UNITED NATIONS
ECONOMIC
AND
SOCIAL COUNCIL

Distr.
GENERAL

E/CN.3/555
15 July 1980

ORIGINAL: ENGLISH

STATISTICAL COMMISSION
Twenty-first session
~~17-26 February~~ 1981
Item 9 (b) of the provisional agenda*

TECHNICAL CO-OPERATION: STATISTICAL DATA PROCESSING

Progress report on statistical data processing

Report of the Secretary-General

SUMMARY

The present document sets out a description of the United Nations technical co-operation activities in data processing which are executed by the Department of Technical Co-operation for Development and technically supported by the United Nations Statistical Office and the statistical divisions of the regional commissions. Paragraphs 3-15 provide a general description of the coverage of the activities. Paragraphs 16-54 provide a description of the United Nations Statistical Office programme for the development of computer software for assisting developing countries in processing their population censuses and surveys data. Paragraphs 55-91 provide a description of present-day data-processing hardware useful to developing countries. Points for discussion by the Commission are included (para. 92).

* E/CN.3/535.

CONTENTS

/...

INTRODUCTION

1. The present document has been prepared in response to the request of the Statistical Commission at its twentieth session. [1] It covers technical co-operation in statistical data processing provided to developing countries by the United Nations, executed by the Department of Technical Co-operation for Development, and substantively supported by the United Nations Statistical Office in collaboration with the regional commissions.

2. The document is divided into three sections. Section I is a general description of technical co-operation in data processing, its coverage and total value. Section II describes the development and dissemination of computer software for population censuses and survey processing. Section III is a review of the development trends in computer hardware and the impact of these trends on the United Nations technical co-operation programme for statistical data processing.

[1] *Official Records of the Economic and Social Council, 1979, Supplement No. 3* (E/1979/23), para. 186 (d).

I. CURRENT TECHNICAL CO-OPERATION ACTIVITIES IN DATA PROCESSING

3. The volume of United Nations technical co-operation activities in data processing has been steadily increasing over the past 10 years. This has largely been owing to the fact that many developing countries have concluded their first complete population census count during the decade and have had to acquire the computer technology for processing the data, thus establishing this necessary capability.

4. Both the United Nations Development Programme (UNDP) and the United Nations Fund for Population Activities (UNFPA) have provided funds for establishing and improving the computer capacities of developing countries for processing their statistical data.

5. While the nature and mix of the various components of such projects differed from country to country, most financial aid was directed towards equipment procurement, computer rentals, expert services and training.

6. In addition, it has now become evident that the usefulness of computers in developing countries depends, to a large extent, on the availability of software to enable countries to process their data. The United Nations Fund for Population Activities has, therefore, provided funds to the Department of Technical Co-operation for Development for the development of suitable editing and data-processing software programme packages. This development is proceeding, and section II below provides a description of the progress.

7. Equipment procured in support of technical co-operation projects includes data entry equipment, integrated computer systems, electrical generation and conditioning equipment, maintenance tools and other auxiliary equipment. Related activities also often include analyses of requirements for physical installation including space and layout, air-conditioning and power supply and control systems. Section III below describes current developments in data-processing hardware of relevance to the technical co-operation activities in statistics.

8. Data-processing experts are recruited and placed in country posts for either short-term or long-term assignments depending on the type of assistance the country requires. The process of recruiting experts consists of evaluating potential candidates, assessing their strengths, interviewing them or establishing and evaluating their interviews by others and matching their skills and availability to particular country posts. Information on a panel of candidates is then forwarded to the country, which makes the final selection.

9. Once the experts are in post, they provide periodic reports which are analysed at United Nations Headquarters and the relevant regional offices. Issues and problems raised in such reports frequently lead to an interchange of technical correspondence between Headquarters anad the expert and, in some cases, leads to a mission from the United Nations or regional headquarters.

/...

10. A training component is included in most country computer technical co-operation projects in order to promote the development of sufficient human capability within the country for continuing the activities in this field in an independent and self-reliant manner. The activity covers the evaluation and placement of candidates for training, planning and execution of training programmes and monitoring the post-training performance on the job.

11. The United Nations Statistical Office, through the Department of Technical Co-operation for Development, has been actively involved over several years in assisting developing countries to establish, upgrade and improve their computer capacity and capability. The total annual expenditure on technical co-operation activities in data processing has increased from $2,624,000 in 1976 to an estimated $7,396,000 in 1980 (see table 1 below). The number of experts assigned to country projects has been between 26 and 33 over the period 1976-1980 and the annual cost of this component has almost doubled from 1976 to 1980.

12. An interregional adviser in computer methods, attached to the United Nations Statistical Office, executes a continual round of short-term consulting missions on uses of electronic data-processing equipment, computing techniques and computer-based systems in statistics. In addition, there are two full-time technical advisers in computer methods and data processing in the Office who provide substantive technical support for country projects. They also undertake country missions to plan and help solve implementation problems. There are also three regional advisers in data processing, located in three regions, who provide substantive technical support to countries in the region. They complement the work of the technical advisers at Headquarters.

13. There also has been a steady increase in the provision of equipment (computer hardware and software, as well as rental costs for hardware) between 1976 and 1980. In 1976, the total expenditure was $1,248,000 and the estimated cost in 1980 is $5,000,000. This expenditure component is expected to increase very significantly in 1981, to over $13 million. This is owing to the fact that UNFPA is providing over $10 million towards the procurement of computers for processing the 1981 population census of China.

14. The annual cost of training local staff, who provide technical continuity after a project has ended, has also steadily increased over the period, from $177,000 in 1976 to $340,000 in 1980.

15. Finally, the United Nations Statistical Office continues to provide substantive technical support to an increasing number of projects in data-processing activities. The total number of projects has ranged between 60 and 75 over the past five years, and it is expected that, although the total number may not substantially increase over the next three to four years, the total estimated cost of these projects is expected to increase substantially as countries participate in the 1980 round of population and housing censuses.

Table 1. Expenditures of the United Nations, including the regional commissions, on technical co-operation in statistical data processing

(cost in thousands of United States dollars)

Activities	1976	1977	1978	1979	1980*
Programme formulation, support and evaluation					
Number of work-months	36	36	49	49	62
Cost	144	144	245	245	310
Regional and interregional advisory staff					
Number of posts	5	5	4	4	4
Cost	240	240	240	240	240
Country experts					
Number of posts	33	26	26	32	32
Cost	813	545	748	1 161	1 500
Equipment (hardware, software and rental)					
Cost	1 248	1 885	1 340	3 060	5 000
Training (fellowships)					
Number	48	42	21	51	50
Cost	177	146	208	333	340
Miscellaneous cost	2	2	5	6	6
Total cost	2 624	2 962	2 786	5 045	7 396

Source: Department of Technical Co-operation for Development, Report 002 for 1979.

* Estimated.

/...

II. DEVELOPMENT OF COMPUTER SOFTWARE FOR POPULATION CENSUSES AND SURVEYS

16. The development of computer software packages by the Statistical Office is being assisted through the provision of UNFPA funds to the Department of Technical Co-operation for Development. This development is aimed at (a) supplementing already available computer software packages with easy-to-learn-and-use packages that can be installed on relatively small and inexpensive computers, and (b) developing new capabilities that are aimed at supporting important component activities of population censuses and surveys and the use of the resulting data.

17. A small computer has been installed at United Nations Headquarters to ensure that the software developed could be easily used on similar computers in developing countries as well as to provide support to countries in developing and modifying computer software programmes when required and for test demonstration and training of visiting staff from developing countries.

18. The Statistical Office has developed and distributed the XTALLY package for producing cross-tabulations, using the types and capacities of computer facilities that are available in most developing countries. A similar package for supporting computer editing and correction of data has been developed by the Statistical Office and is called UNEDIT.

19. The present and planned activities of the Statistical Office focus on (a) the delivery of available software - UNEDIT and XTALLY - to countries wanting it, (b) the enhancement of capabilities of these two packages to better satisfy the needs of users, (c) the development and delivery of a new package called LOGMAP (Logistics Management and Planning) intended to support both the collection and use of resulting data, and (d) the establishment of long-term arrangements for demonstration and training in the use of the software at appropriate regional or national institutions.

20. To complement the development of computer software packages, UNFPA has provided funds to the Department of Technical Co-operation for Development for training staff from developing countries in the use of the computer hardware and software for processing their own data to produce results they may not otherwise obtain on time.

A. Description of XTALLY

1. Capabilities

21. XTALLY can produce multi-dimensional cross-tabulations summing one or two variables or counting records and giving subtotals, percentages, ratios, means, differences, sums or inflated totals at all levels of the hierarchical tabulations. The main features of the XTALLY software package are:

(a) No programming, no compiling, no sorting;

(b) Automatic titles, automatic totals, automatic subtotals, automatic percentages;

(c) Up to 99,999 cells per table, 1 or 2 sums per cell;

(d) Up to seven dimensions per table, up to 126 categories per dimension;

(e) Alphanumeric variables, full character set acceptable in data records;

(f) Up to 15 columns per page, any number of rows, automatic extra pages if over 15 columns;

(g) Data-item and category-set names and definitions automatically filed and need not be repeated when specifying tables;

(h) Automatic generation of A+B, A-B, A/B, A*B, B-A, or B/A for tables summing A and B;

(i) Automatic generation of percentages of total or subtotals at any level and several sets of percentages may be produced for the same table;

(j) Completed table stored on disk:

(i) extra copies immediately available;

(ii) table saved off-line for later use or hierarchical addition;

(k) Tables are defined in seconds or minutes using pre-defined names of data and category sets and

(l) Table production time is a linear function of data file size.

The requirements of the XTALLY package are:

(a) 32K byte primary storage capacity, 4 megabyte disk storage capacity and

(b) RPG-2 or advanced RPG compiler available for type of computer to be used.

2. Ease of learning and using

22. XTALLY can be learned and used by non-programmers having some aptitude for computer use or by other staff at the level of statistical clerk or higher. A few hours of explanation, demonstration and trial use are sufficient for beginning to use XTALLY and, after a few days or weeks of practice, the new user can become quite familiar with the capabilities and limitations of the system. XTALLY has been installed by mail on several occasions and learned entirely from documentation, but on-site explanation and demonstration are highly desirable.

23. There are two major steps in using XTALLY:

(a) Definition of data record variables and their location in the record

/...

and definition of the sets of categories within which sums or frequency counts are to be produced. Performance of this step produces a dictionary containing definitions of variables and category sets which then may be used for producing a large number of tabulations. The dictionary is stored on the XTALLY magnetic disk and is referenced automatically when the XTALLY system interprets table or function specifications;

(b) Specification and production of summary cross-tabulations and certain optional functions, including percentages. Performance of this step is very simple, since only the names of variables and category sets need be stated.

24. Definition of variables and category sets is done using the following statement forms:

(a) For each variable:

(i) name (3 characters);

(ii) start location in source record (2 positions);

(iii) end location in record (2 positions);

(b) For each category of each category set (in ascending order of upper-limit values):

(i) name of category set (3-character variable name + 2 additional characters);

(ii) upper-limit value for this category (1 to 4 characters, depending upon variable length);

(iii) relative print position (top to bottom or left to right) for this category;

(iv) title to be printed for this category (4 characters).

25. Preparation of the statements required for variable and category - set definition is facilitated by interactive procedures with systems having CRT (cathode ray tube) facilities. Otherwise, the statements require one record each using cards or diskettes.

26. Various sets of definitions may be stored on the disk for a variety of record types. When it is time to produce a table, the correct set of definitions is entered into the XTALLY system with a simple procedure, for example, DEDATA CENSUS where DEDATA is the procedure name and CENSUS is the name of the definition file.

27. Cross-tabulations and functions are specified using the names of variables and category sets. Thus, a table is specified with a single statement containing:

(a) Column category sets (names of 1 to 3 category sets, in hierarchal order);

/...

(b) Row category sets (names of 0 to 4 category sets, in hierarchical order);

(c) Names of one or two variables whose values are to be summed (blank means record-count).

A second statement of 96 characters provides the title for the table.

B. Description of UNEDIT

1. Capabilities

28. The UNEDIT system is a portable generalized software package, developed to meet common editing needs in census and survey data processing. The UNEDIT package is written in RPG-II language and can be run on small computers with a minimum internal memory of 32-K bytes. RPG-II is used for programing in order to ensure maximum portability among minicomputers and small computers, for which RPG-II is becoming a very popular programming language.

29. The UNEDIT examines the user's specifications and, if any syntactic errors are found, diagnostic messages are printed. At the execution time, UNEDIT performs the editing operations according to the rules generated from the user's specifications. During the course of execution, error statistics by type of error and by data field name are printed. It contains, at present, no capability for automatic imputation.

30. Although UNEDIT is designed primarily for editing population census and survey data, it can be used for editing data from other statistical surveys as well. UNEDIT is able to process a file consisting of a variety of record types while performing quantitative and qualitative editing. Also, because of the ease with which editing specifications can be entered, UNEDIT can be used for testing and modifying the editing rules during the various stages of data processing from the pilot survey to the final decision on editing rules.

2. Flexibility in input data file structure

31. The UNEDIT can process an input file consisting of a variety of record types. The user specifies in "mode" parameter of form-1 whether the input file is a single-record-type or a multi-record-type file. If it is a multi-record-type file, the user then specifies whether inter-record checking is required. There are four types of file processing:

Mode 1: single-record-type file;

Mode 2: multi-record-type file (inter-record checking not required);

Mode 3: multi-record-type file (inter-record checking required);

Mode 4: household-questionaire-type file (two-record-type file) where one record for household (or housing unit) is followed by one or more person records.

3. Error detection (validity, inconsistency and structural checks)

32. The validity check is a test to determine whether a code or value in the data field is valid or within acceptable range by comparing it with a list of valid codes or values provided by the user. The user simply lists valid codes or acceptable ranges of values by data field name in form-3 (Valid Code Specification Form). When an invalid code or value is detected, an error message is printed and an invalid code indicator (assigned by the user) will be encoded into the erroneous data field. The user may specify either valid codes individually or as an acceptable range of values (lower and upper limits of range).

33. The inconsistency check is a test to determine whether a string of codes or values in two or more data fields are consistent when taken in combination (even though each by itself may be valid and acceptable). The user specifies the inconsistency conditions on form-5 (Inconsistency Check Specification Form). The inconsistency checks are as follows:

(a) Intra-record inconsistency check, that is, a test of data fields within the same record. The UNEDIT performs the check in every "mode";

(b) Inter-record inconsistency check, that is, a test of data fields in more than one record. The UNEDIT performs the check when "mode" of file processing is 3 or 4, that is, when a multi-record-type file or a household-questionaire-type file is being processed;

(c) Quantitative and Qualitative Checks: Data fields tested in the inconsistency check can be quantitative or qualitative;

(d) Arithmetic calculation and comparison of data fields.

In the inconsistency check, sometimes it is necessary to calculate on the field values or to compare two field values in order to determine the inconsistency condition. The UNEDIT provides these capabilities in form-4 (Pre-edit Calculation Form).

34. The structural check is a test to see whether there is any record missing from the file. The structural checks are as follows:

(a) Input count by area: The UNEDIT automatically prints the number of input records for the user designated area, such as enumeration districts;

(b) Detection of missing record-type record: When a multi-record-type file is processed, any missing record-type record will be automatically detected and a message will be printed;

(c) Count of person-records in mode 4 file: In the mode 4 file where a household-questionnaire-type file is processed, the number of person-records within a household is automatically counted and made available to the user for testing in form-5 (Inconsistency Check Specification Form).

4. Automatic assignment of code

35. The user of UNEDIT can assign a particular code or value to the particular data field when a set of conditions specified in form-5 is met. This can be done by entering on form-5 the name of the data field to which a new code is to be assigned. The blank field or the invalid code field, detected in the valid code check and given an error indicator, can also be given a new code at this stage.

5. Output options

36. The user can select the following options for output:

(a) error messages and error statistics only or

(b) (a) plus the corrected record output file.

6. Diagnostic messages on user's specifications

37. The UNEDIT package will check the user's specifications in detail and print out diagnostic messages on the syntax errors, as well as the redundancies and contradictions in the inconsistency condition sets.

7. Portability

38. The UNEDIT package consists of two modules, one for pre-edit preparatory operation and the other for the execution of editing. Both are written in RPG-II language and are completely machine independent. They can be stored on such portable storage media as diskettes, disks, magnetic tapes or cards.

8. Ease of learning and using

39. The users of the UNEDIT are not expected to be familiar with computer programming techniques. UNEDIT is designed to be used by a statistician or a subject-matter specialist knowledgeable of the details of the questionnaire, data-coding rules and final census or survey tables to be processed.

40. In designing the UNEDIT system, special emphasis was placed on developing a system which is suitable for small computers and whose capabilities and required specifications are easy for users to learn. Users are merely required to fill out pre-format editing specification forms; they are not required to write computer-language-like statements. The forms to be filled out are:

Form 1: File Description Specification Form,

Form 2: Input Record Specification Form,

Form 3: Valid Code Specification Form,

Form 4: Pre-edit Calculation Form, and

Form 5: Inconsistency Check Specification Form.

/...

41. The specifications for the editing rules, such as the validity check and the inconsistency check, are written in the specification form in such a manner that statisticians write the editing rules in their own words to the programming staff rather than in the computer language to the computer.

C. Description of LOGMAP

42. A small system called LOGMAP (Logistics Management and Planning) was designed to provide computer support for the following activities common to planning and managing censuses and surveys or to organizing and using the results:

(a) Defining administrative and geographical entities and their composition and hierarchical relationships;

(b) Cataloguing administrative and geographical entities, with their important attributes;

(c) Cataloguing maps and other materials needed for planning, conducting operations and for interpreting and using census and survey results; and

(d) Enriching the summary cross-tabulations or other analyses of census and survey data by adding administrative and geographical data.

43. In operation, LOGMAP allows the user to define the types of aggregate entities comprising the universe to be studied and to specify the important attributes of each entity type. Then, in conversational fashion, the user names each entity of each type and the LOGMAP system compiles a catalogue of entity names by type. When the catalogue is complete, LOGMAP generates a questionnaire for each individual entity, which may be completed at appropriate levels of administration to supply further detailed input to the LOGMAP system. The resulting data base can then support further administrative and planning activities by providing status reports and a variety of important reference information, limited only by the information the user has considered important and has chosen to collect. In particular, the system will provide such computer-stored cross-references as maps for particular localities.

44. For organizing and using the data collected in the census or survey, LOGMAP will include locators for micro-data or summary files to facilitate the production of cross-tabulations and other analyses and will make possible the production of tabulations and analyses relating micro-data to the attributes of the administrative or geographical entities to which they belong.

45. The basic computer program~~me~~s of the LOGMAP system are under development. A draft version of the system is expected to be available early in 1981.

D. Delivery

46. The computer software programme has delivered UNEDIT or XTALLY software to a

/...

number of countries and has outstanding requests from many others. In some cases the software has been installed by local staff following written directions, but in most cases the software has been installed and demonstrated either by staff of the Statistical Office or by a United Nations regional data processing adviser familiar with the details of the programmes and computer-operating systems. Participation of regional advisory staff has been productive for everyone involved, and it is hoped that such participation will be strengthened.

47. Table 2 shows countries to which the software packages were delivered, and table 3 shows the outstanding requests for software packages by country.

Table 2. Delivery of UNEDIT and XTALLY software packages by country

Country	Country	UNEDIT date	XTALLY date	Place of installation and use
IBM S/34	Afghanistan	1979	1979	Various statistical applications
	United Arab Emirates	1980	1980	Central Statistical Dept. for various statistical applications
IBM S/32	Antigua	1979	1979	Place and use not known
	Democratic Yemen	1978	1978/79	Central Board of Statistics for edit and tabulation of census data
	Mauritania	1978	1978/79	Dept. of Statistics for edit and tabulation of census data
	Thailand	1978	1978	ESCAP for various statistical data
	United Republic of Cameroon	1978	1977/79	Bureau Central du Recensement for census and survey edit and tabulation
IBM S/3 (10)	American Samoa		1976	American Samoa Computer Centre for various statistical applications
(10)	Burundi	1978	1976/79	Centre National d'Informatique
(10)	Ghana	1978	1976/79	Volta River Authority for various statistical applications
(15)	Guinea	1977	1977	Pool Comptable National for various statistical applications
(10)	Iran	1978	1978	College of Statistics
(10)	Liberia	1978	1976/79	Ministry of Planning and Economic Affairs

/...

Table 2 (continued)

Computer	Country	UNEDIT date	XTALLY date	Place of installation and use
IBM S/3 (continued)				
(10)	Mauritania	1978	1978	Dept. of Statistics
(10)	Sierra Leone		1979	Place not known; tabulation of census and survey data
IBM S/360 (125)	Dominican Republic	1979	1979	Departamento de Computos
(115)	Oman	1980	1980	Directorate General of Finance
(125)	Sri Lanka	1979	1979	Dept. of Census and Statistics
(135)	Zambia	1980	1980	Central Statistical Office
HB 62/60	Benin	1980	1980	Bureau Central du Recensement for edit and tabulation of sample data
	Tunisia	1980	1980	Centre National de l'Informatique
NCR (201)	Jordan	1978	1978	Royal Scientific Society
CDC CYBER 18-20	Thailand	1979		Mekong Committee

Table 3. Outstanding requests for software packages by country

Computer	Country	UNEDIT	XTALLY	Remarks
IBM S/34	Bangladesh	x	x	
	Gambia	x	x	when computer installed
	Pakistan	x	x	also has S/360-30
IBM S/3 (10)	American Samoa	x	x	
	Ecuador	x	x	cannot install; only 16K

/...

Table 3 (continued)

Computer	Country	UNEDIT	XTALLY	Remarks
IBM S/3 (continued)				
(12)	Samoa	x	x	
IBM S/360 (30)	Ghana	x	x	at the Central Bureau of Statistics
(30)	Philippines	x	x	
(25)	Poland	x	x	
(30)	Sudan	x	x	
IBM S/370 (168)	Israel	x	x	no RPG-2
(125)	Mauritania	x	x	
(135)	Pacific Islands (US Trust Territory)	x	x	
(135)	Philippines	x	x	
(125)	Togo	x	x	
HB 66/10	Iraq	x	x	no RPG-2
62/60	Mali	x	x	
66/20	Niger	x	x	
NCR (200)	Iraq	x	x	
(8250)	Rwanda	x	x	cannot install; version too slow
(101)	Somalia	x	x	
ICL (1902T)	Botswana	x	x	model obsolete and no RPG-2
(2904)	Fiji	x	x	
(2970)	Hong Kong	x	x	no RPG-2
(2903)	Lesotho	x	x	
(1904A)	Malaysia	x	x	model obsolete and no RPG-2
(2903)	Poland	x		
(2903)	Swaziland	x	x	

/...

Table 3 (continued)

Computer	Country	UNEDIT	XTALLY	Remarks
DEC-10	Bolivia	x	x	
Wang 2200VS	Peru	x	x	
Hewlett-Packard	United Kingdom/ World Fertility Survey	x	x	

E. Use and training at United Nations Headquarters

48. The basic activities of the programme are the development and delivery of software to population data-processing projects in developing countries. Other activities that support the basic objectives or take advantage of the project's resources to support related important objectives are described below.

49. The small computer installed at Headquarters, together with the UNEDIT and XTALLY software, comprise a very effective tool for "hands-on" practical training of national staff whose country may have pilot census or survey data needing processing as soon as possible but not having the means to do it. It has met such needs for several countries, including Afghanistan, Democratic Yemen and Gabon. In other cases, such as Haiti and Swaziland, the Headquarters staff have processed data without the participation of national staff. A new scheme, begun in 1980, provides funds that will enable national staff to visit Headquarters or other suitable sites for practical "hands-on" training, strengthens the delivery activity of the main software project and helps ensure that the software itself is well-tested and suited to the needs of developing countries.

50. The software and hardware for processing data of developing countries and for training national staff has thus far been exclusively at Headquarters, but the programme aims to establish collaborative arrangements with regional or national institutions in order to provide as strong and as familiar a base as possible in terms of language, working-space and technical/professional support. Steps in this direction include a recent training workshop at the Centre National de l'Informatique, Tunisia, for participants from French-speaking African countries using Honeywell-Bull equipment. It is hoped that similar workshops can be conducted at selected institutions in other regions and that long-term collaboration can be developed.

F. Other plans and activities

51. The programme aims to inform technical co-operation personnel and organizations in developing countries, both by direct correspondence and by

/...

publication of results through technical and professional channels, what software can be supplied. Also, a number of national staff or experts from developing countries visit Headquarters in the course of their work or training programmes, and demonstrations and explanations of the software packages are frequently given.

52. The software systems developed are suitable for use on many different computers, but the idiosyncrasies of each make and model of computer require preparation of a number of versions of each system, each version having distinct control procedures and device-specific elements for the specific machine. Consultants are employed, as required, for developing these versions.

53. There are plans for the establishment, if possible, of long-term arrangements with regional or national computer-centre institutions which would give demonstrations and training in the use of the software packages, thus providing the software in the language of the country or region and eliminating dependence upon Headquarters for demonstration and training. Towards this end, the feasibility of relying on regional or national institutions for demonstration and training is being tested as opportunities arise.

54. The staff attempts to stay abreast of important software packages available from other international organizations, which can be obtained by developing countries through bilateral assistance or other means and to maintain contact with national offices that have produced useful software and make it available to developing countries. Use of the computer has been made available to the International Statistical Programs Center of the United States Bureau of the Census, for example, to enable the Center to prepare an IBM S/34 version of its COCENTS package, for example. Users of software have been encouraged to become familiar with software tools available from such other organizations as the National Central Bureau of Statistics of Sweden, the Bureau of Labor Statistics of the United States and the Institut national de la Statistique et des Etudes économiques of France.

/...

III. CURRENT DEVELOPMENTS IN DATA-PROCESSING HARDWARE

55. Data-processing hardware has provided an indispensable foundation for statistical data processing for almost 100 years. In this context, the term "data-processing hardware" includes devices based upon mechanical, electromechanical or electronic technology which manipulate numbers in digital form. Modern data-processing devices are often referred to as automatic data-processing equipment, which means that the equipment is capable of interpreting and executing a pre-defined sequence of instructions.

56. The impetus for one of the two major thrusts in the development of computing technology as it exists today came from a statistical data-processing task, the United States 1880 census of population and housing. During the initial phases of manual processing of the 1880 census, it was foreseen that the final counts would not be produced until almost 1890, the year of the next census. Accordingly, the United States Bureau of the Census commissioned a young engineer, Herman Hollerith, to design a series of machines that could be used for tabulation of the census results. Hollerith designed such a series of machines, and they were used successfully with the 1890 census data. In contrast to the census of 1880, which took about eight years to complete, the census of 1890 was finished in about two years.

57. Based upon his early machine designs, Hollerith established a company which became known as the C-T-R (Computing-Tabulating-Recording) Corporation, which later became the International Business Machines Corporation, now called IBM.

58. Modern statistical data processing began with the development of automatic data-processing equipment that was based upon electronic circuitry and was capable of stored program operation. Electronic circuitry, first developed in the 1940s, provided computational speed advantages of 3 to 4 orders of magnitude over previous mechanical devices. The invention of the stored program by Von Neumann allowed for retrieval and execution of program steps at electronic rather than mechanical speeds, and provided the property of self-modification. These advances led to the installation of the first commercial computer - a Univac I - in 1950 at the United States Bureau of the Census for the purpose of processing the data collected in the 1950 census of population and housing.

59. Data-processing hardware is often categorized in terms of the generation to which it belongs. In general, hardware of the first generation was based on vacuum tube technology and hardware of the second generation was based upon transistor technology with discrete components mounted on circuit boards. Successive generation technology has been based upon medium-scale and large-scale integration of electronic components using miniaturized semiconductor circuitry based generally upon photolithographic techniques.

60. The use of generations to identify hardware serves less of a purpose as data-processing technology advances. In the early days of computing, software technology was closely coupled with hardware technology and provided a reasonable typology for computing environments. More recently, both technologies have become considerably

/...

richer in alternatives and have been packaged in various combinations that can be difficult to categorize in terms of technological generation. For users of such equipment, other attributes of equipment are generally more important, such as reliability, capacity, modularity, price and performance characteristics, repair and maintenance services, software choices and ease of operation.

61. Data-processing hardware is typically divided into a number of functional components: (a) processing units; (b) primary memory; (c) secondary memory; (d) input and output devices; and (e) communications equipment. This division forms the basis for the discussion below.

A. Processing units

62. Processing units are currently undergoing significant advances in price-performance behaviour as a result of advances in the underlying LSI (large-scale integration) technology. While central processing units were a very significant expense in first and second generation machines, they are now relatively inexpensive in most computers and may, in the near future, become a small fraction of the total cost of a data-processing system in all but the fastest systems manufactured.

63. Advances in LSI fabrication affect almost the entire range of processors, from the smallest to among the largest. It is now common for processing units to be contained on - at the low end - a single electronic chip about the size of a person's finger or - at the high end - at most, a few moderate size integrated circuit boards. The miniaturization aspect of LSI technology has resulted in lower price per component, lower power consumption and greater reliability. Advances in miniaturization are expected to continue for at least the next 10 years, yielding 20-30 per cent more processing capability each year for the same cost.

64. Of particular importance has been the growth of microprocessors. Such processors are small and generally have limited data path widths and instruction repertories, and they are quite inexpensive. A typical current microprocessor, the Zilog Z-80, has a 250 nanosecond cycle time and 8-bit logic and sells for about $US 20 when purchased in small quantities. (One nanosecond is one billionth of a second, or 10^{-9} seconds.) Currently, 16-bit processors with memory management units are beginning to have an impact on the market; these processors are similar in capability and speed to large minicomputers of several years ago. The next generation of 32-bit microprocessors is technically feasible now; when such a product appears, it is likely that the typical central processor of a small to medium present-day system will exist on one or a few small electronic chips and will be very inexpensive.

65. One important implication of the emergence of inexpensive, relatively powerful microprocessor hardware is that the hardware economies of scale that were believed to exist in earlier generations of hardware either no longer exist or are insignificant for most practical purposes. The observation about hardware economies of scale was formalized many years ago as Grosch's law, which stated that the price of a computer was roughly proportional to the square of its power. This relationship which for many years was believed to exist, was commonly used in

/...

support of centralization of data-processing capability within organizations as the most effective method for efficient use of what have historically been quite expensive and, therefore, generally scarce resources.

66. The availability of a wide variety of microcomputers and minicomputers, combined with inexpensive primary memory, makes it possible to distribute computing power cheaply and effectively. While some modest hardware economies of scale may be lost in doing so, they are generally overwhelmed by advantages in efficiency, owing to more simplified software environments and directness of control. Significant software diseconomies of scale that often existed in large, multi-user multi-purpose machines because of a perceived requirement to make optimal use of the hardware component of a computer system have now been substantially reduced by using smaller hardware configurations for more specific and less general purposes.

B. Primary memory

67. Primary memory is that memory within a computer system that is most immediately accessible to its processor. It is often called "immediate access" memory, connoting that each component of the memory is immediately accessible, that is, in equal intervals of time. While such memory is almost always random access in that patterns of accesses to it of equal length can be completed in approximately the same amount of time, the term "random access" is usually reserved for secondary rotating storage media. In general, programme instructions and data elements of immediate interest are contained in primary memory for rapid programme execution.

68. While primary memory is frequently referred to as "core" memory, connoting the extensive historical use of magnetic core technology for such memory, the bulk of current primary memory technology relies upon active semiconductor circuit technology to maintain memory elements. Thus, the same LSI fabrication techniques that advance processor technology also serve to advance primary memory technology. Currently, commercial semiconductor memory costs about $US 10 for 1 KB (K = 1,024; B = byte = 8 bits), and the cost is being halved approximately every three years. Current memory chips routinely store 16 K bits (binary digits), while 64 K bit chips are now entering the marketplace in commercial products.

69. In addition to providing memory at lower cost, semiconductor technology now provides memory products with greater reliability, less power consumption, increased automatic error-correction features and higher speed. Such trends promote the distribution of increasing amounts of specialized intelligence throughout organizations and processes.

70. The emergence of relatively inexpensive computing systems with increasing primary memory size has direct application to statistical data processing. The most direct implication is that one can expect computing environments to become somewhat larger and more decentralized. Given larger environments, more extensive programmes and integrated systems of programmes will be usable, and more users will be able to make use of statistical software capital such as SPSS and PSTAT that was previously restricted to older machines that could offer large primary memory capacity for use by application programme. Such broader use of existing capital

should also induce further statistical system development, which can only be beneficial to users.

71. The increasing decentralization of computer power could have significant benefits for statistical data processing because such systems can provide relatively convenient mechanisms for functional and geographical distribution of data-processing tasks. Both types of distribution have the advantage of moving the processing function closer to the person, group or organization that has primary responsibility for it. Thus, for example, the ability to decentralize the data-processing function could lead to direct data entry, editing and correction by census and survey organizations at the time when and place where the original source documents are present for reference. The subject-matter knowledge of the specialists could then be used to improve the correctness of the data, rather than relying upon a substantively passive data-recording operation.

C. Secondary memory

72. Secondary memory consists of a variety of storage devices that are used to store data items of less immediate interest to programmes being executed. Compared to primary memory, secondary memory devices are cheaper, larger and slower. Within most computer systems, there is a hierarchy of memory levels both at the primary and at the secondary level; at the top, memory is small, expensive and very fast; at the bottom, memory is voluminous, cheap and slow.

73. The most prevalent current secondary memory devices are magnetic disk and magnetic tape. Disk storage is accessed mostly randomly, while magnetic tape is accessed sequentially. Both rely upon the use of digital magnetic recording technology with extensive error-correction coding. Other forms of secondary storage consist of magnetic bubble memory, charge-coupled devices and variations on standard magnetic tape.

74. Technologically, probably the most important recent developments in disk technology have been the introduction of the sealed disk module and non-removable disks. The sealed disk module technology, often called "Winchester" technology after IBM's code name for its development, provides a hermetically sealed, and therefore non-contaminated, environment for data transfer and storage. The controlled aerodynamic characteristics within the module allow considerably smaller read-write heads with a large potential increase in recording density and module data capacity. For those purposes for which disk modules need not be removed from the computer system, there now exist a variety of non-removable or fixed-module disk products at lower prices and sometimes with better performance for the same capacity than corresponding devices with removable modules.

75. The explosive growth of the microcomputing industry is now beginning to exploit Winchester disk technology in depth, and a wide range of new products based on this technology is becoming available. The range is characterized by a diversity of manufacturers, significant modularity, low cost per unit of capacity and standardization of physical and electronic interfaces to computer systems. The devices are quite suitable for support of decentralized processing operations and

/...

provide sufficient secondary storage capacity for robust operating systems, language processors, utilities, work space and user file systems; in short, they allow replication at the microcomputer level of robust operating systems previously only available at the minicomputer or computer system level. Economies of scale on large systems are now being reduced significantly by the emergence of a mass market for robust microcomputer-based systems.

76. Tape technology is also exhibiting some progress. The emergence of a recording density of 6,250 characters per inch now permits up to approximately 180 MB of information to be stored on one reel of tape, although hardware to support this density is still relatively expensive. More recently, the American National Standards Institute (ANSI) standard tape cassette has been used by one manufacturer to support low cost, block indexed storage of 75 MB per cassette. Further, a new mode of tape recording has emerged, that of streaming mode, in which there is rapid, generally dedicated transfer of information in large quantities between disk and tape. Such a mode is generally used to back up and restore information on non-removable disks of large capacity. Magnetic tape technology still offers extremely inexpensive storage of very large volumes of information both for archival purposes and for routine sequential processing tasks, and it does not now appear that the medium will be displaced in either of these roles in the near future.

D. Input and output devices

77. Input devices have always been of substantial importance in statistical data-processing applications. The most traditional and most common method of accomplishing data-recording tasks has been keyboarding to a machine-sensible medium, such as punch cards, magnetic tape, magnetic diskette or cassette or magnetic disk via a computer-based entry system. Other methods have included direct data entry under utility programme or application programme control, optical mark reading, character recognition and digitization of form co-ordinates.

78. In the domain of keyboard entry systems, magnetic-medium recording devices have substantially displaced punch cards as the preferred medium of use. Magnetic media have a number of advantages in terms of capacity, flexibility, reusability, speed and cost. Keyboarding machines and systems are acquiring greater functionality and, as advances in LSI techniques and microcomputer systems engineering continue, there is every reason to believe that the traditional data-entry station will acquire computer-like characteristics. This development will present significant opportunities for pre- and/or post-processing data at a local data-entry installation and can be used to advantage in many situations. In particular, computer-assisted editing of some complexity becomes possible at low cost at the point of data recording when the source documents that define the data are readily available for reference. While this does not imply that such editing should be done at the time of data entry, the new equipment will offer the possibility and, therefore, enrich the set of alternatives available to the system designer.

79. Research is currently under way in two areas that have long-range promise for data entry, voice-recognition and automatic page-reading devices. Voice-recognition

devices having a moderate vocabulary exist now, but at high cost; more primitive devices exist at much lower cost as microcomputer system accessories. They function, but are not now effective in mass data-entry applications. However, this technology is expected to improve significantly in the long run; extensive research is under way because of the potential size of the market, and advances in circuitry price-performance will directly benefit voice-recognition devices.

80. A significant breakthrough has been made in a minicomputer-controlled optical reader that can recognize and interpret pages printed in mixtures of fonts. An initial version of the device has been combined with a speech-synthesizing device to create a talking reading machine for blind users. While the technology is still very expensive and not now suitable for statistical data-processing applications, its generality and the large potential market for such devices will guarantee substantial research in this area. Like voice recognition, the technology is also compute-intensive, so that advances in LSI and microcomputer system engineering will benefit it directly.

81. The range of output devices continues to expand. Largely because of the growth of minicomputer-based and microcomputer-based systems, the market in low-functionality to medium-functionality VDUs (visual display units) and low-speed to medium-speed printers has exploded, with low and medium resolution graphic output displays available on the same units at moderate cost. At the high end of the spectrum, laser technology is being used to manufacture printers having a significantly higher line-output rate than mechanical devices are capable of.

82. Speech-synthesis devices have been available for quite some time, but at significant cost. The costs are now decreasing significantly, partially in response to an expanding market and partly in response to decreasing LSI costs. Speech output should become commonly available within 5-10 years and may provide acceptable output when the volume of output is low and a written record is not required. Management data retrieval and analysis systems, such as decision support systems, might be positioned to take advantage of this technology. Depending upon cost, it could also possibly be used to advantage to guide the data-entry or edit process.

E. Communications equipment

83. The linkages between communications networks and computer systems continues to strengthen and become more complex over time. In developed countries, this trend has been supported by pressures towards office automation and electronic mail and by the ability to decentralize computer-based intelligence geographically (see paras. 70-71 above). Just recently within the United States of America, a major shift in the regulatory climate promises to allow communications companies to enter the data-processing industry, which should further strengthen computer-communications linkages and encourage more aggressive product development in areas of intersection of these technologies.

84. Technical progress in communications hardware is generally somewhat slower than in the computing industry. Nevertheless, progress is being made in the cost-performance characteristics of statistical multiplexers for making more efficient

/...

shared use of single communication channels for digital transmission and in the use of faster modem (modulator-demodulator) units at speeds of 1200 band (120 characters per second) for data transmission. Of more long-range importance is the initial use of optical fibre cable paths, which promises to increase very substantially both over-all communication capacity and band width available to individual subscribers. Advances in digital circuit technology and increase in demand for digital communications have caused investment in new communications-switching plant and transmission links to become essentially digital rather than analog, promising more direct and efficient data communications interfaces in the future.

85. Of great importance for the average user of digital communications services has been the emergence of digital packet switching services in many of the developed countries. Such services, often called "value-added" networks, act essentially as wholesale agents. They purchase facilities from wideband communications networks, add their own digital packet transmission technology and then retail communications services to individual customers. Packet-switching technology depends upon inexpensive and capacious digital-switching devices which are really minicomputer and microcomputer systems operating at each node of a complex network. Rather than using communication channels, or circuits, that are allocated on a static basis for each transmission requirement, packet-switching networks employ dynamic routing and demand-based allocation techniques to transmit portions of the transmission through the network as required. These portions are called packets. The dynamic routing techniques allow the end user to observe the operation of an effective virtual communications circuit where no correspnnding real circuit exists.

86. Paradoxically, not only do packet network techniques permit more effective use of transmission bandwidth, which is currently a scarce resource, but they provide very significantly greater transmission accuracy and reliability along with a reduction in cost. Advances in LSI technology, coupled with an increase in competition in the market place, should improve packet services even further and increasingly encompass related types of service.

87. Packet-switching networks now extend to most developed countries and to a number of developing countries. Both interfaces to packet networks and international interconnexion of packet networks are currently subjects of intense standardization activity. Networks based on packet-switching are easily extendable to developing countries and even within some developing countries. Such a system would allow effective and relatively inexpensive digital communication for multiple users across sectors both within such a country and with other countries having access to such networks.

F. Synopsis

88. The prognosis is excellent for the increasing effectiveness of computing hardware in statistical data-processing activities. The computing industry is characterized by rapid technical progress in many areas, especially in hardware development, so that, in general, the choice of hardware alternatives available to the system designer becomes both richer and more cost-effective over time.

89. The rate of progress and generation of better and more choices in the hardware domain emphasizes what is becoming well accepted in the industry, namely, that software engineering and data-processing management are major areas of concern for the development and use of computing systems. In other words, effective software systems and effective management of data-processing resources, in general, are critical areas of concern for those who depend upon computer-based systems for their work. This is as true in statistical data processing as it is for the data-processing field in general. Hardware is but one input into the data-processing production function.

90. Statistical offices, especially in developing countries, should benefit significantly from the current trends. Reliability of hardware will increase even as prices decrease. Thus, the computing environment will become more reliable. Decentralization of processing will be easier to support because of (a) more reliable equipment for remote placement; (b) lower prices, allowing purchase of redundant elements in the configuration; and (c) greater ease of transporting components to central maintenance sites for repairs. Within operating environments that are more difficult than the average, the possibility will increase of relying upon support from another country through shipment of components, thus providing sources of supply for data-processing equipment to augment local, and often monopolistic, markets.

91. The challenge of statistical data processing consists of wisely combining good choices from the domain of hardware with appropriate investments in programming so that the resulting products will be reliable, adaptable, easy to use, appropriate and efficient in an economic sense. Within this context, new developments in hardware technology will continue to provide a sturdy, existing and constantly evolving basis for creating effective systems for statistical data processing.

IV. POINTS FOR DISCUSSION

92. The Commission may wish to:

(a) make general comments on the present document;

(b) suggest further means of transferring technology and know-how in statistical data processing to developing countries;

(c) suggest means of strengthening co-ordination between the United Nations Statistical Office and other agencies (multilateral and bilateral) in providing support to developing countries.
