



**Economic and Social
Council**

Distr.
GENERAL

CES/AC.71/2001/8
5 December 2000

ENGLISH ONLY

STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN
COMMUNITIES (EUROSTAT)

CONFERENCE OF EUROPEAN STATISTICIANS

Joint ECE/Eurostat Meeting on the Management of Statistical Information Technology
(Geneva, Switzerland, 14-16 February 2001)

Topic (i): The impact of data warehousing on the management of statistical offices

**THE IMPACT OF DATA WAREHOUSING ON THE MANAGEMENT OF
STATISTICAL OFFICES**

Submitted by ISTAT, Italy ¹

CONTRIBUTED PAPER

**I. THE REASONS FOR THE CONSTRUCTION OF STATISTICAL INFORMATION
SYSTEMS**

1. Traditionally, statistical institutes have concentrated on data collection and processing using single statistical surveys through censuses and sampling surveys. To respond to the new demand for information today, ISTAT looked at the possibility of using an existing survey to gather new information, eventually adding one or more questions to the existing questionnaires. The alternative was to design a new survey, incorporating all the different phases: the definition of contents and sample, the extraction of the sample, dissemination of the paper questionnaires to respondents, data entry, data processing, validation and dissemination of information.

¹ Prepared by Enrico Giovannini, Chief Statistician of OECD, and Alberto Sorce, ISTAT.

2. This approach, in the presence of an increasing demand for statistical information and, at the same time, the dissatisfaction of respondents with the increasing response burden call attention to the weakness of control of the process, duplication of costs, etc. These weaknesses become even more apparent with the decentralisation of the statistical functions to different subjects (local and government bodies, etc.).

3. On the other hand, the development of the administrative registers (fiscal registers, social security registers, etc.) and innovations in computing technologies facilitate the use of a huge volume of administrative information for statistical purposes. This possibility, besides opening new perspectives for statistical analysis, increases the risk of providing inconsistent data on the same phenomenon and confusing users.

4. The necessity to create effective tools in order to co-ordinate the statistical activities within and, in the case of decentralised statistical functions, outside the statistical institute has pushed many statistical institutes to search for a new organisational scheme based on the use of the new computer technologies and statistics, a kind of e-statistics. This implies the redefinition of the production process to make it:

- ◆ efficient, i.e. to minimise (or avoid) duplication within the institute and further reduce the statistical burden on the respondents;
- ◆ flexible, i.e. capable of being transformed on the basis of changes in external conditions or in new demand for information;
- ◆ effective, i.e. enable the production of new categories of information.

5. From an organisational aspect, the first task is to define the areas of production inside a national statistical institute (NSI). On a European level, a certain analogy in the organisation of the institutes of statistics is noticeable, in terms of both diversity and in cultural and institutional terms. Such an analogy (applying to Eurostat as well) can be quite justifiably applied to the evolution of the European regulations and norms in the field of statistics. In fact this norm, particularly in the economic field, has evolved to become the basis of a statistical information system (SIS). The regulations applied to statistical units concerning classification of economic activities, registers of units, structural business statistics, short-term indicators and national accounts, as well as other sectors such as tourism, transport, etc., tend by now to determine the configuration of concepts, definitions and classifications, which represent the building blocks of a complex system of statistical surveys and data processing for statistical purposes.

6. This homogeneity of approach is also due to the similar conditions in various countries where NSIs are working. Among these factors, we can mention:

- ◆ the limited resources, both in financial and human terms, to face the increasing demand for information;
- ◆ the above-mentioned dissatisfaction of respondents with supplying the same information to different administrative and statistical agencies which pushes for improving the use of administrative data for statistical purposes;
- ◆ the rapidity with which the information circulates, also concerning the methodological and scientific aspects of statistics;
- ◆ the dissemination of standards in the development of software, and in statistical methodology.

7. The prevailing orientation is related to the type of statistical activities (short terms surveys, structural surveys, etc.) and not based on thematic issue or group. Such an approach aims to maximise the efficiency of the process, to reduce the pressure on the available resources. In some cases it can require a real revolution in the pre-existing assets. Besides that, it requires structures to integrate the statistical

information across sectors imposing a "matrix" type organisation. Management of such an organisation requires commitments and investments in terms of human resources, training in teamwork, interdisciplinary approach, etc.

8. The answer to such demands is the development of integrated information systems. According to the UN (1999)², "the systems approach is a general human approach for describing, analysing, and controlling complex phenomena. Some basic propositions of the system approach are:

- ◆ a complex phenomenon can be conceptualised as a system, the so-called unperceivable system, since it cannot be fully understood by a single mental act;
- ◆ a system consists of parts;
- ◆ a part of a system is in itself another system, a subsystem of the former system;
- ◆ any system, even the whole, phenomenon first considered, is a part of a wider system, a supersystem or environment, of the former system;
- ◆ the parts of a system are related to each other, and to the system as a whole, and the system is related to its parts as well to other systems in its environment."

9. Given the heterogeneity of the productive process and the thematic processes developed inside a NSI, a SIS is normally composed of subsystems permitting the collection, processing, storage, analysis and diffusion of statistical data. Every subsystem develops a specific task, but within a unique vision that easily allows the interchange and the comparison of data produced by each statistical subsystem.

10. Apart from the immediate aim to provide complex statistical information for users, a SIS has a more general scope, i.e. the creation of a database which provides services to the internal NSI users. In this sense, the SIS becomes a common infrastructure of multiple sources or processes, the expression of which should be coherent with the architecture of the organisation of the NSI. Thus, the SIS can exploit the possibilities offered by its scale and to clearly define the responsibilities of the different subsystems and the relationships between them.

11. In comparison to the past, the recent technological developments make it possible to have a common integrating environment. Naturally, with the growing integration, it becomes more difficult to define the "objects" running in the System. The registers of the units, the classifications, the data dictionary are the traditional means to coordinate the System; generalised tools can be included for the management of the surveys which operate based on the information contained in the System. A good example is the use of a generalised software to extract samples. The software marks the selected units in the register of units to coordinate the sampling process and to reduce the burden on the respondents. The example shows how SIS can become a powerful tool of co-ordination of the whole statistical activity.

12. In addition to its already consolidated position of client-server, the network architecture is technologically strong, using new generation software tools, relational databases, internal and external networks, and allowing the integration of the so-called professional computer science with the "user" ones.

The web architecture provides powerful tools for the statistician, able to handle huge volumes of data and permitting the organisation as a whole to coherently manage increasingly decentralised processes.

2 Information systems architecture for national and international statistical offices. Guidelines and Recommendations (United Nations, Geneva 1999).

II. TOWARD A NEW ORGANISATIONAL CHOICE

13. Figure 1 shows a conceptual scheme, drawn by the United Nations, illustrating the new "vision" of the statistical production process within a NSI. Such a vision can be implemented to reorganise the statistical functions by using the new computer tools; that is what we have defined as e-statistics in para. 4.

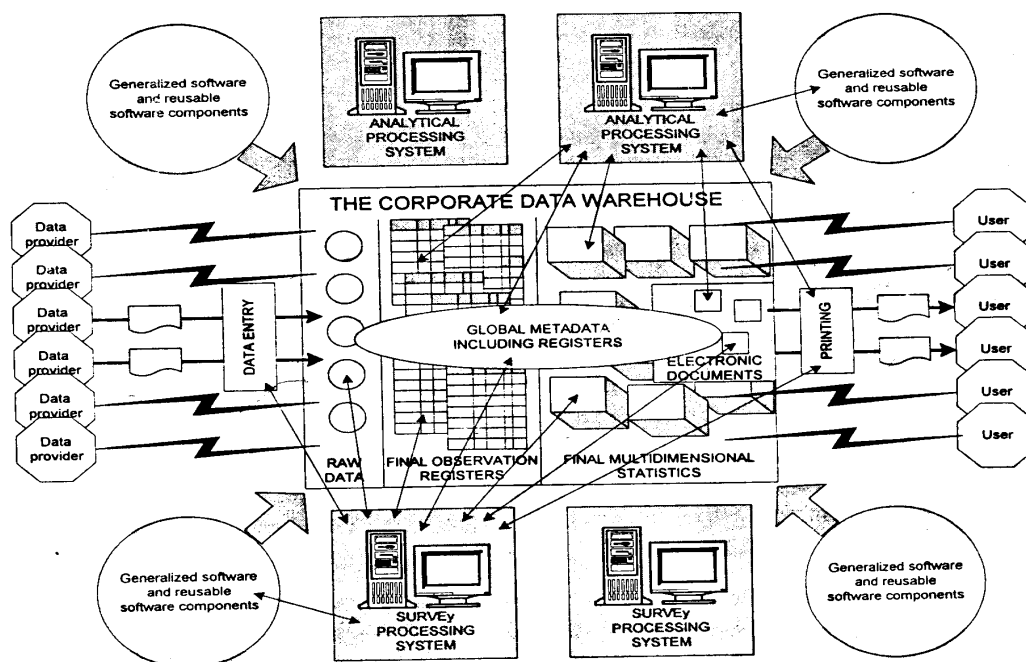
14. The first phase shown in figure 1 illustrates data capture. The information comes through different channels: paper questionnaires, Electronic Data Interchange (EDI), "primary" (through CASIC - computer assisted survey information collection), or "secondary" EDI (that is, through the access to data banks already available in electronic form), etc. Once recorded in electronic form, the information provided by respondents enters the corporate data warehouse (CDW), i.e. "data storage" owned by the NSI, according to default rules in the base requiring a metadata definition for every data item. At this point, every single production process acquires from the corporate data warehouse the required data processes using generalised and specialised software, and sends the data back to the warehouse. From these data are built the so-called "hypercubes", that is, multidimensional structures that allow to access the data according to different keywords (sector, territory, type of variable, etc.), which are used in on-line questionnaires (through Internet or Intranet network) or to derive products in paper or in electronic format (diskettes, CD-ROMs, etc.) for an off-line use.

15. The approach described briefly above is not only convincing theoretically but, as shown by experience, also practically feasible too. For instance, the statistical offices of The Netherlands and Finland have adopted an organisational approach based on the full integration of validated and disseminated information. INSEE in France developed some time ago, at least as business statistics are concerned, a very articulate system based on the integration of administrative register data and statistical data. Following references will be made to the process of deep reorganisation in which the Department for Economic Statistics (DSE) of ISTAT has invested since 1997, together with the development of the Statistical Information System on Enterprises and Institutions (SISSIEI), which is fully consistent with the lines proposed by the United Nations.

16. The DSE, employing 180 researchers out of over 700 persons, is the largest department of ISTAT. This department is in charge of the economic short-term surveys on prices, foreign trade, enterprise activities, employment and labour cost, and structural surveys on agricultural holdings, enterprises and institutions, as well as on the construction of statistical registers of agricultural holdings, enterprises and institutions.

17. In 1997, the DSE adopted a more detailed organisational scheme based on three macro-areas: short term statistics on enterprises; structural statistics on enterprises and institutions; economic censuses and business registers.

Figure 1. An architecture scheme of information system for the statistical organisations



18. The short-term indicators area has been divided into four “Services”:

- ◆ statistics on prices;
- ◆ statistics on external trade;
- ◆ short-term indicators on enterprises’ activity;
- ◆ short-term indicators on employment and labour cost.

19. The area of structural statistics has been further divided into three “Services”:

- ◆ agricultural statistics;
- ◆ structural statistics on industrial and service enterprises;
- ◆ statistics on private and public institutions.

20. The last area deals with the creation and updating of administrative and statistical registers (on agricultural holdings, enterprises, public and private institutions), and economic censuses (on agriculture, industry and services).

21. These three macro-areas are supported by units which coordinate the implementation of information systems, carrying out research in the field of economics and statistical methodology, and responsible for publishing and organisation.

22. The adoption of the "process" approach, rather than by “sector”, in the organisation focussed attention on the improvement of the different phases of data surveying and processing, resulting in more effective use of human resources and more timely data releases. Moreover, the current organisation into

areas follows the pattern suggested by Community regulations and it is easier for the specific Services to meet the Community requirements and co-ordinate relations with Eurostat.

23. A "process" based organisation does not cover all the diverse dimensions of statistical data. Users more frequently ask for integrated overviews (e.g. for labour market) or data on a specific economic sector, and these data can only be provided through a "transversal" reading of short-term and structural data. In order to build these kinds of production processes, each process should store in the corporate data warehouse its own output at elementary and aggregate level linked with the appropriate metadata (definitions, classifications, surveys, questionnaires, etc.).

III. SISSI: A STATISTICAL INFORMATION SYSTEM DEVELOPED IN A DATA WAREHOUSE ENVIRONMENT

24. To transform a classical production system into an integrated information system is an extremely complex operation. On the one hand, new tools must be developed for the management of surveys, data processing and disseminating information; on the other hand, the existing production process must be managed and adjusted to the new information requirements. In 1997, when the reorganization of the DSE and the designing of SISSI (the subsystem for enterprises) began, not all the phases of the process of migration to the new "vision" were defined; neither were the human resources and the professional know-how on a level necessary to start a complete reconversion of the production processes available within ISTAT itself. However, a strategy to radically modify the computing system had already been defined. This involved switching from the mainframe system to a client-server architecture. The following modernisation was foreseen:

- ◆ the use of relational database (Oracle) for the storage of the microdata and the checked macrodata;
- ◆ the use of powerful statistical software (SAS, Speakeasy) for the carrying out of the processes of statistical processing of the data;
- ◆ the use of the Microsoft tools for the office automation.

25. The completion of the project involving the migration of the computer system was predicted for December 1999, taking into account the millennium bug and the introduction of the Euro as currency for the filling in of questionnaires from enterprises and institutions.

26. The decision was taken in March 1997 to proceed with the construction of a relational database (Oracle) in which to store the microdata related to the annual and multi-annual surveys on enterprises. The new statistical register of active enterprises (ASIA), built according to EC regulation n.2186/93 was used as reference. The construction of a second relational database was begun (based on Access and denominated ConIstat), the idea being to assemble all short term indicators produced and updated by the DSE (industrial production, foreign trade, prices, etc.), in order to facilitate access via Internet. In the meantime, the re-engineering of the single productive processes was started.

27. In doing so, all the problems typical of the construction of complex statistical information systems had to be overcome: the management of a huge database, the definition of the structure of metadata (data dictionary, classifications, etc.), the performance of the computing system, the use of the web-server technology, etc. Having determined the structure of the main database for the storage of the microdata and the macrodata, it was possible to provide a "benchmark" for the projects oriented at the reorganisation of the production process, that is the structure towards which all the output had to converge.

28. The dissemination of information through new computer technologies was the particular focus for ISTAT in 1998. During this time, the database ConIstat was also released to users; this database now contains over 9.000 monthly time series updated through Internet immediately after the press release. During the month of December, the database of intermediate census on industry and services, based on a Data Warehouse querying on line via Internet, was released to users.

29. The second phase of the construction of SISSI project was begun in 1998 and was finalized with the development of some generalised tools to manage surveys and carry out the production process. The project, now over 50% realized, forecasts the creation of tools for planning new surveys in terms of defining variables and sample strategy, extraction of samples, checking and correcting of collected data, quality analysis, and dissemination of results, etc. For example:

- ◆ before the beginning of a new survey, it is possible to use software for the survey design, which permits the simulation of different strategies in terms of costs and response burden; the software can check whether a certain enterprise was already covered in past surveys, in order to reduce the response burden (coordinated samples strategy);
- ◆ once the survey is designed, it is necessary to send questionnaires to the enterprises. Because the register is usually based on information referred to 24 months previously, it is possible that addresses of enterprises have not been updated. For this reason, SISSI contains a preliminary version of the register ASIA, in which the identification characters are continuously updated using all the information collected by current surveys, in particular short-term surveys, and administrative sources pertaining to these characters;
- ◆ finally, the operation of sending questionnaires is carried out using generalised procedures, which permits contacting enterprises by mail, by fax or e-mail.

30. To date, SISSIEI is composed of databases related to the following phenomena:

- ◆ ASIA, covers around 3.500.000 industrial and service enterprises;
- ◆ ASIP1, contains around 13.000 public institutions;
- ◆ a first version of ASIP2, related to around 400.000 subject "selected" to be private institutions, which it now under analysis, through a special survey on the field;
- ◆ a first version of ASAIA, related to around 2.700.000 agricultural firms, which it is currently in phase of verification on the field within the 2000 agricultural census activities.

31. The ASIA register (with about 3.500.000 enterprises and 3.900.000 local units) is the main reference register of SISSIEI: it is updated continuously and gives the codes (fiscal code, chambers of commerce code, social security code, etc.) for linking all other statistical information derived from different surveys. In particular, SISSI now integrates the following surveys on enterprises, over the period 1989-00:

- ◆ annual surveys on balance sheets of large enterprises (70.000 units per year) and of small enterprises (50.000 units per year);
- ◆ annual survey on preliminary estimates of balance sheets of very large enterprises (8.000 units per year);
- ◆ occasional surveys on technological innovation of industrial enterprises (5.000 units per wave) and of services enterprises (6.000 units per wave);
- ◆ occasional survey for labour cost (12.000 units);
- ◆ annual survey on scientific research (2.000 units per year);
- ◆ occasional multipurpose survey (300.000 units);
- ◆ monthly external trade statistics (intrastat and extrastat, for about 300.000 units per month);

- ◆ monthly survey on orders and turnover of industrial enterprises (14.000 units per month);
- ◆ monthly survey on retail sales (6.000 outlets per month);
- ◆ monthly survey on employment and labour cost in large enterprises (1.000 enterprises per month).

32. From the point of view of the users, the achieved advantages have been remarkable. As well as the database ConIstat and that of the intermediary census, in the year 2000, Data Warehouse of foreign trade was released on the Intranet. This allows the ISTAT regional offices to satisfy any demand of data processing of import-export data in real time; a reduced version of the Data Warehouse will be released on Internet by the end of 2000. The database on structural statistics of enterprises, from which users, via Internet, can extract the derived data from different annual surveys (on economic accounts, labour cost, etc.) in a fully integrated mode with common classifications and definitions, is expected to be released in December 2000.

IV. CONCLUSIONS

33. As can be seen, the activity oriented towards the construction of the statistical information system on enterprises and institutions at ISTAT has produced numerous and important results, primarily the in-depth modification of the productive process from which the Italian economic statistics are derived. The year 2001 will be the year in which the System will be fully operational. The users will have final integrated databases on: short term-economic statistics, foreign trade, structural statistics on enterprises, censuses on enterprises and annual updating of the productive structure based on ASIA register.

34. After these first results, there remains much to do in three dimensions: the methodological aspect, technological innovation and economic research. The development of a statistical information system particularly requires that the statisticians make a significant investment in the methodology of integration of data deriving from different sources, in order to produce statistical data that exploits all available information. The rich foundation of information available in the administrative sources imposes the adoption of statistical tools able to adequately select and treat the required information.

35. Technological innovation must define the best tools in order to assure the transmission and the processing of the statistical data in full security, entirely protecting the privacy of the respondent. The year 2001, from this point of view, will show a strong increase in ISTAT investments for telematic data capture from enterprises, following the wave of success already achieved by the fiscal authorities during 1999 in the data collection of around 28.000.000 respondent declarations made electronically. Security of access to the corporate data warehouse must be fully guaranteed, given the content of this information.

36. Finally, within the institutional framework, SIS can represent a powerful tool of coordination of decentralised activities within the national statistical system with the participation of local public government agencies. For example, using Extranet it is possible to extract samples from available statistical registers using generalized software, or to use these files related to the official classifications, or to download the collected data in the corporate data warehouse. This is only a further example of how the new technologies and methodological achievements oblige the NSIs to look for new organizational approaches based on e-statistics.