UNITED NATIONS

E

Economic and Social Council

Distr.
GENERAL

CES/AC.71/1999/4 (SUM)
25 November 1998

ORIGINAL: ENGLISH

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Meeting on the Management of Statistical Information Technology
(Geneva, Switzerland, 15-17 February 1999)

Topic (iii): Integration of statistical activities at the national and international levels, including data modelling strategies and standards needed for statistical data integration

AN INFORMATION SYSTEMS ARCHITECTURE FOR NATIONAL AND INTERNATIONAL STATISTICAL ORGANIZATIONS

Submitted by Statistics Sweden[1]

INVITED PAPER

SUMMARY

1. The paper is a methodological report, based upon the author's long-standing experience in designing statistical information systems (SIS) architectures for national and international statistical organizations. Its goal is to assist statistical offices in designing an efficient information systems architecture under conditions of growing users’ demands, increasing international cooperation and constant changes in information technology.

2. The material provides a comprehensive analysis of the existing types of statistical information systems in national and international statistical offices, and outlines development directions for the future. The paper examines

[1] Prepared by Bo Sundgren.

the technical aspects of the proposed SIS architecture, and gives practical recommendations on how to implement a realistic information technology (IT) strategy under conditions of ongoing rapid technological development.

3. The **first chapter** explains the basic concepts used in the report: different types of statistical processes and statistical data, statistical applications and infrastructure, and the flow of data and metadata through the survey process.

4. A statistical information system is composed of subsystems (applications) which collect, process, store, retrieve, analyse and disseminate statistical data. The information systems architecture (ISA) of a statistical office is a common framework within which different subsystems have their respective roles and interact mutually. The paper also analyses some relations between ISA and organization architecture of a statistical office, and describes some specifics of an ISA in statistics.

5. The ISA should reflect the purposes and tasks of the statistical office. One of the reasons for discrepancies between ISA and the existing organizational architecture of statistical offices could be the conflict between the traditional survey-oriented organization of statistical offices and the cross-cutting information needs of statistics users. Since surveys are very often navigated by data collections, the organization of statistical offices is also "input-oriented". This makes it difficult to achieve desirable co-ordination and control across subject-matter areas.

6. Some statistical offices have created special units aimed at servicing special user categories. In view of the technological developments, it may be a more feasible solution to organize a user-oriented clearing-house (as a single unit) with a flexible and open-ended infrastructure.

7. The **second chapter** provides an extensive analysis the following major types of SIS:

survey processing systems,
clearing-house systems, “data warehouses”,
registers,
analytical processing systems.

8. The author reviews the tasks, functions and requirements of each of these major information systems. Special attention is drawn to the **survey processing systems** covering the full life-cycle of a statistical survey: its planning, operation and evaluation.

9. During the planning phase, the designers of the survey make decisions concerning the major purposes and users of the survey, major inputs and outputs, procedures for obtaining the inputs and transforming them into outputs. It is useful if the designers of a statistical survey have access to a knowledge base, containing information about the design of similar or related surveys. To enable

to learn from the experiences gained, all important information on statistical survey design should be documented. Metadata on quality and contents of data and processes, and feed-back from users are a very important part of such documentation.

10. The survey operation phase consists of the following main processes: frame creation, sampling, measurement, data preparation (data entry, coding and data editing), creation of the observation register, estimation, analysis and the presentation and dissemination of results. The estimation process is often combined with production-oriented analysis aimed at improving the quality and efficiency of future surveys.

11. The results of the survey should be made available in user-acceptable form via appropriate distribution channels. In principle, a survey production cycle is completed once the results of the survey have been published. A trend to include the electronic dissemination of the results in the publishing concept can be observed. Important components of the new publishing system could be the clearing-house function and Internet.

12. The survey evaluation consists of checking and evaluating whether the specified end-products have been delivered, the outputs properly published and advertised, the metadata documented and stored, and of the assessment of the production-oriented metadata and user feedback.

13. The **register** function lies in maintaining up-to-date information on all objects belonging to a certain population. The registers can serve as sampling frames for surveys. In addition to maintaining the current status of the population, the register should permit the reconstruction of the population of objects at any point in time, and to reproduce the original status and all events that have affected the objects. A special kind of registers are those containing metadata, such as definitions, links to surveys and data sets, standard formats, value sets, etc..

14. The **clearing-house** function facilitates the exchange of data and metadata between different surveys, registers and analysis functions, including external users. Another label for this function is “data warehouse”. The clearing-house function receives and delivers data and metadata according to specified standard formats, following specified delivery procedures.

15. In addition to the production-oriented **analysis** mentioned above, statistical offices perform some more user-oriented analysis. When such an analysis uses data from several statistical surveys and other sources, the analytical processing system can be regarded as a separate system with interfaces to the survey production systems.

16. The **third chapter** outlines a future information systems architecture for a statistical organization. It is based upon the four major types of statistical information systems specified in the previous chapter. An important component of the proposed architecture is a corporate data warehouse, encompassing all

clearing-house functions and register functions.

17. The future corporate **data warehouse** of a statistical organization includes five compartments:

raw data and metadata;
final observation registers;
final multidimensional statistics;
electronic documents;
global metadata, including registers.

18. Data and metadata in the raw data compartment will not always be in a standardized form. There should be generalized software supporting the standardization of data and metadata. In addition, there should be generalized software tools supporting all important processes and sub-processes in survey processing systems and analytical processing systems.

19. In the case of international organizations, most of the member countries deliver data and at least some metadata electronically. Even so, the data may arrive in many different formats. Thus, a first step will be to standardize incoming data and metadata. This step can be avoided, if member countries agree to provide data and metadata according to some international standard, e.g. the EDIFACT standard for GEneric Statistical MESsages (GESMES). It is important to note that a standard format must include standards for both data and metadata.

20. **Chapter 4** analyses the technical aspects of the proposed architecture. As the statistical information systems should provide information for many different kinds of users, with different and sometimes contradictory needs, flexibility is a particularly important consideration for all the hardware, software and data components.

21. A statistical office very often runs a relatively large number of different statistical applications. However, many of these applications are rather similar in the sense that they perform a limited number of functions, which are typical for information systems supporting statistical surveys, i.e. survey planning, survey operation and survey evaluation functions. When a statistical function or subfunction is analysed, at some stage a level is reached at which the software components need not necessarily be tailored to the needs of statistical applications. Instead, general purpose standard software components may be used. Nowadays, this "general purpose level" may appear rather high up in the systems architecture of a statistical application.

22. The same principle of preferring standard and re-usable components applies to the hardware. The IBM compatible PC has long since become a de facto standard hardware component for statistical organizations and for the users and customers of statistical organizations.

23. The data components of an information system are stored either as physically integrated parts of the application software system or as separate files or databases. Program/data independence is an important requirement

meaning that the software and data components of an information system may be developed and maintained relatively independently of each other. A modification of the contents, structure, or storage of data should not necessitate modifications of programs using the data. On the other hand, it should be possible to modify or add software components without having to redefine data components.

24. Analysis of different information systems architectures leads to a proposal for a multi-tier network-based information systems architecture that balances the needs for centralization and decentralization in a modern statistical organization.

25. **Chapter 5** focuses on the implementation aspects of the proposed IT architecture. Under the conditions of rapid IT development, there must be a realistic plan for implementation which is able to accommodate changes that happen during the implementation process itself. Some recommended development principles are the following:

(i) as the price/performance ratios of standard hardware and software improve all the time, it is better to buy standard components off the shelf rather than develop one's own solutions, and to spend more on hardware capacity rather than complicating a simple software solution;
(ii) it is safer to standardize in terms of interfaces between components rather than in components themselves;
(iii) instead of waiting for better standards, better hardware and software, buy the state-of-the-art hardware and software components, and replace a component with a better one as soon as possible, without having to change any other components;
(iv) have a clear picture of an organization's overall information systems architecture and define a number of strategically important interfaces;
(v) the maximum time-frame for development projects is not more than two years, complex projects should be divided into subprojects with clearly defined results and deadlines, too many and over-ambitious goals should be avoided;
(vi) while migrating to a new technical platform, one can take the opportunity to improve the contents and quality of statistics at the same time but only to the extent that such activities do not threaten the time schedule of the project; depending on how important these improvements really are, some deficiencies might be acceptable and enhanced in the long run by sustainable improvement.

26. The role of top management in this process is essential. However, top management needs support from the subject-matter statisticians. The project should focus on statistical tasks; IT serves as a major instrument which the project has at its disposal. Possibilities to improve statistical co-ordination should be noticed and actively exploited, e.g. by means of the global metadata component of the data warehouse.