1650

UNITED NATIONS

ECONOMIC AND SOCIAL COUNCIL

Distr.
LIMITED

E/ECA/PSD.4/44
2 January 1986

ENGLISH
~~Original: FRENCH~~

ECONOMIC COMMISSION FOR AFRICA

Fourth session of the Joint Conference of African Planners, Statisticians and Demographers

Addis Ababa, 3-12 March 1986

SELECTED ISSUES IN THE DEVELOPMENT OF STATISTICAL DATA BASES

C O N T E N T S

I. INTRODUCTION

1. It was in December 1982 that the Statistics Division of ECA embarked on a study of the establishment of a regional statistical data bank. One year later, a document entitled "Report on the Development of ECA's Statistical Data Base", describing the conceptual organization of the data base, was submitted to the third session of the Joint Conference of African Planners, Statisticians and Demographers, held at Addis Ababa from 5 to 14 March 1984. The years 1984 and 1985 were devoted to installing under Image 3000 the necessary basic upstream (establishment, updating, and retrieval of data from certain production files) and downstream (search/selection/tabulation) facilities. The first part of this paper deals with the assessment of these facilities and gives some indications concerning future activities.

2. During the discussion of the first report mentioned above, the Joint Conference expressed the view that ECA should associate the African countries more closely with the implementation of the ECA data-processing project, called PADIS-STAT, and in particular, help them set up their own statistical data bases. Accordingly, the second part of this document deals with some methodological aspects relative to the conceptual basis of the establishment of national statistical data bases. This second part of the paper addresses the issue of conceptual organization and installation procedures.

II. THE PADIS-STAT SYSTEM

3. The PADIS-STAT system was established to:

(1) Provide ECA statisticians and economists with:

- direct speedy access to African statistical data;
- data processing and analysis services;

(2) Automate the publications of ECA's statistical information system,

(3) Assist African countries planning to set up statistical data bases.

4. PADIS-STAT has three levels, of which levels I and III may be considered as sub-banks structurally linked with the master bank which is level II. Level I is designed to contain master data which can be used for the automatic production of country profiles. Level II will contain the bulk of the statistical data available over a long period. Level III, the only one presently in place, is geared towards dissemination and analysis.

(a) Structure and content of Level III

5. The data base comprising Level III covers the following sectors: population and employment, national accounts, agriculture, industry, transport and communications, finance, prices, foreign trade, education and health statistics. At present, it contains more than 100,000 time series covering the period 1970-1985. This data base is installed on PADIS' Hewlett Packard HP 3000/III mini-computer, managed by DBMS Image 3000, with programmes written in Cobol by the computer expert of the ECA Statistics Division. These programmes are used for real-time and batch mode processing. They are accessed interactively according to a main menu.

6. Structurally, Level III is made up to 9 sub-files or tables linked by pointers. The data files are:

- Time series
- Status of series
- Origin of series
- Mode of observation
- Reporting country code
- Partner country code
- Unit of observation
- Producer identifiers
- Password

7. Access to the data is locked by a security system open only to authorized persons.

(b) Principal uses of Level III

8. Level III is now operational. It can be used for automatic generation of the tables for the Statistical Yearbook for Africa, those on socio-economic indicators and certain tables for the African Foreign Trade publications (Series A and C). At the moment, the types of calculation possible concern averages, percentages, ratios and trend indices. It is expected to broaden the scope of activities by establishing procedures for statistical and econometric analysis. To avoid the system becoming unwieldy to manage, however, the latter calculations will not be integrated into it. They will have to use general interfaced software such as SPSS.

(c) Principal constraints to be overcome

9. As can be seen from the above, the basic Level III management operations have been completed. Currently, using the various modules that have been installed, the bank can be interrogated directly, series can be selected and the results published. However, other significant tasks remain to be performed. Firstly, a satisfactory solution is required to the problem of updating the base. The countries of the region are PADIS-STAT's major source of information, since it relies basically on the statistics production system of ECA, which in turn, is totally dependent on the national statistical systems. But however, ECA faces the following difficulties in compiling statistical information from African countries:

- Undue delays in the publication of statistica data and their transmission between the countries and ECA;
- Long delays in data entry arising from the fact that the information from the countries is stored almost exclusively on paper.

10. In order to ensure a measure of quality and usefulness of the data bank, these difficulties must be solved. ECA is seeking to make as much use as possible of all available data sources - national publications, international sources, and information gathered in the course of missions. The effective solution, however, would be for the African States to record their statistical information on magnetic media for ease of reproduction and transmission. They should therefore integrate data processing techniques more systematically in their overall statistics production process.

11. The second problem requiring urgent resolution is that of documentation The data fed into the bank will have to be documented before it can be correctly interpreted and used. This means that the bank should contain a data base on the information it contains which can be interrogated on-line. Consequently, the following should be provided:

(i) Data dictionary;

(ii) Documentation for non-computer users (language and query procedure);

(iii) Acquisition documentation (acquisition formats, acquisition procedures and check-listed);

(iv) Core documentation (conduct of basic operations, safety procedures, definition language and data manipulation).

12. A third problem is the dissemination of data by the bank. The foremost attribute of a data bank is the capacity to store and transmit to the user a considerably larger volume of data than was previously published. This presumes the existence of a data transmission medium to convey data to the user. This can be provided through telecommunications (telephone, radio link or satellites). Here data dissemination and transmission become one and the same thing. However, it appears unrealistic at the moment to adopt this mode of dissemination for the obvious reasons of the considerable volume of data and the high cost of transmission. On the other hand, the dissemination of data from the base on magnetic (diskettes, tapes) or other media (microfilms and microfiche) should be considered and established along with the dissemination of information stored on paper media.

13. Finally, considerable improvements in access to the data base are required so that final users should as much as possible be relieved of programming constraints and the mastery of data processing skills will not be a pre-requisite. To this end, the year 1986 will be devoted to the installation of a menu-oriented interrogation system which will allow users without data-processing knowledge to manipulate data through simple commands written in English and French.

III. ESTABLISHMENT OF NATIONAL STATISTICAL DATA BASES

14. One of the objectives of PADIS-STAT is to assist African countries that are planning to set up their statistical data bases. In view of the fact that the installation and development of a statistical data base is much more dependent on organization than on data-processing, special attention should be given to the conceptual phase. It is on the basis of the organizational structure decided upon that the data-processing equipment will be installed. Certain conceptual elements of the establishment of national statistical data bases are described in this chapter.

(a) Conceptual phase

15. A statistical data base should be designed as a normal extension of the traditional production, analysis, forecasting and dissemination activities of a national statistical service. In other words, the data base must serve both as a working tool for statisticians and as a user-oriented data processing system. The conception of such a base should be defined in specifications which will include the following:

- Determination of users and their needs;
- Determination of the data to be handled. A data dictionary needs to be compiled based on the uses foreseen. Two classes of data will be considered, namely those to be stored, and those to be derived through calculation. Among the data to be stored, a distinction should be made between those to be recorded on disc because they are more frequently required, and those for which magnetic tapes will suffice as the storage medium;
- Definition, in broad outline, of use scenarios and updating procedures;
- Determination of the data-processing equipment needed in the light of the bank's technical requirements (storage volume, number of movements to be processed, volumes printed and volumes to be acquired);
- Preparation of a budget for the establishment and operation of a data bank (including the costs of data-processing equipment and software).

(b) Logical design phase

16. A national statistical system undertakes four major functions:

- Collection and processing of census or survey data or information extracted from administrative records;
- Data analysis and interpretation;
- Data dissemination;
- The Metabase, consisting of information on the information.

17. A statistical data base affords possibilities for carrying out these functions. It is made up of a set of data-processing services using data files organized in such a way as to allow the storage and retrieval of all kinds of information for publications and other needs. Such data can be individual data (primary product of census or survey) or aggregate data (time series or tables). In the present African context, however, the highest priority of a national statistical base should be the compilation of aggregate data produced or centralized by a national statistics office. It should, moreover, at least in the initial stages, act simply as a file manager. Hence the integrated services available to users will be confined to:

- Search in the proper sense of the word, making it possible to isolate a sub-set of a series;
- Selection which makes it possible to choose, from sub-sets assembled by means of search, the series as well as the characteristics and the periods on which work is desired;
- Transfer of the results of selection, where necessary, to a private work area;
- Simple statistical calculation;
- Submission of the results (reports, tables, graphs)

18. All functions that are not integrated into the bank, such as modelling, should be handled by means of outside software, specific or general.

1. Structure of data

19. The major services to be provided for in an aggregate statistical data base are:

- Time series;
- Observations;
- Nomenclatures;
- Linkage tables;
- Tables;
- Standard tables;
- Menu.

(a) Time series

20. The principal object of a macro-economic data base is the time series. Its structure should include the following elements:

- Statistical source of the series (census, survey, administrative files);
- Country/region/province concerned;
- Regularity (code indicating if the series is regular or not);
- Periodicity of the series;
- Time unit covered by observations;
- Flow or stock series;
- Type of values of the series (physical unit, monetary unit, etc.);
- Mode of value representation - absolute, index, ratio, etc.
- Type of correction (raw, correction for seasonal variations);
- Taxation system (duty-free, all taxes included);

- Observation values;
- Origin of the series;
- Precision of unit;
- Sign of series (possibility of observations being positive or negative);
- Power to which the unit is raised;
- Data base identification of series;
- Title of series;
- Base of a series of indices;
- Comments, if any, on series;
- Date of first observation;
- Date of last available observation;
- Identifier of series producer;
- Identifier of series file;
- Protection of series.

21. Nearly all these items of information will be in coded form. The others will be represented by short texts or integers.

(b) Observations

22. An observation must be represented either by a numerical value or by a conventional symbol. It may be accompanied by a comment. The following information should accompany an observation:

- Value: numerical representation of the value of an observation; it may be positive or negative; the position of the decimal point if any will be defined in the series;
- Status of observation: Every observation will have a status in terms of its stage of statistical preparation: definitive, provisional, semi-difinitive, revised, estimated;
- Confidentiality: degree of confidentiality of an observation;
- Comments, if any, attached to an observation. The non-numerical representations of an observation will be made using coded elements taken from a table (for instance: insignificant, not calculated, not available, etc.)

(c) Nomenclatures

23. All nomenclatures which serve as an element in the definition of the series should be identified, characterized and managed by the base. Each nomenclature will have to be characterized by:

- Nomenclature item code;
- Item title (short text);
- Comments, if any, on a nomenclature, which will consist of text;
- Identifier of the nomenclature expressed in code from a nomenclature table;
- Field of application of the nomenclature, expressed in code;
- Status of nomenclature (official, specific, management nomenclature) expressed in code.

(d) Linkage tables

24. The linkage tables translate relationships between two nomenclatures that are considered to be respectively input and output nomenclatures.

(e) Tables

25. Tables are a significant element in the management of a statistical data base. They make it possible, on the one hand, to identify categories of entities, and on the other hand, to catalogue the entities within their categories. The tables are indexed with the assistance of a Meta table which guides the user in his enquiries.

(f) Menu

26. The purpose of the menu is to offer a logical approach to the data base to guide the user step by step in selecting series groupings and lastly to designate the series of his choice. The following elements should be used in structuring the menu:

- Domain or node of menu (set of data corresponding to a domain or series grouping);
- Identifier of a domain series grouping;
- Title of the domain or grouping;
- Number of series attached to a domain or to a series grouping;
- Domain protection code;
- Domain password;
- Comments, if any, on the domain.

2. Computerization of production files

27. A statistical information system comprises two groups of activities which are different but closely related:

- The first group can be considered as a source-oriented sub-system. Its function are broadly the compilation, initial processing and storage of data corrected according to their primary structure. These are individual data items, also known as "primary data".
- The main purpose of the second group, the user-oriented sub-system, is to organize and store primary data in the form of files called production files geared towards the presentation of data, most often in the form of publications with a standardized content. The first problem in establishing a statistical data bank is to transform the production files into data bases. The files take different forms not only in physical terms (manual or computerized), but also in terms of logical structure.

28. If a file is computerized, the equipment on which it is prepared should be described along with the access mode, the method of recording, and the location of the data on the recording. If it is on paper, a clear description should be given of how the information is to be read.

29. The logical structure of the files should be standardized by applying the following guidelines:

- Determination of the sets of entities (series, observations, nomenclatures, tables, etc.)
- Determination of the associations between the sets of entities (an association is a binary relationship between two sets of entities);
- Determination of the integrity conditions which cannot be deduced directly from the definition of the sets;
- Establishment of a conceptual diagram of the data (a simplified representation of the sets of overall entities and associations);
- Definition of data handling procedures. In addition, each production file should be fully documented on the various subjects of the file. This documentation must be arranged in chapters corresponding as far as possible to the various domains of the menu.

IV. CONCLUSION

30. An attempt has been made in this document to give, first, a brief assessment of the existing facilities and the tasks remaining to be performed with regard to the Statistical Data Base of ECA.

31. In this connection, the future measures should permit a gradual transition from a data base that is regarded as a working tool for ECA statisticians to a data base that is accessible to a larger group of users, especially member States. To facilitate this transformation, an African network of statistical data bases needs to be established with PADIS-STAT as the central element. The installation of such a network must be preceded by the establishment of national statistical data bases. Accordingly, this document has also attempted to define a kind of strategy for the formulation of a user-oriented national statistical information system. This strategy rests on a basic principle, namely, that a national statistical data base should serve both as a working tool for statisticians and as an instrument for disseminating statistical information.