UNITED NATIONS
ECONOMIC AND SOCIAL COUNCIL

ECONOMIC COMMISSION FOR AFRICA

Second Meeting of the Committee on
Development Information (CODI)

4-7 September 2001
Addis Ababa, Ethiopia

Report on ECA Database Management Development Activities.

I.   Introduction

1. The original recommendations for the United Nations Economic Commission for Africa (ECA) to design, develop and operate regional statistical databases go back to 1959.   These recommendations were formulated pursuant to resolutions adopted by the First Conference of African Statisticians with a view to instituting procedures for an effective response to various requests for structured, easily accessible and usable statistical information as a tool for decision making and as factor inputs to research within and outside Africa.   Subsequently, several initiatives came to confirm the urgency of this undertaking and to enrich the quality of statistical data required on Africa.   The most important of these was the formulation by the Sixth Session of the Joint Conference of African Planners, Statisticians, Population and Information Scientists of the Addis Ababa Plan of Action for Statistical Development in Africa in the 1990s.

2. Statistical design and development work itself only began in 1983 and the first structure became operational in 1985.   The prevailing computing environment was dominated by the Pan African Documentation and Information System (PADIS) project and the limited range, if not actual lack of skills in these areas which weighed heavily on PADIS – STAT design and development activities.   Subsequently, these activities were transferred as a priority programme element to the Statistics Division without ever having had the material and human support required to make the endeavour a success.

3. With the institution of new structures at ECA on 1 January 1997, the activities of the Statistics Division were fully taken over by the new Development Information Services Division (DISD). In the new structure, the development of regional statistical databases fully assumed its original multi-sectoral dimension, integrating as it did economic, bibliographic as well as geographical information.

4. This paper gives a comprehensive picture of the first – generation database structure developed on the Hewlett Packard 3000 mainframe computer, since migration to temporary MS-ACCESS platforms has merely transposed on to the basic structure.   This presentation provides an understanding of the characteristics of the existing modules.   Also described are the key recommendations made by the consultant who conducted the feasibility study for the development of the ECA Statistical Information System (ECA-SIS), as well as similar activities undertaken in other operational Divisions of the Commission, including the Subregional Development Centres (SRDCs) and by other external partners.

II.   The first-generation statistical database

5. This database was designed and developed under the PADIS project.   The computing, human and financial resources available at the time were limited relative to the magnitude of the task to be accomplished.   The database was installed on the HP 3000/58 computer available at the time with a mainframe memory of 5 megabytes, for diskette readers that could each stock 404 megabytes of data, two magnetic tape players with the capacity for recording 1600 and 1600/6250 bpi, printers and monitoring terminals.   It was accessed from all terminals connected to the system in the ECA premises and from OAU by means of a special telephone line.

6. The HP3000 computer was managed by an MPE4 operating system whereas the data file management software included IMAGE 3000, QUERY 3000 and KSAM. The user programmes included INSIGHT, VIEW, TDP (text processing) SORT, MERGE AND FCOPY. Programming languages like BASIC, COBOL, FORTRAN, RPG and SPL were used for the various components.

7. The design development and maintenance of the database was led by a restricted team of one computer scientist working full time and two statisticians working part time to develop the database while other statisticians in the Division took care to manage their own data.

8. The first statistical database was designed in three integrated levels which were to form the nucleus of a comprehensive statistical information system for statistical data retrieval, storage, management, processing and dissemination. The first level gives a brief picture, in 2 – 3 page monographs, of the economic and social situation in each member State of the Commission. The second contains a set of detailed statistics at the lowest level of disaggregation. It was used as the main source of data for the two other levels and its development was to be the first stage in the building of the entire system, but that did not happen because of inadequate resources for electronic processing of the data and the luck of skilled manpower. In the form that it was installed on the HP 3000 computer, this level only contained data on international trade and national accounting statistics stored on magnetic tape for future batch processing. The third level consisted of time series focusing on 13 different sectors, namely agriculture, transport and communication, education, finance, health, industry, employment, population, international trade, civil registration statistics, prices, the environment and national accounts. These were the series most frequently requested by users. Sectoral content, however, developed to an uneven pattern. This component of the information system was also developed as a priority, given the strong internal demand for time series data and the need to speed up the publication of ECA statistical information. The initial capacity was set at 150,000 entries for 120,000 series. The languages to be used were French and English but only English became operational.

9. The logical structure is described in table 1 of the Annex. The codification was, to the extent feasible, made to follow international nomenclatures such as the Standard International Trade Classification (SITC), the partial list of products and material derived from the Standard International Classification of Industry (SICI) for all types of economic activity, the International Statistical Yearbook and the standard country code of the United Nations Statistics Division (UNSD). Codes were introduced to cover subregional and regional groupings. The physical structure of the data follows the rules that apply to databases managed by IMAGE, namely the organisation of data files in two sets: a master set which can be operated manually or automatically and a detail subset comprising a collection of data-related recordings. From the former, access can be gained to the later through identifiers.

10. The collection of basic data on ECA member countries has neither been systematic nor regular. It is currently conducted on an ad-hoc basis and depends on the needs for annual publications of the Commission such as the Statistical Yearbook, the Directory of the African Statisticians or of African Data Processing Centres, the survey of Economic and Social Conditions in Africa and the Africa Economic Report. For the collection of data, questionnaires are sent out, missions are fielded to member countries (a practice severely limited by the scarcity of financial resources) and use is made of publications issued by other specialised agencies such as FAO,

UNIDO, OECD, UNCTAD and the UNSD. The system of validation is decentralised to the workstations of statistical assistants attached to the various domains of the database.

11. The third level which has been the most operational was designed and developed with the HP IMAGE 3000 database management system which is used to define and create the database and also to access it. The system of general database management is not suited to the specific needs of statisticians so special programmes have been developed to tailor the characteristics to such specific request as those for statistical data processing, table creation, basic statistical computing, database maintenance and the regular updating of derived data.

12. The database has been used to set up the nucleus of a rational system for publishing statistics on subjects such as statistical yearbooks for Africa, socio-economic indicators, external trade statistics, the statistical annexes of the Survey of Economic and Social conditions in Africa and various other tables produced and sent to countries and international organisations in various formats. The use of the system has been limited by the following constraints: (a) running out of disk space given the large size of the series; (b) the fact that its establishment in the member countries has been limited by lack of portability; (c) the slow pace at which statistical tables are published; (d) the constraining procedures for data updating created by the unwieldy modes of access; (e) the marginalisation of member countries in the design, development and even the updating process: (f) the frequent risks of error in aggregates which, unfortunately, are not calculated at the time the tables are published but beforehand and entered into the database; (g) the limited possibilities of interactivity which can allow for series-by-series consultation but not for batch processing; (h) the fact that there is no procedure for monitoring the data content; and (i) the absence of an integrated meta-info system in the database. Mention should also be made of the limited possibilities of computing functions and graphic displays. Because of these constraints and the continuing lack of skilled manpower, levels 1 and 2 have not been developed subsequently.

III. The migration from the HP 3000 mainframe to personal computers

13. When in the mid 90s, the HP 3000/59 reached its limit both in terms of life expectancy and storage capacity, interruptions became frequent and the secretariat had to initiate the process of decommissioning. It became urgent to transfer the resident data to an albeit temporary platform, pending the building of the physical infrastructure needed to develop an operational network and the coming on stream of skills for database, interface and applications design and development.

14. With the spread of personal computers (PCs) within the ECA secretariat, the bulk of data stored on the HP 3000 was transferred to the workstations of research assistants pending migration of the entire statistical database. The transfer operations were conducted in various formats (QUATTRO PRO, PARADOX and EXCEL) and disparately without co-ordination and coherence in the classification standards and validation procedures. The development of the local area network (LAN) and the installation of a server for statistical activities have made it possible to migrate the integral database to the network.

15. For this phase, the data model and relational database management system were selected as the standard and migration platform and as the software for managing the database. This option

appears justified because it is more suited to a computing environment built around PCs operating in a network and using the MS windows management system that ECA has. Furthermore, the maintenance staff, have acquired solid experience with this software, thus saving time and resources during the transitional phase.

16. The substantial volume and magnitude of the series, however, make it difficult to work on the whole database using MS ACCESS resources and tools. For that reason, the initial base was separated into 12 more manageable and PC-friendly sectoral databases. These components have to do with the following sectors (each having its own database): agriculture, transport and communications, population and social statistics, finance, industry, trade, prices, national accounts, energy, environment and tourism. Energy, environment and tourism modules transferred from the HP 3000 mainframe have been emptied of any recording, as was the case for the original data files. This situation will continue until the response rate to questionnaires sent out to member States on the two sectors increases. The only resources available at the moment are the existing national web sites, the Yearbook of World Tourism Statistics, the African Development Bank and the AFRISTAT web sites. The coding classifications for all the data are the same as those for the old database.

17. The central database, together with the sectoral components in MS ACCESS format, have now been designed around six sets of tables similar in structure to the database on the HP 3000 mainframe. There is a detail statistical data table with 6,289 recordings and 5 master data set tables having 122,007 recordings. The latter set comprises the classification code for statistical series, the series observation modes, the codes of reporting countries, names of countries and their currencies, the observed values of the series and the table of measurement units. Data on this part are available both in text and in MS ACCESS format. In addition, the design and structure of all the sectoral databases and their master tables are identical and make both for easy processing of the sectoral modules at independent work stations and for subsequent integration of the component.

18. Highly user-friendly menus have been created in the MS-ACCESS tools, forms, requests, reports and macros and procedures developed for easy use of the sectoral databases. The same applies to the generation of standard tables for the Statistical Yearbook, the Compendium of External Trade Statistics and for the aggregation procedures.

19. The following problems and constraints imerged generally during the period of migration. Indeed, in a progress report on the development of ECA's regional statistical database, the number of recordings was estimated at 160,000 with series covering the period 1965-1990. This figure was only 120,000 at the time of migration when the period covered was from 1970 to 1995. ECA has therefore lost a substantial amount of data. The external trade data which have been temporarily stored on more that 370 magnetic tapes as well as documentation on the database itself have disappeared. And yet, the available data can be used with different types of database management systems and applications in developing the nucleus of the ECA information system.

20. With regard to the sectoral components, some updating of the standards will be required as for example with the conversion of the national accounting series of the 1968 System of National

Accounts (SNA 68) to SNA 93; a new writing of data processing programmes especially for national accounts, price indices and/or the acquisition of appropriate software such as ERETES.

21. The most recent accessible data for transport and communications date back to 1997 (with few country exceptions) and major differences continue to exist among the various national and international sources. Attempts to improve the coverage have proved futile because the response rate to questionnaires sent to ECA member States has been only 23 per cent. ECA has, therefore, been relying on international sources such as UNSD's publications on road, rail transport and communications statistics, UNESCO's on communications infrastructure and the World Bank's on rail transport databases. The coverage of maritime and air transport statistics remains very poor. It is equally important to expand the database content with a view to reflecting technological innovations such as mobile telephone, and internet connectivity.

22. Most of the available price statistics rely on consumer price indexes and are often confined to urban consumption. The most recent information dates back to 1995/96 but updated corresponding series have been extracted mostly from the ILO Labour Statistics Yearbook, national web sites and the AFRISTAT web site where time series covering as many as 30 years are published for all the States members of the Observatory. The price statistics are those most regularly published in most of the countries and which very often appear monthly. These data resources are fairly widely available on the Internet and the updating of this component of the database is not a serious problem. The secretariat undertook a comprehensive national data collection exercise in this sector and the result has been one example of success in regular publication of reliable and timely national statistics. Finally, the maintenance of the energy, industry and the international trade modules has lagged behind because recent data are not available and problems have been experienced accessing external data sources and resources such as those of the International Computing Centre in Geneva.

23. In conclusion, the main problems holding back the development of a reliable and operational statistical database have to do with the multiplicity of validation posts and migration of data from the HP 3000, the maintenance of the sectoral components of the current database, lack of meta-info documentation, difficulties accessing external sources and of accessing available data, which are more specifically a matter of user connectivity. A meta-data and informational system monitoring the reliability and compilation methodologies should therefore be built into the database and into its various components. More intensive use of resources available on the Internet and exchange of electronic data among ECA secretariat, member States and those international institutions which have developed meta-info systems for generating statistics should be promoted. The current activities of the team working on database development, maintenance and management are geared to achieving this objective.

IV.   Micro Databases

24. The secretariat has also developed thematic databases using information from the demographic and health surveys conducted in Africa. The corresponding forms and data files are useful for comparing and harmonising social, demographic and health statistics as well as for developing a non-monitory approach to poverty reduction. They contain indicators on practices and health access, behaviours relating to sexually transmitted diseases including HIV-AIDS, fertility and its determinants as well as many other social and demographic variables. Databases had been

built for 29 African countries in which a total of 48 surveys were conducted and 82 indicators defined both for urban and rural areas. These statistics are disseminated on CD ROMs.

25. The content of the databases has been expanded to 136 indicators and other themes have been added to the initial structure. The meta-info presented have also been enriched. Currently, work is being undertaken on the comparability and harmonisation of indicators based on 15 to 20 different types of survey data files. A presentation of the test results will be made during this session.

V.   The current trend; client-server database customised to user needs

26. Reforms in ECA secretariat have led to the refocusing, especially of statistical activities, to reflect the genuine need of users. Such customisation in the development of databases has been made easier by: (a) promotion of the intensive use of new database design and development tools such as relational database management systems and target-oriented programming; (b) the expansion of networks and sharing of statistical data and methodologies; and (c) the general trend towards the use of internet tools and resources in conducting statistical activities. This is how the secretariat launched the design and development of its Statistical Information System, including the multi-sectoral and regional database component.

27. The initial phases ended with:

   a) a review of the existing environment (information system, database and computing environment) within ECA Secretariat and the SRDCs conducted through interviews of the various actors;
   b) the design of modular design data (MCDs) and design processing modules using a representative sub-set identified from the review of the existing environment;
   c) the functional architecture of the new system and proper planning for its operationalisation within the approved scenario for designing and developing ECA-SIS covering such aspects as the technical environment, alternative platforms, software programmes, SGDRs, servers, dissemination modes and costs (human, material and financial).

28. Already, the shortcomings of the current system as revealed by the review in terms of data and processing capacity can be described as follows:

   a) the multiplicity of data sources which expose users to hard choices and lead to publication of contradictory and inconsistent statistics within the same secretariat;
   b) compartmentalised divisions not having an internal system of information sharing with the result that even ECA users have to turn to external sources and unco-ordinated data updating;
   c) limited access to sensitive data, particularly those relating to household surveys and which are essential for poverty analyses;
   d) insufficient, if not total lack of data sharing between headquarters and the SRDCs;
   e) duplication of data treatment and lack of resource sharing and rationalisation; and
   f) the unco-ordinated nature of updating exercises earlier referred to.

29. Recommendations were also made on data management, organisation and substantive focus. The final report of the consultant whose terms of reference are given in annex 2 will shortly become available.

IV.   Conclusions and recommendations

30. Preserving existing data on a temporary platform, reviewing the existing environment of databases and information systems, identifying user needs and proposing technical options for the functional architecture will be the way forward for developing the ECA statistical information system, particularly when the resources and skill constraints have been overcome. The planning for implementation should be flexible enough to allow for customisation to user needs and the existing environment. For this to be done successfully, a multi-disciplinary team made up of experts from the various substantive divisions should be established.

Annex 1

Logical structure for the first generation statistical database

Variables	Description	Bytes	Type
SERCODE	Series identifier	6	Alphanumeric
REPCODE	Reporting country code	2	Alphanumeric
UNIT	Observation unit	2	Digital
MODE	Observation mode	2	Alphanumeric
PARCODE	Partner country code	4	Digital
BASEYR	Base year applicable	2	Digital
PRDTYPE	Period type	2	Alphabetical
STARTYR	Series starting year	2	Digital
YRT	Observation value for year T	9	Signed digital
FNT	Data type and source for year T	2	Alphanumeric
UPDATE	Date updated	6	Digital
DRID	Creator ID	2	Alphabetical
DATATYPE	Confidentiality indicator	2	Alphabetical
EXPDATE	Expiration date for confidential data	6	Digital

Annex 2
Terms of Reference
Design and Development of
ECA's Development/Statistical Information System

Summary of the terms of reference and work assignments (see attachments)

⇒ **Definition of a comprehensive master technical design** for a computerized UNECA statistical/development information system: formulate a complete set of subsystem components. including functional requirements. major information flows. relationships between current and potential uses and users:

⇒ **Compilation of a final technical report,** including: the users' requirements, the master technical design, the schematic diagram of the subsystem components, a comparative analysis and list of benefits. and, an estimate of the cost. person-hours and time-frame for the development and implementation of the recommended system.

Phase I

1. Conduct extensive **interviews** with UNECA staff and management to identify and articulate the full range of their **functional requirements** for development/statistical information/data processing;

2. Identify the **major flows. relationships and current and potential uses and users** of UNECA development information/statistical data:

3. Based on 1 and 2 above. **formulate a complete set of subsystem components** necessary to build a **comprehensive computerized UNECA development information/statistical information system.** providing. for example. for data input. storage and management. and define their characteristics and features:

Phase II

4. While taking into consideration the existing software and hardware in use in UNECA **(the existing computing environment).** the latest developments in database/information systems software tools and information technology and all cost implications. propose the **most suitable database/information technologies. software utilities and packages, GUI tools and operating systems** to serve as the foundation for a flexible and adaptable computerized UNECA statistical/development information system. one requiringin turn minimal software maintenance:

5. Consolidate 1 through 4 above into a **cohesive master technical design** for a computerized UNECA statistical/development information system:

6. Devise a work and cost plan for the development and implementation of the **master technical system design:**

Phase III

7. Combine 1 through 6 above into a **final report.** The final report. in French (to be translated in English), shall be in machine-readable wordprocessor format and shall include. but not be limited to:

    (a)   **An executive summary;**

    (b)   **A compilation of users' requirements.** taking into account. but not limited to the following areas:

(i) accommodation of data/development information currently used at UNECA, as well as data which may become available in the future:

(ii) on-line directory system indicating what data are available and where, together with mechanism to easily access and manipulate them:

(iii) user interfaces:

(iv) data input, including data-entry and verification procedures, processing speed from data entry to data and information output:

(v) data confidentiality and security;

(vi) data storage and management:

(vii) on-line enquiries:

(viii) data output:

(ix) data analysis (e.g. statistical and graphical analysis):

(x) forecasting, modelling and mapping:

(xi) data import and export:

(c) **A master technical design,** which shall include, but not limited to recommendations and justifications for:

(i) The degree of data centralization:

(ii) Flexible and adaptable data storage:

(iii) User interface(s) and their appearance:

(iv) Procedures for inputting statistical data, in machine-readable forms and from the keyboard:

(v) Provisions for data manipulation, retrieval, output, analysis and presentation:

(vi) Linkages to external communications (electronic data interchange);

(vii) Concomitant human resource development:

(viii) Help facilities:

(d) **A schematic diagram of the subsystem components** of the recommended system, including suggestions for the software to be used at each point:

(e) **A comparative analysis and list of benefits** of the propsed system as opposed to the present situation in UNECA, including a comparative review of existing databases, and with other alternative systems available on the market:

(f) **An estimate of the cost, person-hours and time-frame** for development and implementation of the recommended system.

Phase IV

8. Presentation of the report to an evaluation Committee comprising staff from DISD and from other substantive Divisions as well as partners to seek comments.

9. Finalization of the report and submission to the relevant bodies for the launching of the effective development process.

Qualification

University degree in computer sciences, with advanced knowledge of statistics.

Experience

Five to eight years of experience in computer programming, designing and developing information systems and database management systems at national and/or international levels. Familiar with networks and UNIX and NOVELL NETWARE operating systems.

Language

French and/or English as working language.

Duration: 2 to 3 months

Cost:	1.	Honorarium (2 to 3 months)	US$18,000 (US$ Eighteen Thousand)
	2.	Travel and DSA	US$ 6,700 (US$ Six Thousand Seven Hundred)
		Total	US$24,700 (US$ Twenty Four Thousand Seven Hundred)

Duration: 2 to 3 months

Cost:			
	1.	Honorarium (2 months)	US$2*9,000 = US$18,000 (US$ eighteen thousand)
	2.	Travel	US$1,900 (US$ one thousand nine hundred)
	3.	DSA (2 months)	US$2*30*80 = US$4,800 (US$ four thousand eight
		hundred)	
		Total	US$24,700 (US$twenty four thousand seven

hundred)

Desirable beginning of contract: Second half of September