NATIONS UNIES **E**

Conseil économique et social

Distr.
GÉNÉRALE

ECE/CES/2006/29
3 avril 2006

FRANÇAIS
Original: ANGLAIS

COMMISSION ÉCONOMIQUE POUR L'EUROPE **COMMISSION DE STATISTIQUE**

CONFÉRENCE DES STATISTICIENS EUROPÉENS

Cinquante-quatrième réunion plénière
Paris, 13-15 juin 2006
Point 6 de l'ordre du jour provisoire

SÉMINAIRE SUR LES RECENSEMENTS DE LA POPULATION ET DES HABITATIONS TROISIÈME PARTIE

L'entreposage des données dans la diffusion des résultats des recensements[1]

Communication présentée par l'Institut national de statistique de l'Espagne

I. RÉSUMÉ

1. Les systèmes d'information fondés sur la nouvelle démarche technologique de la veille économique peuvent considérablement améliorer l'utilisation des données récoltées lors d'opérations statistiques de grande ampleur, telles que le recensement de la population. La composante principale d'un tel système est la structure des données, les informations étant entreposées de façon à faciliter la recherche des données plutôt que leur traitement. Les entrepôts de données sont conçus dans cette optique sur la base de modèles dénormalisés afin d'obtenir les meilleurs résultats possibles.

2. Les utilisateurs communiquent avec le système à l'aide d'une interface simple mais puissante. Les structures complexes des données leur sont cachées. Ils n'ont qu'à choisir les variables et les conditions de la recherche pour obtenir des résultats. Grâce aux outils de traitement analytique en ligne (OLAP), ils peuvent trouver des informations en fonction de n'importe quel critère qui les intéresse. Par ailleurs, la mise en place du système nécessite une révision des procédures applicables en matière de confidentialité des statistiques et d'édition des données.

[1] Cette communication a été établie à l'invitation du secrétariat.

II LE RECENSEMENT DE 2001: UN DÉFI TECHNIQUE

3. L'Institut national de statistique (INE) d'Espagne s'est appuyé sur une forte composante technologique tout au long de l'opération de recensement. Les principaux aspects à relever à cet égard sont les suivants:

a) Les questionnaires étaient préremplis avec les données personnelles tirées du registre de la population; ils ont été soigneusement élaborés de façon à en faciliter la numérisation grâce à un système très performant qui a permis d'achever le traitement de la majeure partie des données dans les trois mois qui en suivaient la collecte;

b) L'ensemble de la population avait la possibilité de remplir les questionnaires sur Internet (l'Espagne est le premier pays au monde à avoir mis en œuvre cette option pour l'ensemble de la population).

4. Les éléments ci-après ont dû être pris en compte lors de l'organisation de la diffusion des résultats du recensement:

a) Internet serait la principale filière de diffusion. Des livres et des produits électroniques pourraient également être distribués, une partie des utilisateurs ayant déjà l'habitude de ces produits. L'accès à Internet serait gratuit et n'imposerait aucune exigence particulière aux clients en matière de vitesse ou de technique de communication;

b) Le mode de diffusion devait offrir autant d'informations que possible aux utilisateurs, expérimentés ou non;

c) Ainsi qu'il ressortait de l'expérience antérieure, de nombreuses demandes ponctuelles d'information étaient formulées. Celles-ci posaient un problème en raison du surcroît de travail qu'elles représentaient pour les techniciens. De telles demandes devaient être réduites au minimum et facturées.

5. Un système de diffusion classique établi sur la base d'une série de tableaux n'était pas à même de satisfaire de telles exigences. La nouvelle démarche serait celle d'un système en «libre-service» où les utilisateurs pourraient concevoir les tableaux en fonction de leurs propres besoins.

6. Le système de diffusion devait donc présenter les caractéristiques suivantes:

a) Le système en «libre-service» devait être capable de répondre à la plupart des demandes des utilisateurs en moins de 10 secondes, car avec un temps d'attente plus long ils risquaient de quitter le programme;

b) L'interface devait être très intuitive et conviviale;

c) Pour les utilisateurs habitués aux méthodes de diffusion classiques, un sous-système devait générer une série de tableaux prédéfinis contenant les informations les plus pertinentes.

7. Les caractéristiques propres à l'opération de recensement (nombre considérable de variables, niveau de détail géographique très poussé) rendaient celle-ci trop complexe pour être

traitée avec les bases de données relationnelles classiques. La combinaison de tous ces éléments a déterminé la solution: recourir aux techniques de modélisation multidimensionnelle des données et aux outils OLAP afin de créer un système efficace. Or ces outils sont intégrés au système de veille économique. L'INE n'ayant aucune expérience de tels systèmes, leur mise en place a constitué un défi.

III. ÉVALUATION DES DIFFÉRENTES SOLUTIONS ET CHOIX DE LA TECHNOLOGIE

8. Les systèmes d'information de ce type sont surtout utilisés dans le secteur privé où différents aspects de l'activité économique sont analysés en permanence. Il y a au moins trois grandes différences par rapport au travail du service de statistique:

a) Ces systèmes sont équipés de puissants mécanismes de mise à jour, car ils doivent gérer des informations très dynamiques (ventes d'un produit, cours des actions) et être à même d'informer les cadres de ce qui se passe à tout moment dans la société;

b) Les ratios et les rapports qui présentent un intérêt sont clairement définis et ces éléments sont les seuls à être stockés aux fins d'analyse;

c) Les utilisateurs de l'information peuvent être formés au maniement des systèmes et à l'exploitation des données dans un environnement spécifique.

9. En revanche, pour le processus de diffusion des statistiques:

a) Le délai de mise à jour n'est pas un facteur essentiel étant donné que les opérations statistiques ne produisent pas de résultats en continu. Le recensement, notamment, n'a lieu que tous les 10 ans;

b) Il faut prendre en considération de nombreux types d'utilisateurs et satisfaire des intérêts divers;

c) L'interface utilisateur doit être intuitive vu l'impossibilité de former tous les utilisateurs potentiels du système.

10. Pour offrir un service universel qui ne dépende pas de la technologie de l'utilisateur et qui n'exige pas le téléchargement de logiciels lourds (appliquettes, logiciel client, etc.), l'interface utilisateur doit être expressément conçue pour un tel environnement.

11. Il a donc fallu tenir compte de deux aspects importants dans le choix de la technologie:

a) Le temps de réponse aux demandes d'information, qui devait de préférence rester inférieur à cinq secondes;

b) L'existence d'outils permettant la personnalisation de l'interface afin d'obtenir une application utilisateur intuitive et suffisamment puissante.

12. La technologie proposée par l'institut SAS en matière d'outils de veille économique a été choisie pour les raisons suivantes:

a) Le logiciel, indépendant du matériel, offre une solution de bout en bout et permet l'accès à diverses sources de données;

b) C'est une technologie que l'INE connaît déjà, bon nombre de personnes travaillant avec des outils SAS;

c) Ce système s'est avéré le plus performant lors des essais préliminaires;

d) L'élaboration et la mise en place des solutions de veille économique et d'entreposage des données semblaient à la fois simples et rapides grâce à un assortiment d'outils bien rôdés.

13. Ce dernier point a été vérifié avant qu'une décision soit prise. Dans l'appel d'offres envoyé à différentes sociétés, l'INE demandait un prototype du système comportant cinq variables. En quelques jours, les consultants de SAS ont mis au point un système parfaitement fonctionnel et très performant.

14. L'architecture du système repose sur les éléments clefs suivants:

A. Logiciel

15. Scalable Performance Data Server (SPDS, serveur de données à performance adaptable): un système de base de données appartenant à SAS. Il est axé sur les demandes d'information et capable d'exécuter différents processus en parallèle et de travailler avec des objets partagés. Il utilise des index modernes de type arbre balancé.

16. Serveur OLAP: cet élément gère les structures multidimensionnelles dans un modèle HOLAP (OLAP hybride), à savoir non seulement les tableaux de données, comme dans un système de base de données relationnelle, mais également des hypercubes et des objets multidimensionnels. Il fournit un aperçu simple et logique sur les autres processus et agit en tant que serveur intermédiaire en réorientant les interrogations vers l'objet minimal permettant d'y répondre.

17. Administrateur d'entrepôt: l'outil de modélisation des données. Il permet de définir toutes les structures dans un environnement graphique intuitif et génère le code SAS nécessaire pour le processus d'extraction, de transformation et de chargement.

18. AppDev Studio/Integration Technologies: outil de mise au point de l'interface. Il fournit un grand nombre de mini-applications de Java tirant parti de l'environnement SAS.

19. Enterprise Guide: un outil mis à la disposition de l'utilisateur final pour les interrogations et les rapports.

B. Matériel

20. Environnement de production: deux ordinateurs Sun Fire 480 fonctionnant avec le système d'exploitation Solaris configurés en groupe pluricellulaire actif/passif afin de garantir une grande disponibilité. Ces ordinateurs sont reliés à un réseau de stockage Symmetrics d'EMC d'environ 2 To d'espace de stockage pour des objets détaillés et résumés.

IV. MODÉLISATION DES DONNÉES

21. La modélisation des données a été lancée à partir des données des fichiers textes plats résultant des processus de saisie et d'édition des données brutes recueillies entre novembre 2001 et mars 2002.

22. À partir de ces données, on a obtenu un premier niveau d'agrégation en tenant compte de toutes les combinaisons existantes des variables étudiées et en calculant différentes mesures pour toutes ces combinaisons. Ce niveau correspond à ce que l'on appelle un tableau à N-entrées.

23. Le fichier étant très lourd et difficile à interroger directement, un processus d'agrégation se révèle indispensable. Seuls quelques sous-ensembles de variables sont pris en compte pour les niveaux d'agrégation suivants. Le résultat de ce processus est un groupe d'objets plus «petits». Ils sont tous reliés les uns aux autres dans une structure logique dans laquelle le serveur OLAP peut répondre à n'importe quelle interrogation concernant les variables envisagées.

24. Dans les stratégies de modélisation de données, il est important de déterminer:

a) Comment les variables sont agrégées, en tenant compte de la façon dont elles sont reliées entre elles, de celles qui seront le plus souvent demandées, des relations hiérarchiques éventuelles, etc.;

b) L'équilibre à trouver entre le niveau d'agrégation et les ressources utilisées: un plus haut niveau d'agrégation permet d'obtenir de meilleurs résultats dans les demandes de renseignements mais nécessite plus de temps pour le chargement et l'actualisation des structures des entrepôts de données, ainsi que plus d'espace pour stocker les informations.

25. Dans le projet de diffusion des données du recensement, les aspects suivants ont été dûment établis:

a) Les hiérarchies «naturelles» sont prises en compte (par exemple: âge dans les groupes ≥ âge année par année, niveaux géographiques);

b) D'autres hiérarchies déterminantes sont utilisées pour gérer telle ou telle variable ou rubrique du questionnaire en fonction des différents niveaux de détail de l'information (par exemple, études de troisième cycle, titre décerné à la fin des études, niveau d'activité économique selon la nomenclature NACE);

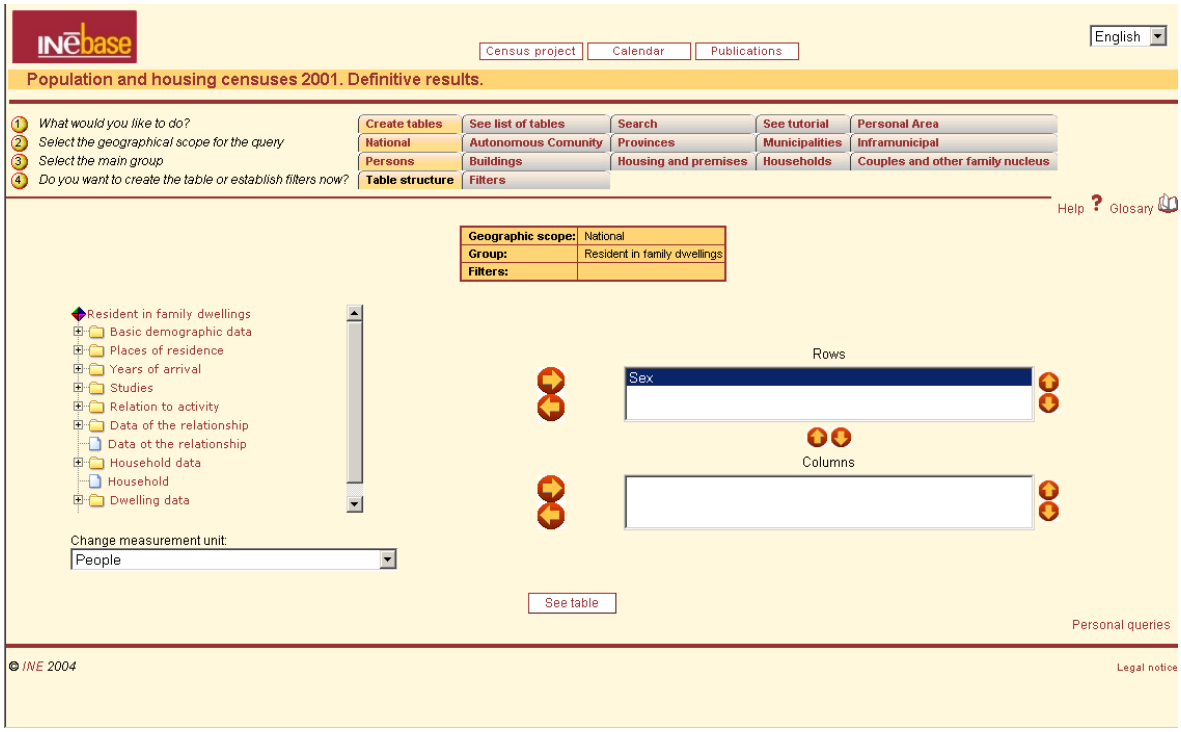
c) Des groupes thématiques de variables ont été constitués lors de la phase de modélisation des données, et ce pour deux raisons: un problème de cardinalité se pose lorsqu'il s'agit d'évaluer toutes les combinaisons possibles de l'ensemble des variables si celles-ci sont considérées comme indépendantes; les variables associées par sens pourraient probablement être consultées conjointement (par exemple, études des parents: la mère et le père sont considérés ensemble dans le modèle).

26. En construisant le modèle multidimensionnel, il faut prévoir une application permettant de tirer parti de toutes les informations qu'il contient.

V. CONCEPTION DE L'INTERFACE

27. Pour disposer d'un outil à la fois puissant et simple à utiliser, il convient de prendre en considération les aspects suivants:

a) Tout d'abord, l'assistance offerte par le système doit être orientée par des questions simples lorsque l'utilisateur commence à communiquer avec lui. Par exemple, la demande «Je souhaite poser une question sur la dimension du territoire/concernant un groupe particulier (personnes, bâtiments, logements)...» oriente les utilisateurs vers un écran sur lequel ils peuvent choisir la (les) variable(s) à disposer sur les lignes et dans les colonnes du tableau; un autre écran auxiliaire est utilisé pour définir les filtres (conditions) à appliquer à la demande.



b) Lorsque la demande est définie, la fonctionnalité la plus importante apparaît sur l'écran où sont affichés les résultats. Tout d'abord, les utilisateurs peuvent vérifier si l'interrogation est comprise correctement par le système, car toutes les sélections effectuées sur les écrans précédents figurent dans un cadre. Une barre d'outils est affichée, ce qui permet aux utilisateurs d'exporter les résultats vers différents modes de présentation bien connus, d'obtenir des représentations graphiques des données affichées (pyramide des âges, cartes, diagramme à barres...), d'ajouter une variable géographique, de modifier la variable analysée, de revenir à la définition du tableau, etc. Le tableau de données contient des éléments dynamiques dans l'intitulé de toutes les lignes et colonnes, ce qui permet aux utilisateurs de naviguer parmi les informations. Ils peuvent détailler, étendre, dérouler les dimensions présentées dans le tableau en cliquant simplement sur la catégorie choisie. Ils peuvent également modifier les variables présentées s'ils s'intéressent à un aspect particulier. Métadonnées: pour mieux comprendre les informations, les utilisateurs ont la possibilité de chercher dans le glossaire la signification d'une variable ou d'une catégorie affichée ou, dans le cas de zones de dimensions restreintes

par exemple, d'en demander la définition ou de consulter une carte permettant de la localiser à l'intérieur d'une ville ou d'une province.

c) Restrictions: Même si le système peut gérer des interrogations très complexes, il a fallu prévoir certaines restrictions: trois variables au maximum peuvent être emboîtées dans les lignes ou les colonnes d'un tableau. Un tableau contenant plus de variables deviendrait difficile à lire. L'interface se limite à des tableaux comportant moins de 10 000 cellules. Il n'est guère commode de transmettre et d'afficher sur Internet des tableaux plus importants en raison du temps que cela prendrait. Le cas échéant, les utilisateurs sont invités à remplir un formulaire de demande et à télécharger les résultats de leur interrogation à partir d'un site ftp; certaines interrogations ne peuvent pas être formulées en ligne: pour protéger le système en cas de calculs excessivement complexes, il est proposé aux utilisateurs de se procurer un fichier à télécharger à partir d'un site ftp lorsqu'il faut du temps pour traiter l'interrogation; il n'est pas donné suite aux interrogations portant sur plus d'un million de cellules. Cette règle vise à protéger le système contre des demandes d'information démesurées provenant d'un seul utilisateur, afin d'éviter que ce service public ne fasse l'objet d'un usage abusif. Interrogations spécifiques: lorsque les utilisateurs ne savent pas comment procéder pour formuler une interrogation spécifique, ils ont la possibilité de décrire leur question. Ces interrogations sont traitées par un groupe d'experts à l'aide d'Enterprise Guide comme outil de base. Restrictions applicables à l'information pour des raisons de confidentialité: ce sujet est traité de façon plus détaillée ci-après.

d) Autres fonctionnalités: comme on l'a vu, le système offre également une vaste collection de plus de 10 000 tableaux de données classés hiérarchiquement selon des critères géographiques et par sujet afin de satisfaire les utilisateurs traditionnels du système de diffusion. Un outil de recherche permet de localiser facilement une variable voire une catégorie précise.

VI. CONFIDENTIALITÉ DES DONNÉES

28. À la différence du mode de diffusion classique de tableaux préétablis, où chacun peut être contrôlé du point de vue de la confidentialité, le système permet de combiner des informations selon deux ou trois variables ou plus et de formuler des interrogations successives, ce qui pourrait conduire à l'identification de caractéristiques individuelles.

29. Un ensemble de principes a donc été établi afin de protéger le caractère confidentiel des données. Ces principes n'envisagent pas la confidentialité de façon excessivement restrictive, car cela pourrait limiter considérablement les possibilités de diffusion des informations relatives aux zones géographiques restreintes, qui comptent parmi les données de recensement les plus intéressantes et les plus demandées.

30. En bref, les principes de confidentialité sont les suivants:

a) Les distributions à une variable peuvent être diffusées quel que soit le niveau de désagrégation géographique car elles ne révèlent pas d'informations individualisées à moins que d'autres caractéristiques ne soient déjà connues;

b) L'identification d'unités individuelles est considérée comme une violation du principe de confidentialité uniquement si des informations individualisées exactes sont révélées.

La présomption que certaines données publiées pourraient correspondre à telle ou telle personne n'est pas assimilable à l'identification de ladite personne;

c) Les informations diffusées comportent un certain degré d'incertitude, qui est dû aux données manquantes, à l'imputation statistique, etc.

31. Compte tenu de ces principes, les règles suivantes ont été appliquées aux interrogations afin de protéger la confidentialité des données:

a) Le nombre de variables comprises dans une interrogation est fonction de la zone la plus restreinte à laquelle elles se réfèrent, explicitement ou implicitement. Quatre niveaux de population, comportant chacun un nombre de variables limité, ont été déterminés:

Population	Nombre maximal de variables comprises dans une interrogation
Jusqu'à 100 habitants	1 variable
De 101 à 5 000 habitants	2 variables
De 5 001 à 20 000 habitants	3 variables
Plus de 20 000 habitants	Illimité

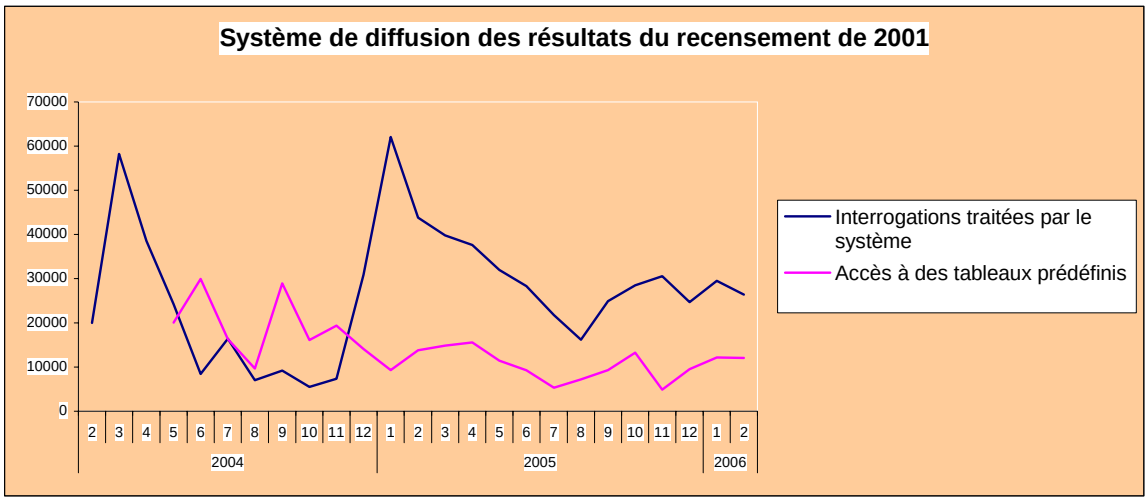
b) Pour certaines variables, des valeurs très détaillées sont groupées lorsque l'on descend dans les niveaux géographiques. Par exemple:

Code d'occupation jusqu'à deux chiffres	Tous les niveaux de population
Code d'occupation jusqu'à trois chiffres	Plus de 20 000 habitants

c) Certaines variables ne sont pas disponibles au-dessous d'un seuil déterminé. Par exemple, la nationalité n'est indiquée que pour les zones de plus de 100 habitants.

VII. EXPÉRIENCE FOURNIE PAR LE SYSTÈME DE DIFFUSION

32. Plus d'un million d'interrogations, comprenant tant les questions formulées que des consultations de tableaux prédéfinis, ont été traitées. Les utilisateurs ont apprécié d'emblée le système et, avec le temps, la plupart d'entre eux optent pour une interrogation directe du système plutôt que pour la recherche d'informations dans les tableaux préétablis. Le graphique ci-dessous montre l'évolution des interrogations mensuelles.



33. Les données concernant des zones restreintes (inframunicipales) ont été publiées en décembre 2004, mais uniquement sur le Web. Le système a été fortement sollicité car les utilisateurs attendaient ces informations. Des produits spécifiques ont été créés pour satisfaire les utilisateurs qui exploitent intensivement ces données.

34. Le temps de réponse est resté dans la fourchette prévue: 75 % des interrogations sont traitées en moins de quatre secondes et 90 % le sont en moins de 11 secondes.

VIII. DÉFIS À RELEVER

35. Le transfert des données du recensement de 1991 dans le système est en cours. Cela permettra de fournir bien plus d'informations qu'auparavant.

36. L'INE d'Espagne s'attache actuellement à incorporer une dimension temporelle dans les données de façon à pouvoir effectuer des comparaisons entre les recensements. Des difficultés restent à surmonter car il faut trouver un niveau de métadonnées commun pour pouvoir comparer des caractéristiques identiques dans des opérations différentes. D'ici à mai 2006, il sera possible de comparer les données des recensements de 1991 et de 2001.

37. Le système de diffusion s'étant révélé efficace, d'autres opérations statistiques pourraient être intégrées à brève échéance à cette plate-forme.

IX. LEÇONS TIRÉES DU PROJET

38. Pour promouvoir le projet, il importe au plus haut point de pouvoir compter sur un groupe de personnes enthousiastes, en envisageant le point de vue des utilisateurs afin de prévoir un éventail complet de fonctionnalités et en assimilant bien la technologie pour atteindre les objectifs fixés dans les délais. Trois départements (recensement, diffusion et calculs) ont directement participé au projet.

39. Le recensement est le plus vaste projet qu'un service de la statistique puisse entreprendre. Un tel projet dispose d'un budget considérable, d'où les fortes pressions exercées pour obtenir des résultats rapides et détaillés. Sans doute aurait-il été souhaitable de débuter avec des projets plus restreints, mais l'envergure de ce système a permis d'explorer les nombreuses possibilités techniques qu'offrent les systèmes de veille économique.

40. Les processus d'édition des données sont nettement plus importants que dans un système de diffusion statistique classique fondé sur des tableaux. Il faut veiller à l'intégrité et à la cohérence de toutes les données car n'importe quelle interrogation peut être formulée. Lorsqu'on publie un ensemble de tableaux limité, il faut être très attentif aux données qui y figurent.

41. Des règles claires doivent être établies en vue de faire respecter le caractère confidentiel des statistiques, les informations disponibles pouvant être analysées de multiples façons.
