



**Economic and Social  
Council**

Distr.  
GENERAL

CES/AC.71/2001/28  
6 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 (iv): Integration of statistical (survey) data with registers (administrative) data

**INTEGRATION OF STATISTICAL DATA WITH ADMINISTRATIVE DATA**

Submitted by ISTAT, Italy <sup>1</sup>

**CONTRIBUTED PAPER**

**I. INTRODUCTION**

1. Many important factors have induced the National Statistical Institutes (NSIs) of Europe to modify their mode of operation and inside organisation. Among these factors we can name users' requirements of information, the need to reduce the burden on respondents of statistical surveys, the continuous innovation in the field of information technology (IT), and the necessity of a new data capture strategy responding to increasingly segmented demands to expand the field of observation of statistics, timeliness of information, and monitoring of the emergent aspects of economic reality. As well as the continuous demand for information from national accounts (to provide estimations on main sector aggregates), there is an increasing demand by the statistical offices of the European Community, as per community regulations, for statistics with a detailed breakdown according to sector, size and territory, representing economic phenomena that are difficult to measure using conventional statistical surveys.

---

<sup>1</sup> Prepared by Enrico Giovannini, Chief Statistician of OECD, and Alberto Sorce, ISTAT.

2. These developments increasingly require a change of mentality among the NSIs' statisticians, with an important impact on the methodological and organisational framework.

3. The growth of this information at the NSI level, in terms of quality and quantity of data for public and private decision-making and deriving from the process of European unification, raises problems related to the efficiency of the production of statistical information. This is particularly relevant for a country like Italy, where the number of small and very small enterprises is notably high. This imposes further statistical burden on respondents in comparison to countries with a large concentration of enterprises. The greatest problems are those connected to the monitoring of the productive units, characterised by high birth and mortality rates and multiple samples necessary to guarantee the quality of sectorial aggregate estimations.

## **II. OPPORTUNITIES TO RENEW THE EXISTING STATISTICAL SYSTEM ON INSTITUTIONS AND BUSINESS**

4. The new demand for information requires notable efforts for the production of indicators according to a detailed sector, size and territorial level, indicators that would allow to analyse both the structure and the dynamics of business system, co-existence of preliminary time series estimations for a subset of indicators and final estimations for the whole set of variables and sectors. It is therefore essential to find an information source that could be integrated with direct statistical surveys. Such a solution was found using an administrative source (the ASIA archive), that is strategically and functionally accurate and with a controlled quality. It expands the use of administrative sources in sectors characterised by a high number of direct surveys. Access is provided to administrative databases including:

- ◆ Balance sheet Variables (extracted through the Chamber of Commerce network);
- ◆ Data related to employment, salary and wages (extracted from National Social Security Institute's records);
- ◆ Fiscal data (collected by the Department of Income, Ministry of Finance).

5. This data is available rapidly and thus allows to re-structure the surveys on enterprises and institutions based on the integration of statistical and administrative sources, and methodological innovations in estimation of aggregates. Only by means of an integrated use of statistical and administrative data will it be possible both to increase the efficiency of estimations and to reduce the statistical burden on respondents. With the new data collecting methods, this will result in the optimisation of the data related to:

- ◆ Employment, salaries and wages;
- ◆ Balance sheet data of enterprises;
- ◆ Fiscal data related to the direct and indirect taxes (VAT) available in the archives of Finance Ministry.

6. The administrative data can be used on at least two levels:

- ◆ Ex-post:
  - for the production of a preliminary estimation of the principal variables within 10 months from the reference period;
  - for the processing of the missing data (total and partial answers) of the surveys on accounts of enterprises, to provide meaningful results within 18 months from the period of reference;
- ◆ Ex-ante:
  - for the integration of administrative data in the design of the surveys in order to subsequently decrease the sample size of the survey on enterprises within a range of 1-99 employees. This survey would be applied to those enterprises for which data from both sources - Inps and Chamber of Commerce – are

simultaneously available. The missing data would be filled in with variables not extracted from administrative sources. By doing so, the reference populations of single segments of the estimation system must be redesigned.

7. As far as the processing of fiscal data is concerned, we have to consider the problem from a different angle; in this case, the information is substantially exhaustive. The use of administrative data for statistical purposes is a reality today; there are, however, inherent risks associated with such an approach, as is shown by ISTAT's experiences and by the research activities conducted by Italian and international researchers.

8. The increasing use of administrative sources is common to all industrialised countries, and it is strongly recommended to further diminish the statistical burden on respondents. A certain delay can be seen in the use of sources such as fiscal data in Italy. This experience has hindered the use of the fiscal register to establish the Statistic File of Active Enterprises (ASIA) and foreign trade data. The same problem is encountered in the use of tributary and extra-tributary incomes data in the framework of national accounts.

9. The reasons for such a delay are both in the very prudent attitude of statisticians and a certain reluctance of the fiscal authorities to allow full access to their information. Serious delays have accumulated in the past in the management of the declarations of the contributors by the financial administration. So the database can be effectively used only with several years' delay from the reference period. Therefore these data have not been used for the production of current statistics.

10. The situation is, however, changing rapidly. Concerning standards, in 1992 ISTAT and the Ministry of Finance agreed to a memorandum of understanding and common interest in this matter and partially defined the guidelines of the working plan. In reference to Law n. 681/97, the Institute has full access rights to all administrative sources.

11. The rapid development in the last year of transmission telematics for fiscal declarations has placed Italy in an excellent position at the international level, and has resulted in drastic time reduction on the availability of information provided by contributors. In January 2000, for example, the number of declarations reached 22 million (while during February the number reached 26 million declarations). The declarations related to 1998 sent electronically to the Ministry of finance numbered more than 67.000. The structure was as follows:

Decl. form 730	9.332.406	Decl. Form 770	1.017.476
Of which: "unico" physical persons	8.563.791	Of which: annual VAT	236.799
"unico" persons enterprise	771.979	Periodical VAT	1.669.758
"unico" capital enterprise	516.379	Communications: 4/8 per thousand	470.215
"unico" non commercial institutions	57.920		

12. According to first analyses of the transmitted data, the percentage of formal errors has diminished from 28% to 3% and the delay has been reduced from 3-4 years to a few months. In fact, the Ministry of Finance is able to publish the data related to tax declaration of physical persons within one year after submission, and within one and a half years those related to legal persons. Thus, the 1998 data will be available at the end of 2000 according to experiences so far; and the data published in September 1999 were VAT data reported during 1996 and 1994 referring to the declarations of physical and legal persons.

13. According to the actual enterprise distribution in the ASIA archive (built on the basis of different fiscal rules), the following should be noted: of the approximately 3.5 million enterprises with less than 10 employees, 2.4 million respond to the declaration called "Unico" for physical persons, with the compilation of the respective form E and G (simplified accounting) for free lance category; and 500.000 respondents must compile the same "Unico" form corresponding to legal persons. It is therefore evident that it is in ISTAT's interest to not only access the new available sources, but also the potential effect of the reduction of statistical burden on small enterprises is to the Institute's advantage.

14. In these conditions, the acquisition of the data collected by the Ministry of Finance through the so-called "sector studies" will be a great relief. Starting with 1999, with reference to 1998, around 1.500.000 enterprises have begun to compile sectorial questionnaires containing much data of statistical interest, to which, with reference to 1999, another 700.000 contributors are added. Other sector studies were begun this year, so from the beginning of 2000 there will be around 2.200.000 subjects annually completing the forms provided by the Ministry of Finance. In this case, the greater number of declarations are to be provided in electronic format (attached to the form as shown above), so it is expected that the data reported during 1999 will be available at the end of 2000.

15. The second area for which ISTAT has already started collecting data is related to the financial data reported by the enterprises in balance sheets. These balance sheets should be deposited at the respective Chamber of Commerce. There are around 600.000 enterprises which present the balance sheet according to the IVth European Directive, many of which attach the financial statement to the balance sheet as well.

16. During 1998 ISTAT appraised, through a special pilot study, the feasibility of this source for statistical purposes and particularly the importance of the differences which exist between the definitions of the balance sheet data items and statistical variables. The study, carried out on 1.000 enterprises, has shown that 160 of approximately 400 statistical variables collected through the annual survey on accounting systems of enterprises with more than 20 employees can be extracted on the base of the existing schemes of balance sheets. Consequently, ISTAT has started the procedures for the acquisition of the balance sheets in electronic format, realised annually by specialised companies. These companies record the checked data of the balance sheets in optical form in conformity with a general standard before these electronic balance sheet data are submitted to the Chamber of Commerce. From the time schedule point of view, ISTAT hopes to acquire the whole set of balance sheets, starting with those referring to 1999 normally available at the specialised accounting company level, during September-October of the year following the year of reference.

17. The acquisition of data from fiscal sources and annual balance sheets will involve an in-depth review of the organisation of structural surveys on enterprises, speeding up the process begun during 1999. Starting with 1998 as the year of reference, the threshold of enterprises included in the existing annual survey on the accounting system of enterprises has been increased from the enterprises with 20 employees to those of 100 employees, with a net reduction of 60 000 enterprises involved in the survey. For this group of enterprises (20-99 employees), a sample survey has been undertaken using a simplified questionnaire that has been used for the enterprises with a range of 1-19 employees. Starting with the reference year 1999, the questionnaire for the enterprises with 1-99 employees will be simplified, with the perspective of a full integration of data derived from the budgets and from the statistical surveys.

18. Therefore, the availability of more data sources and the necessity to reduce the statistical burden on respondents will force statistical institutes to adopt efficient techniques of integration of administrative and statistical data, and to build an information system that allows the processing of great volumes of data. The advantage in terms of reduction of the burden on respondents, of improved timeliness and quality of the data appears to be remarkable, taking into account the investment in computing technology (IT technology) and, above all, in the methodological field.

### **III. NEW METHODOLOGIES FOR THE DATA PROCESSING**

#### **III.1 The correction phase**

19. The administrative files can not be acquired in the same manner as that used to acquire survey data. The acquisition should be made using telematics or magnetic support. This results in two types of possible errors:

- ◆ conceptual error due to errors in metadata attribution;
- ◆ processing error during the information storage from a communication change or because of a physical defect of the transfer medium.

20. The data, therefore, must be preliminarily analysed, and any errors corrected. The result of this phase is the input for use of the data.

#### **III.2 The phase of consolidation in the database**

21. As has already been mentioned, it is possible to obtain a great deal of data from statistical sources. The main sources are administrative, but other sources can also be of interest, such as, for example, those deriving from the National Professional Associations or from the Chambers of Commerce. Combining the data derived from these sources with those derived from direct surveys creates a very rich statistical information environment. But this environment is not homogeneous. A lot of surveys are sample surveys and can be characterised by a large number of recorded economic variables and a very limited number of respondents compared to the size of the universe of enterprises. The administrative data have a tendency to cover the whole universe, but they have, in comparison to the statistical surveys, few observations. To combine the data derived from surveys with those derived from registers means linking the survey data with the correspondents' data from the registers. This leads to the following situation:

- ◆ The variables measured from both the surveys and the registers consist of very densely populated records, and this whole entity is statistically exhaustive;
- ◆ The variables that are measured only from surveys consist of poorly populated records, and in this subset only partial information is available.

22. Using estimation methods, it is possible to fill in the missing answers with information available in administrative files. The problem becomes complicated when the derived data are combined by the surveys with data derived from more than one administrative register, in this case the same variable can have different values.

### **III.3. Database of reference**

23. To load the data from surveys with those derived from registers into a database, it is necessary to classify the data according to their level of accuracy and source. In order to compile a system of reference, it is necessary to build a second database at microdata level, that contains only final values. Such a database is similar to the first one, self-consistent, and not modifiable. The information contained in this database will be used for dissemination, as a reference tool for the correction phase and for estimation of missing answers.

24. The data collected from the statistical surveys can also be used to check the information derived from different registers for the same variables. It is in fact possible to verify the data derived from the registers through models built with data derived from surveys. The process of integration among the archives will have a positive effect on the quality of statistics. The combination of the derived information from the varied sources will further strengthen statistical output. The integration process reveals the consistence or the insubstantiality of the various sources and presents the opportunity to resolve the latter during a preliminary phase, without waiting for the results calculated by national accounting.

### **III.4 The metadata**

25. The consolidation of all the inherent variables to a specific process within the same database of microdata requires a high degree of standardization of concepts, in particular the classification. This means maintaining the structure of the metadata. It should not be possible to read/write data without the metadata. The metadata should be a system of reference for the database. Each variable can be read or written through its respective metadata definition and this requires having a full exclusive and exhaustive gset of metadata. The metadata, because of the way they have been developed in the Institute, contain traces of their origin from surveys. An information process must now be developed which describes the origin of the data, in order to define the operations and the necessary tools for collection.

### **III.5 The use of the new technologies in the data capture**

26. The traditional technology of data collection used paper questionnaires. Only some statistical surveys on enterprises run electronically: e-mail, electronic questionnaires, interactive forms with a related database for storage. This new method has a notable impact on the methods of work; it not only represents a new tool for the collection of information, but also represents a more effective instrument. If all statistical surveys refer to a unique archive, we could use one source for the data capture, and the respondent would communicate with one organizational unit of the NSI. The same information could then be used to produce additional statistics. A further step would be the transmission of the information of huge archives by telematics. This means using administrative data (mainly public registers), those of taxes (Office of the Department of Internal Revenue) or of the Social Security. The whole process would therefore involve the data exchange process with both respondents and Government bodies. The introduction of the EDI (Electronic Date Interchange) has a great impact on the organization of the work, integrating the organizational units involved in data capture. The following aspect should be pointed out:

- ◆ the quality of the statistics will benefit from this process, in particular regarding consistency of information. The concentration of information in a unique archive emphasises the need for standardization of the statistical concepts and questionnaires;

- ◆ it will have positive effects on the collaboration of ISTAT with multiple respondents by sharply reducing the statistical burden.

27. It is a fact that the paper questionnaires used up to now for statistical data collection are entirely inadequate, both from the point of view of timeliness, and for the lack of a metadata which could be continuously updated, guaranteeing the correct capture of information.