Distr.
LIMITEE

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

Original: FRANCAIS

NATIONS UNIES

CONSEIL ÉCONOMIQUE ET SOCIAL

COMMISSION ECONOMIQUE POUR L'AFRIQUE

Quatrième session de la Conférence commune des Planificateurs, Statisticiens et Démographes africains

Addis Abeba, 3-12 mars 1986

QUELQUES PROBLEMES RELATIFS AU DEVELOPPEMENT DE BASES DE DONNEES STATISTIQUES

Table des matières

I. INTRODUCTION

1. C'est en décembre 1982 que la division de la Statistique de la CEA a commencé à étudier la mise en place d'une base de données statistiques régionale. Un an plus tard, un premier rapport intitulé "Rapport sur le développement de la base de données statistiques de la CEA" décrivant les aspects conceptuels de cette base a été présenté à la Conférence commune des planificateurs, des statisticiens et des démographes africains lors de sa troisième session qui s'est tenue à Addis Ababa du 5 au 14 mars 1984. Les années 1984 et 1985 ont été consacrées à implémenter sous Image 3000 les outils de base nécessaires en amont (cération, mise à jour récupération des données de certains fichiers producteurs) et en aval (recherche/sélection/tabulation). Le présent document traite dans sa première partie de l'évaluation de ces outils et de quelques indications quant aux travaux futurs.

2. Lors de l'examen du premier rapport mentionné ci dessus, la troisième session de la Conférence commune avait émis le voeu que la CEA associe plus étroitement les pays africains à la mise en oeuvre de la base de données de la CEA également appelée PADIS-STAT, en particulier en les aidant à se doter eux-mêmes de bases de données statistiques. Dans cette perspective, le présent document traite dans sa deuxième partie de quelques éléments méthodologiques pouvant servir de base conceptuelle à la mise en place de bases de données statistiques au niveau national. Cette deuxième partie s'articule autour des points suivants : éléments conceptuels; procédure de mise en place.

II. LE SYSTEME PADIS-STAT

3. Le système PADIS-STAT a été mis en place pour :

1) offrir aux statisticiens et économistes de la CEA :

- un accès direct et rapide aux données statistiques africaines;
- des services informatiques de traitement et d'analyse de données.

2) permettre une automatisation des publications du système statistique de la CEA;

3) fournir des services d'assistance aux pays africains qui envisagent la création de bases de données statistiques.

4. PADIS-STAT est hiérarchisé en trois niveaux dont le niveau I et le niveau III peuvent être considérés comme des sous banques organiquement reliées à une banque maîtresse qui est le niveau II. Le niveau I est prévu pour contenir des données maîtresses permettant d'éditer automatiquement des fiches signalétiques pays. Le niveau II envisagé contiendrait tout le fond des données statistiques disponibles sur une longue période. Le niveau III qui est le seul niveau actuellement implanté, est orienté vers la diffusion et l'analyse.

a) Contenu et structure du niveau III

5. La base de données que constitue le niveau III intéresse les domaines suivants: population et emploi; comptabilité nationale; agriculture; industrie; transports et communications; finances; prix; commerce extérieur; éducation et santé. Elle contient actuellement plus de 100.000 séries chronologiques couvrant une période allant de 1970 à 1985. Cette base est installée sur le mini-ordinateur Hewlett Packard (HP) 3000 Série III de PADIS et gérée par le SGBD HP Image 3000 et des programmes écrits en Cobol par l'informaticien de la division de la statistique de la CEA. Ces programmes sont utilisés pour des traitements en temps réel et par lots. Ils sont accédés de façon interactive selon un menu global.

6. Le niveau III, sur le plan structurel est constitué par 9 sous fichiers ou table reliés par des pointeurs. Ces sous fichiers sont:

- Séries chronologiques
- Statut de la série
- Source
- Type de série
- Pays déclarant
- Pays partenaire
- Unité de mesure
- Identifiant producteur
- Mot de passe

7. L'accès aux données est verrouillé par un système de sécurité. Seules les personnes autorisées peuvent y avoir accès.

b) Principales utilisations du niveau III

8. Le niveau III est actuellement opérationnel. Il permet de générer automatiquement les tableaux de l'annuaire statistique africain, ceux concernant les indicateurs socio-économiques africains et certains tableaux des publications du commerce extérieur africain (séries A et C). Les calculs possibles pour l'instant se ramènent à ceux de moyenne, de pourcentage, de ratio, d'indice d'évolution. Il est envisagé un élargissement de ce champ en mettant en place des procédures d'estimation (interpolation et extrapolation) et des modules d'analyse statistique et économétrique. Cependant afin de ne pas alourdir la gestion du système, ces derniers éléments de calcul ne seront pas intégrés dans le système. Ils seront fournis par des progiciels généraux interfacés comme le SPSS.

c) Principaux problèmes à résoudre

9. Comme nous venons brièvement de le voir, les opérations fondamentales de gestion du niveau III ont été réalisées. Actuellement, grâce aux divers modules installés, on peut interroger en direct la banque, y sélectionner des séries et éditer les résultats de cette sélection. Cependant, d'autres actions importantes restent encore

à être menées à bien. En premier lieu, il s'agit de résoudre d'une manière satisfaisante le problème de la mise à jour de la base. Les pays de la région constituent la principale source de renseignements de PADIS-STAT. En effet, ce dernier repose essentiellement sur le système de production statistique de la CEA, lui même totalement dépendant des appareils statistiques nationaux. Or, la récupération par la CEA de la production statistique africaine rencontre les difficultés suivantes :

- délais trop longs en ce qui concerne la publication et l'acheminement des données statistiques entre les pays et la CEA;
- l'utilisation quasi totale du support papier entrainant des délais de saisie assez longs.

10. Afin de conserver une certaine qualité et une certaine utilité à la banque, il importe de résoudre ces difficultés. En ce domaine, la CEA essaye d'utiliser le plus largement possible toutes les sources de données (publications nationales, sources internationales, renseignements rassemblés au cours de missions). Cependant, la solution efficace serait que les pays africains enregistrent leurs informations statistiques sur supports magnétiques facilement reproductibles et transmissibles. A cet effet, il faudrait qu'ils intègrent d'une manière plus systématique la technique informatique dans leur processus de production statistique.

11. Le second problème à résoudre urgemment, concerne la documentation. Il importe en effet que les données introduites dans la base soient renseignées afin de pouvoir être utilisées et interprétées correctement. C'est dire que la base doit contenir un système d'information sur l'information qu'elle contient, interrogeable en ligne par les utilisateurs. Il s'agit donc de mettre en place :

i) un dictionnaire de données;

ii) la documentation des utilisateurs non informaticiens (language et procédure d'interrogation);

iii) la documentation de saisie (formats de saisie, procédure de saisie, liste des contrôles à effectuer);

iv) la documentation concernant le noyau (mise en oeuvre des opérations fondamentales, procédures de sécurité, langage de définition et manipulation des données).

12. Un troisième problème à résoudre est celui de la diffusion des données par la base. Une base de données devrait avoir pour premier mérite de permettre les stockage et le transport jusque chez l'utilisateurs d'un volume de données bien plus considérable que celui auparavant publié. Ceci suppose l'existence d'un support technique du transport des données jusqu'au terminal de l'utilisateur. Ce support peut être fourni par les télécommunications (fil, ondes herziennes, sattelites). Ici la diffusion des données se confond avec le transport. Cependant, il paraît irréaliste

pour l'heure d'envisager tout de suite ce mode de diffusion pour plusieurs raisons évidentes : données volumineuses et coût élevé du transport. Par contre, la diffusion des données de la base par utilisation de supports magnétiques (disquettes, bandes magnétiques) ou d'autres supports d'enregistrement (microfilms et microfiches) doit être envisagée et mise en place, parallèlement à la diffusion sur support papier, en priorité.

13. Enfin, il importe d'améliorer considérablement l'accès à la base pour que les utilisateurs finals soient autant que possible dispensés des contraintes de programmations et que la maîtrise des connaissances des techniques informatisées ne soit pas un préalable. A cette fin, l'année 1986 sera consacrée à la mise au point d'un système d'interrogation par menus permettant aux utilisateurs sans connaissance informatique de manipuler les données au moyen de commandes simples libellées en anglais et en français.

III. MISE EN PLACE DE BASES DE DONNEES STATISTIQUES NATIONALES

14. Un des objectifs de PADIS-STAT est de fournir des services d'assistance aux pays africains qui envisagent la création de banques de données statistiques. Etant donné que la construction et le développement d'une base de données statistiques relèvent beaucoup plus de l'organisation que de l'informatique, il importe d'apporter une attention particulière à la phase conceptuelle. En effet, c'est à partir de l'organisation retenue que les outils informatiques seront mis en place. Dans ce chapitre on trouvera quelques éléments conceptuels concernant la création de bases de données statistiques au niveau national.

a) Phase conceptuelle

15. Une base de données statistiques doit être conçue comme le prolongement normal des activités traditionnelles de production, d'analyse, de prévision et de diffusion d'un service national de statistique. C'est dire que la base doit être à la fois un outil de travail pour les statisticiens et un système d'information orienté vers les utilisateurs. La conception d'une telle base de données doit faire l'objet d'un cahier de charges incluant les éléments suivants :

- déterminer les utilisateurs éventuels et leurs besoins;
- déterminer les données qu'on manipulera. Il s'agit ici de dresser un inventaire des données à partir des utilisations prévues. On considérera deux classes de données, celles à stocker et celles à restituer par calcul. Parmi les données à stocker, il est nécessaire de distinguer celles qui seront sur disque parce que plus fréquemment utilisées et celles qui pourraient se contenter de bandes magnétiques comme support de stockage;
- déterminer dans leurs grandes lignes les scénarios d'utilisation et les procédures de mise à jour;

- déterminer le matériel informatique nécessaire à partir des exigences techniques de la banque (volume à stocker, nombre de transactions à traiter, les volumes imprimés, les volumes saisis);
- établir un budget pour créer la banque et pour la faire fonctionner (budget incluant le matériel et les logiciels).

b) Phase d'implantation logique

16. Un système statistique national est composé de quatre principales fonctions :

- la collecte et le traitement des données de recensement ou d'enquête ou de données tirées des dossiers administratifs;
- l'analyse et l'interprétation des données;
- la diffusion des données;
- la méta base qui est l'information sur l'information.

17. Une base de données statistiques ouvre la possibilité d'effectuer ces fonctions. En effet, elle est constituée d'un ensemble de services informatiques utilisant des fichiers de données organisés de manière à permettre le stockage et l'extraction de renseignements de toutes sortes aux fins de publications ou d'autres besoins. Les données peuvent être des données individuelles (produit primaire de recensement ou d'enquête) ou des données agrégées (séries chronologiques ou tableaux). Cependant, dans le contexte actuel africain, une base de données statistiques nationale devrait en priorité rassembler les données agrégées produites ou centralisées par un bureau national de statistique. D'autre part, elle ne devrait être, tout au moins dans un premier temps, qu'un gestionnaire de fichiers. En conséquence, les fonctions intégrées mises à la disposition des utilisateurs se limiteront à :

- la recherche proprement dite permettant d'isoler un sous ensemble de séries;
- la sélection permettant de choisir dans les sous ensemble constitué après une recherche, les séries ainsi que les caractéristiques et les périodes sur lesquelles on désire travailler;
- au transfert éventuel du résultat de la sélection dans une zone de travail privative;
- au calcul statistique simple;
- la présentation des résultats (rapports, tableaux, graphiques).

18. Toutes les fonctions non intégrées à la banque, la modélisation par exemple, devront faire appel à des logiciels extérieurs, spécifiques ou généraux.

1. Structure des données

19. Les principaux objets à prévoir dans une base de données statistiques agrégées sont :

- les séries
- les observations
- les nomenclatures
- les tables de passage
- les tables
- les tableaux standards
- le menu

a) Les séries

20. L'objet central d'une base de données macroéconomiques est la série chronologique. Sa structure devra prendre en compte les éléents suivants :

- source statistique de la série (recensement, enquête, fichier daministratif);
- pays/région/province concernés par la série;
- régularité (code indiquant si la série est régulière ou non);
- périodicité de la série;
- unité de temps sur laquelle porte les observations;
- série de flux ou de stock;
- type des valeurs de la série (unité physique, unité monétaire, etc.);
- mode de représentation (valeur absolue, indice, ratio, etc.);
- type de correction (brut, corrigé de variations saisonnière, etc.);
- système de taxes (HT, TTC);
- unité de mesure (unité dans laquelle sont exprimées les observations;
- origine de la série;
- précision affectée à l'unité;
- signe de la série (possibilité pour les observations d'être signées ou non);
- puissance affectée à l'unité;
- identification de la série;
- intitulé de la série;
- base d'une série d'indice;

- commentaire éventuel attaché à une série;
- date première observation;
- date dernière observation disponible;
- identifiant producteur de la série;
- protection de la série.

21. La presque totalité de ces informations seront des éléments codés. Les autres informations seront représentées par des textes courts ou par des entiers.

b) les observations

22. Une observation devra être représentée soit par une valeur numérique, soit par un signe conventionnel. Elle peut être accompagnée d'un commentaire. Les informations suivantes devront accompagner une observation :

- valeur : représentation numérique de la valeur d'une observation; elle peut être signée; la position de la virgule éventuelle sera définie au niveau de la série;
- état de l'observation : toute observation aura un statut relatif à son degré d'élaboration statistique : définitive, provisoire, semi-définitive, révisée, estimée;
- confidentialité : degré de confidentialité d'une observation;
- commentaire attaché éventuellement à une observation. Les représentations non numériques d'une observation se feront à l'aide de codes éléments tirés d'une table (exemple : non significatif, non calculé, non disponible etc.).

c) les nomenclatures

23. Toute nomenclature utilisée comme élément participant à la définition d'une série devra être identifiée, caractérisée et gérée par la base. Toute nomenclature devra être caractérisée par :

- code d'une nodalité de nomenclature;
- intitulé d'une modalité (texte court);
- commentaire éventuel d'une nomenclature qui sera texte;
- identifiant de la nomenclature représenté parr une code éléent d'une table de nomenclatures;
- champ d'application de la nomenclature représenté par un code élément;
- statut de la nomenclature (officielle, spécifique, nomenclature de gestion) représenté par un code élément.

d) les tables de passage

24. Les tables de passage traduiront les correspondances entre les postes de deux nomenclatures considérées respectivement comme nomenclature de départ et nomenclature d'arrivée.

e) les tables

25. Les tables constituent un élément important de la gestion d'une base de données statistiques. Elles permettent d'une part d'identifier les catégories d'objet et d'autre part de répertorier les objets au sein de leurs catégories. Les tables sont répertoriées à l'aide d'une table appelée table des tables permettant de guider l'utilisateur dans sa consultation.

f) le menu

26. Le menu a pour objet de proposer une approche logique de la base permettant de guider l'utilisateur d'étape en étape, vers la sélection de regroupements de séries et de désigner au dernier niveau la série de son choix. Les éléments suivants devront être utilisés pour structurer le menu :

- domaine ou noeud du menu (ensemble des données correspondant à un domaine, un regroupement de séries);
- identifiant du domaine ou du regroupement de séries;
- intitulé du domaine ou du regroupement;
- nombre de séries rattachées à un domaine ou à un regroupement de séries;
- code de protection du domaine;
- mot de passe du domaine;
- commentaire éventuel sur un domaine.

2) Récupération des fichiers producteurs

27. Un système d'information statistique comprend deux groupes d'activités différents cependant fortement liés :

- le premier groupe peut-être considéré comme un sous-système orienté source. Il a pour principales fonctions la collection, le premier traitement et le stockage des données corrigées selon leur structure primaire. Il s'agit ici de données individuelles dites encore données primaires;
- le second groupe ou sous système orienté utilisateur a pour objet principal d'organiser et de stocker les données primaires sous forme de fichiers dits fichiers producteurs orientés vers la présentation des données prenant le plus souvent le caractère de publications à contenus standardisés. Le premier

problème à résoudre pour mettre en place une base de données statistiques est de transformer ces fichiers producteurs en bases de données. Les fichiers producteurs se trouvent sous des formes diverses, non seulement du point de vue physique (support manuel ou informatique) mais du point de vue de leur structure logique.

28. Si un fichier est un support informatique, on devra décrire le matériel sur lequel il s'imploie, le système d'accès, la forme des enregistrements et la position des données dans ces enregistrements. S'il est sur papier, il faudra donner une description claire de la façon de lire les informations.

29. La présentation logique des fichiers devra être harmonisée en utilisant le schéma suivant :

- détermination des ensembles d'entités (séries, observations, nomenclatures, tables etc.);
- détermination des associations entre les ensembles d'entités (une association est une relation binaire entre deux ensembles d'entités);
- détermination des conditions d'intégrité qui ne se déduisent pas directement de la définition des ensembles;
- établissement d'un diagramme conceptuel des données (réprésentation simplifiée des ensembles d'entités et des associations);
- définition des procédures de manipulation des données. D'autre part, chaque fichier producteur devra faire l'objet d'une documentation complète portant sur les différents objets du fichier. Cette documentation devra être structurée en chapitres correspondant dans la mesure du possible aux différents domaines du menu.

IV. CONCLUSION

30. Nous avons essayé dans ce document de présenter d'abord une évaluation sommaire des outils mis en place et des actions qui restent encore à mener à bien en ce qui concerne la base de données statistiques de la CEA.

31. Dans ce domaine, nous pouvons dire que les travaux futurs devront permettre le passage graduel d'une base de données considérée comme outil de production pour les statisticiens de la CEA à une base de données ouverte à un plus large groupe d'utilisateurs et en premier lieu aux Etats membres. Pour favoriser ce passage, il importe de pouvoir créer un réseau africain de bases de données statistiques ayant comme élément central PADIS-STAT. L'implantation d'un tel réseau passe nécessairement par la mise en place de bases de données statistiques nationales. Aussi, avons nous également essayé dans ce document de définir une sorte de stratégie pour concevoir un système national d'information statistique orientée vers l'utilisateur. Cette stratégie se fonde sur le principe de base suivant : une base de données statistiques au niveau national doit être à la fois un outil de travail pour les statisticiens et un instrument de diffusion de l'information statistique.