent how the data will be utilised, in what kind of data aggregations (thematic, geographic, longitudinal) they will be involved, and to what other information this collection will be linked.

28. **Candidates for electronic data reporting.** In the recent past, economic surveys were considered better candidates for the application of new electronic data reporting technologies because firms are more likely than households to be equipped with access to Internet, personal computers and other relevant equipment. The use of Internet questionnaires, as well as the design to support them, may be quite different between an economic and a demographic survey, for example. An economic survey often requires historical data to be transmitted to the respondent, a lengthy intermittent survey completion time and relatively little concern about non-response bias due to access to a computer. A demographic survey, on the other hand, rarely requires historical files, but may create significant coverage concerns due to access to a computer with an up-to-date browser. The coverage concerns would be overcome by offering the Internet questionnaire as an optional method of reporting, with a paper questionnaire as backup.

29. Findings from diverse trials pushed statistical agencies to investigate several uses of the Internet as a response option for business companies and other organisations. Many of these activities are of a research nature and, therefore, should be elucidated in close cooperation between statistical agencies and research institutes. Research is very often conducted to quantify business concern regarding Internet reporting and also to ensure that statistical agencies do not overlook any special requirements or features. Ultimately, the use of Internet questionnaires, as well as the design needed to support them, may be quite different between an economic and a demographic kind of statistical survey. For surveys based on Internet data collection, a special security system protecting confidentiality of data collected via Internet should be developed.

30. **The Internet will reinforce the integration of business enterprises** for collection operations within statistical agencies. Collection will become more complex as multiple modes, including the Internet, are offered to respondents. Internet has the potential to reduce respondent burden by direct interaction with business application systems.

31. **Privacy, confidentiality and security** will be strategic issues requiring public debate, new legislation and innovative approaches. New skills will be required in Internet technology, questionnaire design and security management. These skills will also be in high demand in industry, resulting in competition for these skilled resources. This is likely to lead to more outsourcing.

32. It is also important for the respondent to know who has access to their data, what protection is in place (legislation) and the redress available to them if this confidentiality is breached. Respondents should be informed about the level of protection provided by security implementations. For example, how is the data encrypted and who has access to the keys? At what point is the data held in clear text format and how secure is this location?

33. **The role of metadata in data collection via Internet.** Metadata and its management are the key issue. Recent goals in R&D on statistical metadata are clearly influenced by the rapidly growing number of diverse client and respondent groups. Metadata as an information tool is a key process within the national 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 the solutions used in other governmental organisations.

34. To cope with this trend, the convergence of statistical metadata producers, users and archives is indispensable. This is, however, a very demanding and expensive process. National statistical agencies in many cases lack resources as well as specialists with the necessary qualifications for their own research and, therefore, the role of R&D in solving this task is vital. The development of common terminology on statistical metadata and the development of metadata catalogues and thesauri as a basis for corporate statistical metadata repositories are the most important tasks. Metadata, registers and directories should be designed to be non-confidential and shareable wherever possible. They should either not contain sensitive material at all or should carefully segregate such data with secure access capabilities.

35. **Data quality.** The effects of online collection on data quality are largely unknown. This implies an increased emphasis on cognitive studies and comparisons among alternative collection channels. We need to determine guidelines for CSAQ design and general interface design. For example, how are respondents informed about data use and privacy? Do users have the flexibility to make comments, receive advice and consult metadata while they respond? Can respondents complete an interaction in multiple sessions? Can respondents review earlier answers and alter their responses? The quality impact of providing these choices must be studied.

IV. INTERNET AND DATA DISSEMINATION

36. Perhaps no infrastructure change has had as significant an impact as the growth of the Internet and the accompanying quantum increase in the number of data customers who now expect and demand full access to an almost unlimited selection of statistical information. Statistical agencies have, of course, responded to the opportunities and challenges presented by this new data dissemination medium. Whereas in the early ‘nineties only a trickle of economic statistics were available "on-line" through dial-in bulletin boards, today every agency of note has a Website full of the most recent numbers. The World Wide Web has brought some existing challenges into sharper focus. Long-standing requirements for accurate, timely and reliable figures face more demanding scrutiny than ever before.

37. **Internet users.** As the audience grows larger it also grows much more diverse. In the past data dissemination vehicles typically assumed a certain level of economic and statistical sophistication. Frequently, the user base for a given set of statistics was as knowledgeable of the subject matter as the

producer themselves. This is clearly no longer the case. The audience for economic statistics, for example, can range from professional economists and policy makers, to interested members of a lay population, to young students working on school assignments. Economic and statistical literacy can no longer be taken for granted.

38. Furthermore, the Internet provides a means for closing the loop between statistical data suppliers and their users. Some respondents may be motivated to respond if they can see the kind of statistical outputs that are generated from their input. Some institutional statistical activities, particularly in the health, education and justice communities, require shared access to microdata. In these cases, the respondents are both suppliers and users of the data.

39. A special effort should be made in statistical agencies to understand who the Internet users are, since their requirements may vary substantially. In general, according to the targets of statistical agencies, the users of statistical information on Internet can be either internal, i.e. statisticians responsible for the production of statistical information, or external users of statistical information.

40. Subject-matter researchers, political decision-makers, public officials, executives, teachers, students, librarians, journalists and others can be identified among external users. It is therefore essential that the statistical agency take account of the variety of users. Another way to classify external users could be by their level of skill in statistics and their interpretation. Bearing this in mind, the following groups can be distinguished: users with limited skills in statistical analysis (general public); skilled users with limited inclination to read the metadata, e.g. preferring ready made compilations; and expert users skilled in searching, retrieving, assessing quality, interpreting and eventually producing statistical information on their own.

41. The range of users with varying interest and skill levels and the availability of large quantities of data will make it imperative that users be able to exercise control over their interaction by personalising the interface. Irrespective of relevance to a specific group, the user of statistical information on Internet, in general, needs metadata for the following functions: to see what data are available, to assist in the search for information, to interpret the information and, if necessary, to assist post processing of the information (downloading and further application).

42. **Stakeholder relations.** Together with a dissemination policy, there has to be a policy for user support in the statistical office. This support should cover both technical and functional issues. It has to be clear to users what level of service they can expect from the statistical agency for each of these areas. It is essential that statistical agencies develop and maintain a mechanism for interacting with the above-mentioned user groups. An important point is that user feedback cannot only provide valuable information for the greater satisfaction of users in future but also for improving the whole production process. Advisory committees and councils, federal, provincial and territorial committees, bilateral relationships with key federal departments, and various user and distributor consultation forums could constitute such mechanisms. User needs can be assessed through surveys and special studies. As programs move online, ways of measuring user satisfaction with the online offering have to be found.

43. Some examples of how a statistical agency might carry out such interacting mechanisms in an online environment are mentioned below:

- Advisory activities. The establishment of a Website for all advisory groups would facilitate an exchange of dialogue and would serve as a repository for all communications and documents.
- Joint program and research activities. Joint program and research initiatives are of particular importance in areas such as justice, education and health that involve other levels of government and the exchange of administrative and survey data. The development of Extranets or virtual private networks will be required to ensure the secure exchange of information. There are often no ready-made solutions for implementing Extranet and they have to be tailor-made. One of the promising new developments is to use the XML format for real time message delivery with administrative partners. This would make it possible to mutually use administrative sources within public

administration and the basic nationwide registers such as population register, business register, etc.

- User consultation activity. Users need to be consulted with respect to survey content and to develop the Agency's products and services. An increasing number of products and services are moving online and this has an impact on libraries and universities, as well as on independent researchers. The Internet can provide unprecedented opportunities to involve these communities in program planning.
- Online training. An online training capability to help support the dissemination of information and to provide an educational opportunity for the general public is a new opportunity provided by the Internet.

44. **Designing Web pages.** The World Wide Web has great potential for improving dissemination of statistical products. It is likely that in the future the quality of services of the statistical agencies will often be judged by the quality of their Web pages. Designing Web pages is a similar craft to that of designing a survey, and statistical agencies should pay appropriate attention to this task. IT managers should bear in mind that to build efficient and user-friendly Websites requires resources. A capable, trained staff is necessary to properly implement a human computer interaction including usability testing of designed Websites. It requires professionals in relevant software engineering methods. Statistical organisations which may already have a pool of evaluators to draw upon can confirm that to be accustomed to evaluating and field testing questionnaire design requires the necessary resources and culture to perform systems usability analysis. Experience also shows, however, that usability engineering can be implemented gradually, starting with a small core of interested personnel and expanding as the efforts demonstrate their usefulness. There are sufficient reference materials available, as well as industry based courses and academic programs to begin.

45. **Data warehouses.** Significant consequences can result for statistical management caused by the movement from subject-oriented collection, production and dissemination of statistics to an integrated approach under the conditions of Internet networking. The development of output databases and data warehouses may change the paradigm of the statistics production process that is traditionally subject-oriented. Contrary to the subject-oriented statistical systems, the development and maintenance of data warehouses and centralised databases could no longer be the responsibility of individual subject-matter statistical departments. The major role of a data warehouse is to combine data and make it as accessible as possible to as many users as possible. A well-designed data warehouse administration with a hierarchical structure up to the top-level management of a statistical agency will be an important success factor in this process.

46. A data warehouse allows the possibility of accumulating data over time and presenting statistics on a continuous basis to the public rather than at discrete intervals. This also means that the trade-off between timeliness and precision could also be continuous. The user would have the choice of gaining early access to preliminary statistics, or to wait until sufficient data could be accumulated to reduce the margin of error.

47. The main problem when developing a statistical data warehouse remains the consistency of data stored and maintained in this vehicle. A sophisticated metadata system and harmonised classifications are needed to support data consistency. Although statistical metadata are increasingly used in the production process, there is still in many cases a lack of unification inside and outside the statistical office. The use of relevant metadata should cover the whole production process, starting with data collection. Setting up a centrally managed metadata system should be the responsibility of the management of the statistical office.

48. **Metadata for dissemination.** Consistency of data should be considered at both national and international levels. The metadata and data requirements, collection and exchange have to be coordinated between international organisations to facilitate data and metadata exchange and not to overburden countries with duplicate requests from different international agencies. Furthermore, international organisations should provide national statistical agencies with tools for dissemination of metadata to users. In order to fulfil this task, greater integration of metadata and data between international organisations is needed.

49. We can already identify what metadata could be shared. 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 solid basis for developing such 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. Such an agreement was already reached among IMF, OECD, Eurostat and other international organisations at the UN/ECE Work Session on Statistical Metadata held in November 2000 in Washington.

V. CONCLUDING REMARKS

50. After all considerations mentioned above there are still many questions to be discussed:

- Should the Internet be regarded as the primary channel for dissemination and collection in statistical agencies?
- Is the Internet transforming our whole approach to statistical production, or is it merely an alternative for input and output?
- Is the Internet impacting the marketing of statistics?
- Should statistical agencies be proactive in establishing privacy practices for the Internet?
- What should be our role in promoting international standards for presentation and description of statistical data?
- What should be role of international cooperation in preparation of guidelines, recommendations and best practices?
- Is the Internet contributing to, or is it a solution for, the technology gap between developing and developed countries?
- Is strengthening cooperation with research and academia needed to solve some new tasks caused by the growing role of Internet in statistical production?

- - - - -