UNITED NATIONS **E**

Economic and Social Council

Distr.
General

ECE/CES/2006/29
3 April 2006

Original: ENGLISH

ECONOMIC COMMISSION FOR EUROPE **STATISTICAL COMMISSION**

CONFERENCE OF EUROPEAN STATISTICIANS

Fifty-fourth plenary session
Paris, 13-15 June 2006
Item 6 of the provisional agenda

SEMINAR ON POPULATION AND HOUSING CENSUSES
SESSION III

Data warehousing in census dissemination[1]

Submitted by the National Statistical Institute, Spain

I. ABSTRACT

1. The information systems based on the new technological approach - Business Intelligence, can significantly improve the use of data collected in large-scale statistical operations, such as the Population Census. The main component of this system is the data strucure where information is stored in such a way as to facilitate data queries rather than data processing. Data warehouses are designed for this purpose, based on denormalized models to get the best performance.

2. Users interact with the system via a simple but powerful interface. The complex data structures are hidden from users. They only need to choose variables and conditions to obtain results. With On Line Analytic Processing (OLAP) tools, users can find information according to any aspect of interest. Furthermore, this system makes it necessary to revise the statistical confidentiality and data editing processes.

[1] This paper has been prepared at the invitation of the secretariat.

II. CENSUS 2001: A TECHNOLOGICAL CHALLENGE.

3. The National Statistical Institute of Spain (INE) relied on a strong technological component throughout the whole census process. Main issues to be pointed out in this relation are:

(a) questionnaires were pre-filled with personal data from the Population Register; they were carefully designed to facilitate digitalisation with a high performance system that allowed to finish processing most of the information within three months after collection;

(b) the possibility to fill in questionnaires through Internet was available for the whole population (first country in the world to implement this option for the whole population).

4. The organization of disseminating the census results had to take into account the following considerations:

(a) Internet would be the main dissemination channel. Books and electronic products could be also distributed because part of users are familiar with these products. Internet access would be free of charge and not pose particular requirements on clients concerning communication speed or technology;

(b) dissemination should offer as much information as possible for both general and expert users;

(c) earlier experience showed that many ad-hoc requests for information had been made. These created a problem because of the extra workload on technicians. Such requests should be reduced to a minimum and provided for a fee.

5. A traditional dissemination system based on a collection of tables could not satisfy these requirements. The new approach would be a "self-service system" where users could design the tables based on their own needs.

6. Consequently, the dissemination system has to have the following characteristics:

(a) the "self-service system" should be capable of solving most of users' queries in less than 10 seconds because with the longer response time users may abandon the program;

(b) the interface should be very intuitive and user friendly;

(c) for the users who were accustomed to traditional dissemination methods, one subsystem should generate a set of predefined tables with the most relevant information.

7. Particular characteristics of the census operation (a huge number of variables, highly detailed level in geography) made it too complex to be treated with traditional relational databases. The combination of all these elements defined the solution: to use the multidimensional data modelling techniques and OLAP tools to build a successful system. Now these tools are included in the Business Intelligence concept. INE Spain did not have any previous experience with these systems and therefore their implementation was a challenge.

III. EVALUATING THE ALTERNATIVE SOLUTIONS AND CHOOSING TECHNOLOGY

8. Experience with using such information systems can be found mainly in the private sector where different aspects of business are being continuously analyzed. There are at least three important differences compared to the statistical office's work:

(a) these systems have powerful updating mechanisms because of the need to maintain very dynamic information (sales of a product, share prices) and to be able to inform the managers about what is happening in the company at any moment;
(b) ratios and reports of interest are well defined and only those aspects are stored for analysis;
(c) users of information can be trained in a specific environment on using the systems and operating data.

9. In contrast, for the statistical dissemination process:
(a) updating time is not a critical factor because statistical operations are not producing results continuously. In particular, census is done once every 10 years;
(b) many kinds of users have to be considered and their different interests satisfied;
(c) user interface has to be intuitive because it is not possible to train all possible users of the system.

10. Because of the need to offer a universal service, independent of user's technology and without heavy downloads of software (like applets, clients, etc.), the user interface has to be designed specifically for this environment.

11. So, two important considerations had to be taken into account in choosing the technology:
(a) response time to queries: more than 5 seconds was considered undesirable;
(b) existence of tools allowing to make a personal design of the interface to get an intuitive and powerful user application.

12. The technology provided by SAS Institute in Business Intelligence (BI) Components was chosen because:
(a) the software is hardware independent, offers an end-to-end solution and access to a wide variety of data sources;
(b) it is a known technology for INE, a significant amount of people work with SAS tools;
(c) it showed best performance in the pre-test;
(d) development and deployment of BI/Data Warehouse solutions seemed to be easy and quick because of a set of well-developed tools.

13. This last point was tested before the decision was taken. In the request for proposals to different companies, INE asked for a prototype of the system with five variables. In a few days SAS consultants developed a fully functional system with very good performance.

14. The key elements of the system architecture are:

A. Software

15. Scalable Performance Data Server (SPDS): a database system owned by SAS. It is query-oriented and capable of running parallel processes and work with partitioned objects. It uses modern b-tree indexes.

16. OLAP Server: This component manages the multidimensional structures in a HOLAP (Hybrid OLAP) model, that is, not only data tables as in a relational database system, but hypercubes or multidimensional objects. It provides a simple logical view to other processes and acts as a proxy redirecting the queries to the minimum object where it can be solved.

17. Warehouse administrator: the data-modelling tool. It allows to define all the structures in a graphical-intuitive environment and generates SAS code needed for ETL process.

18. AppDev Studio/Integration Technologies: interface development tool. It provides a lot of Java beans that make use of SAS environment.

19. Enterprise Guide: an end user tool for querying and reporting.

B. Hardware

20. Production environment: Two Sun Fire 480 computers with Solaris operating system arranged in an active/passive cluster to ensure high availability. These computers are connected to a SAN – Symmetrics from EMC with approximately 2TB of storage space for detailed and summarized objects.

IV. DATA MODELLING PROCESS

21. The data modelling process started from data in flat text files that result from capturing and editing processes of primary data collected from November 2001 to March 2002.

22. From this data, a first aggregation level was obtained considering all the existing combinations in the studied variables and calculating different measures for all these combinations. This level corresponds to the so-called N-way table.

23. As the file is very big and difficult to query directly, an aggregation process is needed. Only different subsets of variables are taken into account for the next levels of aggregation. A group of "smaller" objects are obtained as the result of this process. All of them are linked in a logical structure where the OLAP server can solve any query about the variables considered.

24. In data modelling strategies it is important to define:
    (a) how variables are aggregated: to take into account how the variables are related, what variables are going to be more frequently asked, if there are hierarchical relations between them, etc.;
    (b) balance between the level of aggregation and resources used: the higher level of aggregation gives best performance in queries but requires more time to load or refresh the data warehouse structures and more space to store the information.

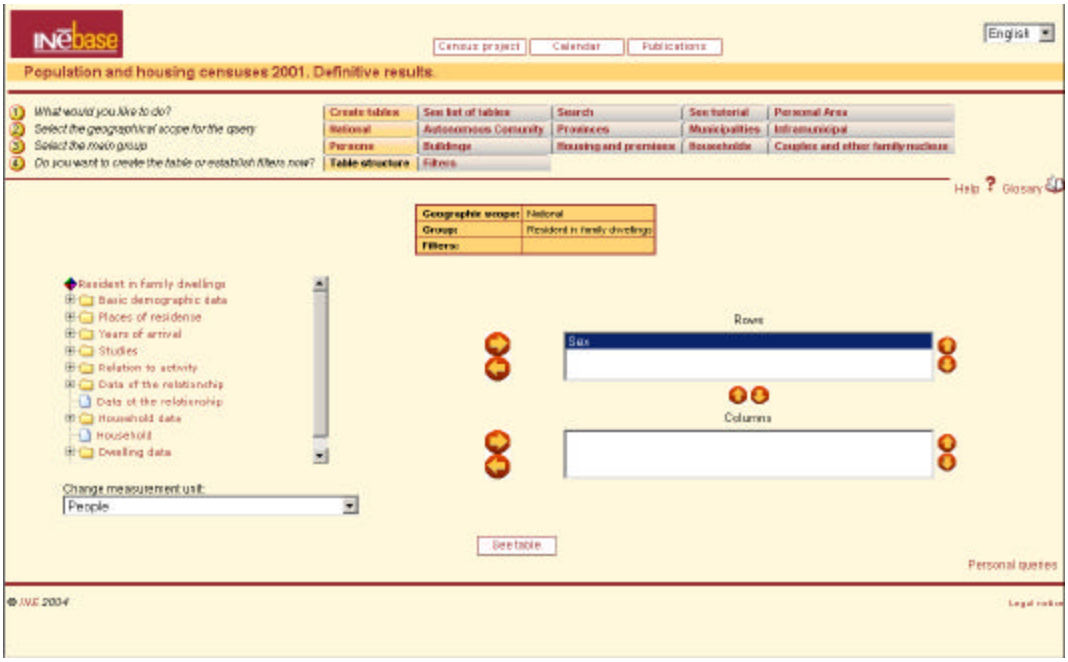
25. In the Census dissemination project, these points were well established:
    (a) "natural" hierarchies are considered: (ex. Age in groups->age year-by-year, geographic levels).
    (b) other instrumental hierarchies are used to manage a variable or item of questionnaire

according to different levels of detail of information (ex. Post-graduate studies, the title you can get from the studies, different level on economic activities by NACE classification)
(c) thematic groups of variables were formed in data modelling stage for two reasons: a cardinality problem to evaluate all the possible combinations of all variables if they were considered independent; the variables associated by their meaning could probably be consulted jointly (ex. Parents studies: mother and father are considered jointly in the model).

26. When the multidimensional model is built, an application should be designed to make use of all information available in the model.

V. DESIGNING INTERFACE

27. To have a powerful but easy-to-use tool, two aspects should be considered.
(a) First assistance should be guided by elementary questions to help the user to start interacting with the system. For example, the request 'I want to make a query/over the territory dimension/referred to a particular group (people, buildings, dwellings)…' brings the users to a screen where they can choose the variable(s) to be disposed on the rows and columns of the table; another auxiliary screen is used to define filters (conditions) to be applied to the query.



(b) When a query is defined, the most important functionality is contained in the screen where the results are displayed. First, users can test whether the query is correctly understood by the system because all the selections made in previous screens are shown in a card. A tools bar is displayed where the users can export the results to different well-

known formats, make graphical representations on the displayed data (population pyramids, maps, bar charts…), add a geographic variable, change the variable analysed, go back to table definition, etc. The data table contains dynamic elements in all row and column headings. By these headings, users can navigate in the information. They can drill down, expand, roll up the dimensions presented in the table simply by clicking on the chosen category. They can also change the presented variables looking for a particular aspect of interest. Metadata: To help to understand the information, users can look up in the glossary the meaning of one variable or one category displayed, or in some cases, like small areas, request for the area definition or view a map to locate it in the city or province.

(c) Restrictions - even though the system can manage very complex queries, some restrictions had to be made: up to three variables can be nested in table rows or columns. A table with more variables is difficult to read; Interface has been limited to show tables with less than 10000 cells. The time required to transmit and show bigger tables on Internet makes it not practic al. In this case, users are invited to fill an application form and download the results of their query from an ftp site; some queries cannot be performed online: to protect the system from too complex calculations, the users are offered to be provided a file to be downloaded from ftp site when a query will have a long time to process; queries with more than a million cells are not executed. This rule has been set up to protect the system from enormous requests that a single user can make to avoid abusing the public service; ad-hoc queries: When users do not know how to make up a specific query, they can describe their question. These queries are treated by a team of experts using Enterprise Guide as basic tool; limits on the information because of statistical confidentiality: This topic is explained in more detail below.

(d) Other utilities: as mentioned before, the system also offers an enormous collection of over 10000 of data tables classified hierarchically by geography and topic to satisfy the traditional users of our dissemination system. Searches allow the easy location of one variable or even a specified category.

VI. DATA CONFIDENTIALITY

28. Unlike traditional dissemination of ready-made tables where each table can be controlled from the confidentiality viewpoint, this system allows to combine information on two, three or more variables and to make repeated queries that may lead to identification of individual characteristics.

29. Therefore, a set of principles have been established to protect the confidentiality of data. These principles do not take a very restrictive approach to confidentiality because it may lead to very limited possibilities to disseminate information on small geographical areas, some of the most interesting census data in demand.

30. In summary, the confidentiality principles are:

(a) univariate distributions may be disseminated on any geographical breakdown level because they do not reveal individual information unless other characteristics were previously known;

(b) identification of individual units is considered a breach of confidentiality only if the exact individual information is revealed. Suspecting that certain published data could

correspond to someone is not considered as identification;

(c) the disseminated information includes a level of uncertainty by nature because of missing data, statistical imputation, etc.…

31. With these principles in mind, the following rules have been applied to the queries to protect data confidentiality:

(a) number of variables involved in a query depends on the smallest area they refer to, explicitly or implicitly. Four population levels have been determined related to which the number of variables are limited:

Population	**Maximum number of variables in a query**
Up to 100 inhabitants	1 variable
Between 101 and 5000 inhab.	2 variables
Between 5001 and 20000 inhab.	3 variables
More than 20000 inhab.	No limit

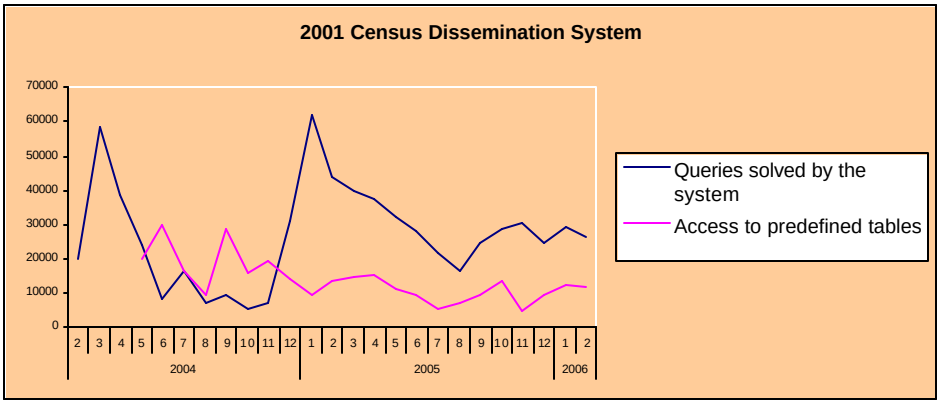
(b) very detailed values of some variables are grouped when drilling down on geography. For example:

Occupation code to 2 digits	All population scopes
Occupation code to 3 digits	More than 20000 inhab.

(c) some variables are not given below a certain threshold. For example: Country of nationality is only given for areas with more than 100 inhabitants.

VII. EXPERIENCE ACQUIRED WITH THE DISSEMINATION SYSTEM

32. More than a million requests have been treated: including both formulated queries and hits on predefined tables. Users have been comfortable with the system from the beginning and, as time goes by, most of them prefer to make their own queries instead of looking for information in the ready-made tables. The graph below shows the monthly queries.



33. Data related to small areas (inframunicipal) were published on December 2004 but only via web. It caused high pressure on the system as users were expecting this information. Specific products have been generated to satisfy those users who work intensively with these data.

34. Response time has been kept in the expected range: 75% of queries are solved in less than 4 seconds and 90% are done in less than 11 seconds.

VIII. NEXT CHALLENGES

35. Uploading data from the 1991 Census is now in progress in this system. It will provide much more information than has been published until now.

36. INE-Spain is now working to incorporate temporal dimension to the data to allow comparisons between censuses. There are some difficulties because it is necessary to find a common metadata level to compare the same characteristics in different operations. By May 2006, it will be possible to compare the 1991 and 2001 census data.

37. As this new dissemination system has been successful, other statistical operations could be included in a short term in this platform.

IX. LESSONS LEARNED FROM THIS PROJECT:

38. It is important to have an enthusiastic group of people to promote the project: to think from the viewpoint of users to have a complete set of functionalities, and to be confident with technology to achieve the planned goals within the deadline. Three departments have been directly involved in this project: census department, dissemination and computing departments.

39. Census is the biggest project that a Statistical Office can undertake. It has an enormous budget and as a consequence, there is a great pressure to get quick and detailed results. Perhaps it would have been good to begin with smaller projects but the big scale of this system allowed to explore the many technological possibilities of BI systems.

40. Data editing processes are much more important than in a traditional statistical dissemination system based on tables. The data integrity and consistency has to be maintained all over the data because any query can be formulated. When you publish a limited set of tables, you have to be alert mainly in the data that you are publishing.

41. It is very important to specify clear rules to protect statistical confidentiality because information can be explored in many ways.

* * * * *