13th Voorburg Group Meeting

Rome, 21-24 September 1998

An informative system for enterprise statistic integration

(SISSI project)

A. Sorce – G. Capasso

Istituto Nazionale di Statistica (ISTAT) - Italy

Abstract

Introduction

Complex statistic surveys often are a simple enumeration of acts and services without that they form an integrated and systematic view of the realty. They are seen only in management function. Seldom, it is considered the need to supply a statistic that documents the survey activity.

Limit of these statistics resides in their scarce systematically. Often they represent a body at themselves. Therefore they are difficulty comparable and seldom they allow an international comparison.

Existence of this heterogeneous statistic activity requires the presence of a reference picture on which inserting the different information. The different surveys actually present in Institute have often their own story and their own autonomy that privileges them inside coherence, also timely, but always these statistics don't answer to homogeneous judgements in the scope of classification, terminology, standards, etc. Often, the aggregation criteria is not respected or the same phenomenon on years old is not defined with the same name. As consequence, the passage from elementary data to macrodata always doesn't allows the following phases of integration and therefore a part of information power is lost.

It has been therefore necessary to define a statistical system able to answer to demands of decisional process, able to inform on consistence and evolution of taken over phenomenon, and able to allows searchers to characterise patterns of behaviour, uniformity and regularity of evolution; besides, information access has to be the easiest possible.

With these objectives, a statistic has to be complete, coherent, trustworthy, checked and controllable, accessible and timely.

Till now, role of Institute has been to produce statistics in a systematic way but with a modest attention to their utilisation: it is been negligent action of goad and of co-ordination of statistics activity. It is to be hoped that this way to work changes. It occurs that gathered information is elaborated in a coherent picture and in fast times, and that gathered results are diffused in reasonable times in way to be able to supply an useful information at all users.Besides, availability of an enterprises archive allows a rigorous planning of the different investigations on field and application of methodologically more adequate procedures of sampling and esteems of taken over data.

The Idea to realise an integrated database arisen by need to have on line to the inside of Institute all available information relative to activity of each single enterprise. This allows,

The system allows the on-line access to the microdata. This characteristic is very important on the checking and validation processes of the single investigations. In fact, the checking programs that will be developed in future they will interface to system by SQL embedded commands: in such way, they will be able reaching all information relative to the enterprise of which are checking findings.

By a product of On Line Analytical Processing (OLAP) it will be also possible viewing on a screen window all indexes and middle values relative to the others firms of the same economic sector, of the same dimension and of the same geographical location. Drill-down and drill-up operation will allow to choose the aggregation level by mouse clicking.

Integration of enterprise surveys

Till now, investigations were single processes, separated the one from each others; in such a system there wasn't conversation and information exchange. Personnel in charge to survey production wasn't able to exploit information gathered by parallel investigations.

On the new Institute informative system, we can distinguish more subsystems:

- the informative subsystems of each single survey, that they continue to be distinguished by central system, but aren't separate;
- SISSI system, that run like collector of all enterprise data, whether anagraphic or economic, gathered by Institute;
- SIDI system (settles informative system of survey documentation) that it contains all logical ties between data gathered by investigations;
- system for the outer information diffusion, constituted by ConIstat.

On the new architecture, SISSI is hearth of the whole system: it acts like collector of data coming from the investigations and like distributor of information for all Institute organs. Its primary job is to allow operators that checks the investigations always to have on line, in real time, all information relative to the enterprise that he is valuing and all information relative to sector in which it operates. The single investigations informative subsystem, in fact, they will be realised with rapid application development tools and the user interfaces will allow to view in a window all information supplied by SISSI.

Figure 1

At the end of check process the survey results go into SISSI for updating of registry data and for saving process of new microdata.

Figure 2

Particular importance for the system running ceils the enterprises registry updating process. With the new system it is possible always to have postponed the enterprises archives, with the possibility to have registry information replied on the basis of the surveys carried out: so it is possible to have on line more addresses and phone numbers of the references for every investigation.

Figure 3

New data analysis operations allowed by the system

The archiving of all data of every italian enterprise allows the Cross-sections analysis in a manner practically snapshot, meanwhile with the old informatic instruments till now available ones this operation expected of a few men/work weeks programming and running times.

The ipercube structure that saves aggregated information will allows implementation of data mining techniques to make further analysis in quick and efficient manner.

Figure 4

The dissemination of data produced by the Institute

The realised system allows an immediate diffusion through all the most modern technologies available today.

Figure 5

From this point of view, standard defined by ConIstat ceils fundamental importance for strategies adopted by Institute. ConIstat is a generalised program for saving and diffusion phase of cyclical statistics distributed to the outer one and utilisable on personal computer. It

allows the table and graphic visualisation of data and their exportation towards the more one widespread office automation products.

Actually the updating phase of ConIstat is based on sequential files but it is already in study a new version able to interface directly to SISSI system for the updating and automatic discharge of data by INTERNET.

ConIstat: introduction

*Con*Istat is a system to manage, record and present statistical data organised in historical series, and in particular in this phase of the project, of cyclical historical series.

It is quite a long time since inside the National Institute of Statistics and both traditional and non traditional users of ISTAT data expressed the need to have data organised in historical series which could be easily managed and retrieved. The typical questions we asked ourselves, and that users ask themselves, are:

- Can I know every detail concerning a specific product?
    - production
    - consumption
    - export
    - etc.
- Can I compare a characteristic of a specific product with characteristics of other products?
- Can I have a picture of a specific period for one or more historical series?
- Can I compare the rough trend of a specific product with its cyclical and tendentious variations?
- Can I supply the collected data to a software for analysis, without any need for complex format translations?
- Can I do all these things without having to learn how to be a skilled computer user, quickly and with a tutorial?

The main goal was that of organising cyclical data in a relational database, develop data loading procedures which could reduce, if not eliminate, possible errors due to oversights and to carelessness, supply a software to navigate simply and intuitively among data. Therefore the effort was focused on data integration through appropriate relations in such a

way that every element could be observed from different viewpoints, so that everything the end user needs is.

Product analysis

We tried to outline the requirements inside the direction and expectations that could arise from ConIstat. In particular, summing up the points that had been underlined:

- at least at the beginning the system should be oriented to the end user outside the Institute, however the whole project should be developed to be easily fitted inside the Institute. Therefore the database should be designed to satisfy internal needs. For example, it is not important that the external user has different versions of same datum, provisional 1, provisional 2, etc. only the definitive datum is important; whereas provisional data are necessary for the internal user. Furthermore confidential data will not be supplied to external users, while these data are to handled inside the Institute, even if complying with appropriate security provisions.
- The system should allow to:
    - search one or more products using its code,
    - search one or more products using the aggregates of which these products are a part,
    - compare more products which belong to the same or to different aggregates,
    - choose the time interval to be examined once products are selected,
    - contemporarily manage different time intervals (e.g. monthly, quarterly, yearly),
    - manage both rough and de-seasonalised historical series,
    - compare series as absolute value with their tendentious and cyclical variation (the system must compute these variations),
    - have the full “visibility” of metadata (which should supply information concerning unit of measure...),
    - have the possibility of quickly plotting one or more historical series,
    - quickly print the examined series,
    - export data to sequential files,
    - dialogue directly with MS-Excel electronic spreadsheet,

- load the values of future periods directly from the user site using automated procedures,
- search the required information using an on line help,
- navigate among data through a graphic user interface according to the standard GUI.

- Moreover the system should be set for future developments, such as:
  - embedding a macro language,
  - adding other data codification system apart from ATECO codification, which represents the codification in release 1.0 of system,
  - Internet portability, so that the database can be navigated with any browser.

Therefore from this analysis the direction for the system general architecture ensued and the following functional blocks were underlined:

a) database

b) data import

c) navigator

The database

The main purpose was that of building a model to study problems connected with the organisation and recording of data in a DBMS, the studied data are part of historical series. In particular statistical data modelling was considered.

The database was built using a multidimensional structure. The observation of a specific element is the basic element. Any observation depends on different factors which determine the n-dimension space. In any case observations are always time-dependent. When observations of the same event, made at different times, are put together we have time series. These series are put together to create a statistical object which is often called statistical table, the database is the result of the union of these tables. Statistical tables have a multidimensional structure and always collect time series with the same structure. Each dimension can assume a set of values; the code lists are these values. Each statistical object (observations, time series, tables, database, dimensions, code lists,...) can have its own text, called “notepad”. A statistical table is a collection time series which are at least structurally homogeneous. This means that all time series which belong to the same table should have the same multidimensional representation.

In modelling the following limits were considered:

- Data can be easily extracted. Using the appropriate interface, users should be able to:
    - display information (e.g. data description), data characteristics (e.g. details on seasonality) and values;
    - perform quality controls (check of missing data, breaking points, etc.)
    - select appropriate quantitative processing (e.g. tendentious variation, cyclical and inclusive variation)
- Database tables should be easily fed in a controlled way, according to requirements of data flow involved.

The logic structure of the database consists of tables grouped by type of economic aggregate; each table represents an economic event, divided into the composing economic activities. This architecture has the advantage of dividing information into aggregates easy to understand, and subsequent new information or update entry would not lead to implementing too big tables, difficult to handle and with not really effective update or query functions.

Each table has the following code structure:

- Domain: set of historical series with the same key structure or the same set of attributes.
- Set: within the same key of domain, it groups one level of the economic activity sector.
- Category: within the economic activity, this code identifies the disaggregation level.
- Type of data: absolute or relative.
- Measure unit: identification code of the measure unit expressed by numbers.
- ATECO91 code: it is the economic activity code (maximum disaggregation of 3 digits) of the examined quantity.
- Year: examined year.
- Period of time: examined period, 1 if yearly, 4 if quarterly, 12 if monthly
- Value: observation value.

Then, within this structure, information is obtained when are known its 7 coordinates.

When implemented, the answer time of a written query to have a time variable, for a Pentium 166platform, is 0.75 seconds.

Product use

The product is supplied in CDs or it can be downloaded from the Internet site www.istat.it. The set up can install this program as stand-alone program or as local network program (an NT server is required in this case). At regular intervals, when connected to the mentioned Internet site, the user can download updated releases of the application program and updated files with the database data.

These are the functions available to the user:

- database management
- database update
- database query

Database management

There are two routines to maintain the database: the first routine is used to compress a defragmented database which occupies a lot of disk space. The second routine is used to repair the database in case it is damaged, as for example when there is a sudden power black-out during a work session.

Database update

At regular intervals, when connected to the Internet site www.istat.it, the user can update the database by downloading files with updated data (usually six times a month) and files with the schedule of update (usually twice a year). The program automatically recognises new files with updated data and it automatically loads them into the database. The system implements the commit and rollback device, so that any malfunction would not affect the integrity of the database itself, because the rollback device can be activated to restore the situation existing before the crashed operation.

Database query

Navigation inside the database is guided and has two modes: by Economic activity Code (Ateco) and by Aggregates:

- search by ateco is done by selecting each time the economic activity Section, then the ateco class and eventually the concerned series
- search by aggregate is done following a guided path, through a structure of sets and categories till the concerned series is reached

With this system, historical series can be compared starting from different Ateco and aggregates.

Searches made can be recorded so that, in the future, the user can retrieve a specific search without going through the database structure.

When the concerned historical series are selected, the relevant time range can be chosen by selecting the year, the beginning and the end of period, and if required cyclical, tendentious, inclusive and cumulative variations. Data are shown in tables.

The user can:

- print selected data, thus data and displayed descriptions can be immediately printed
- export selected data to a sequential file, so historical series can be saved into a file in table format or in a structure with records divided by separators or with fixed length fields
- print charts with selected data, so that charts of one or more series can be contemporarily displayed, saved and printed

Service statistics

SISSI enterprises registry allows the immediate identification of enterprises which operate in services sector. All microdata, historical series and aggregated data analysis and extraction operations can be seen exclusively for services enterprises.

Special investigations directed to findings on services sector are treated by the system in equal manner to the others statistics surveys. In such a way, you can use all the SISSI power.

Discussion and conclusions

The integration of enterprise statistics by a new informative system opens new frontiers in the statistic world. An immediate advantage is the reduction of duplicated information gathered by each survey. The next steps are the study of new aggregations made possible by the system (i.e. cross-section analysis) and the use of new data analysis tecniques available today. The statistical researchers can now use all the system power to make their analysis: their work is limited only by their fantasy.

Bibliography

D. Archer. *Maintenance of Business Registers*. John Wiley & Sons, 1995

W. H. Inmon. *Building the Operational Data Store*. John Wiley & Sons, 1995

W. H. Inmon. *Building the Data Warehouse*. Wiley Computer Publishing, 1996

E. Ohlsson. *The System for Co-ordination of Samples from the Business Register at Statistics Sweden*. Statistics Sweden, 1992

Guido M. Rey. *Le statistiche ufficiali e l'attività della Pubblica Amministrazion*e. ISTAT – Quaderni di discussione, Jan 1984

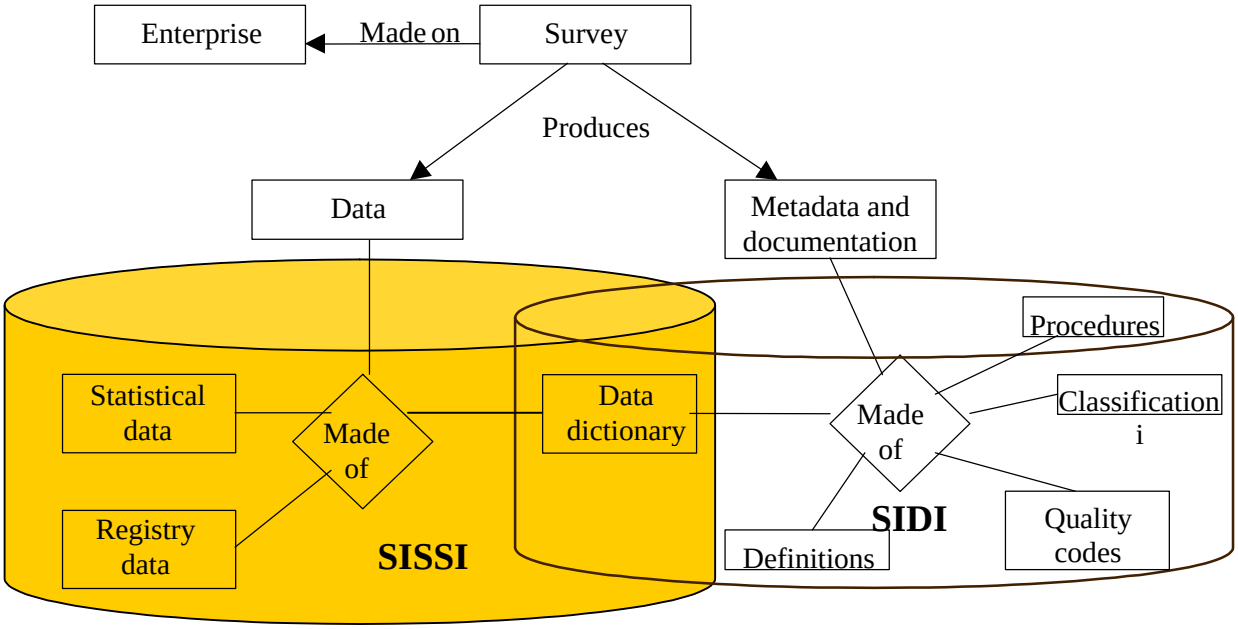

Fig. 1 *- Flow chart of information coming from enterprise surveys*

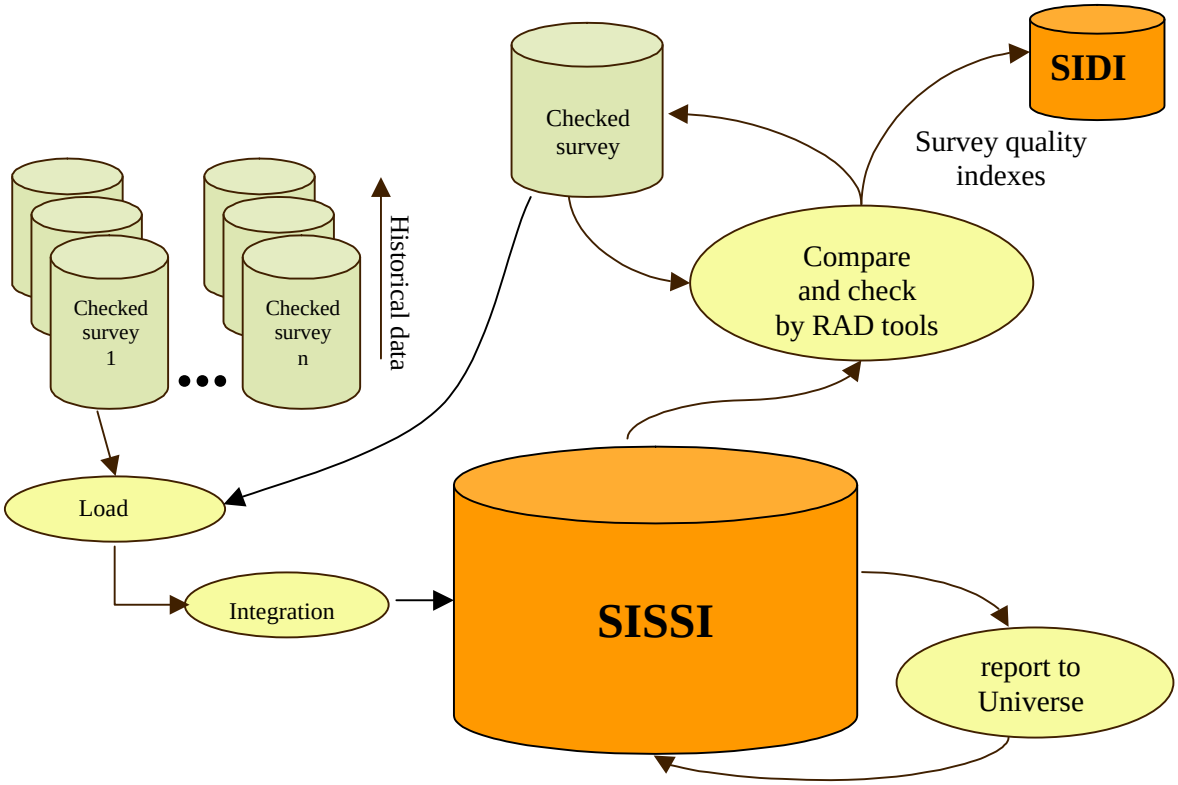

Fig. 2 *- Survey revision and checking process*

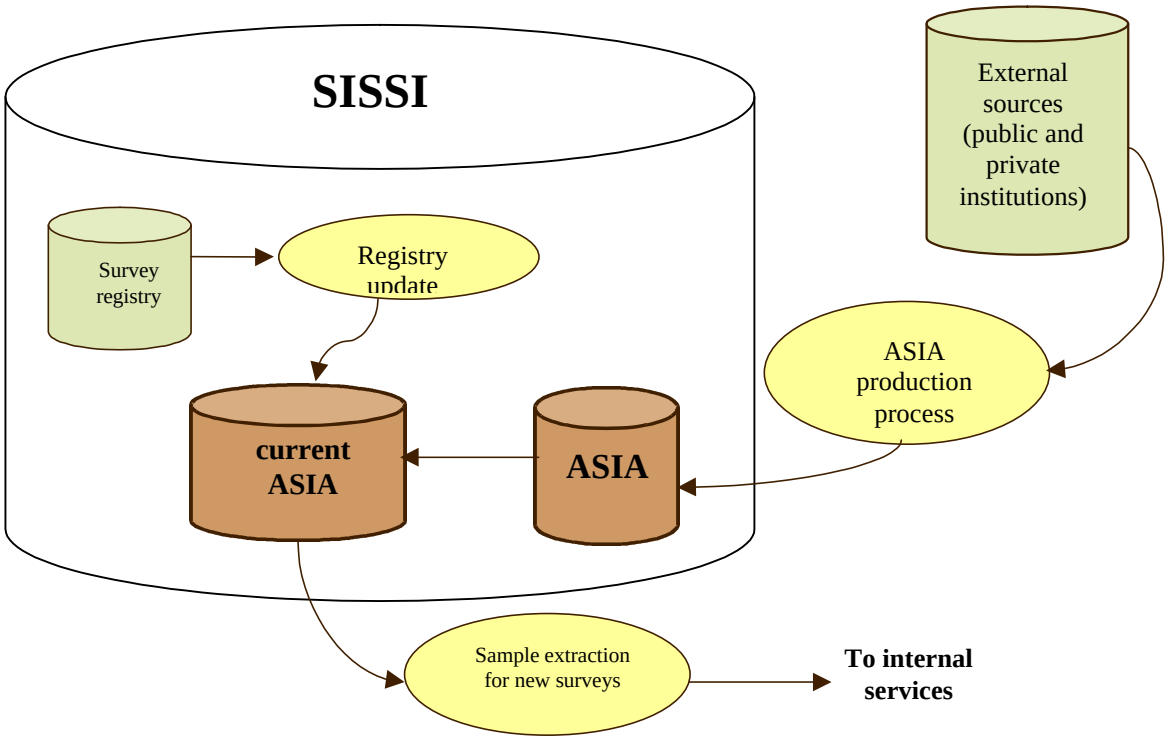

Fig. 3 *- Registry information updating*

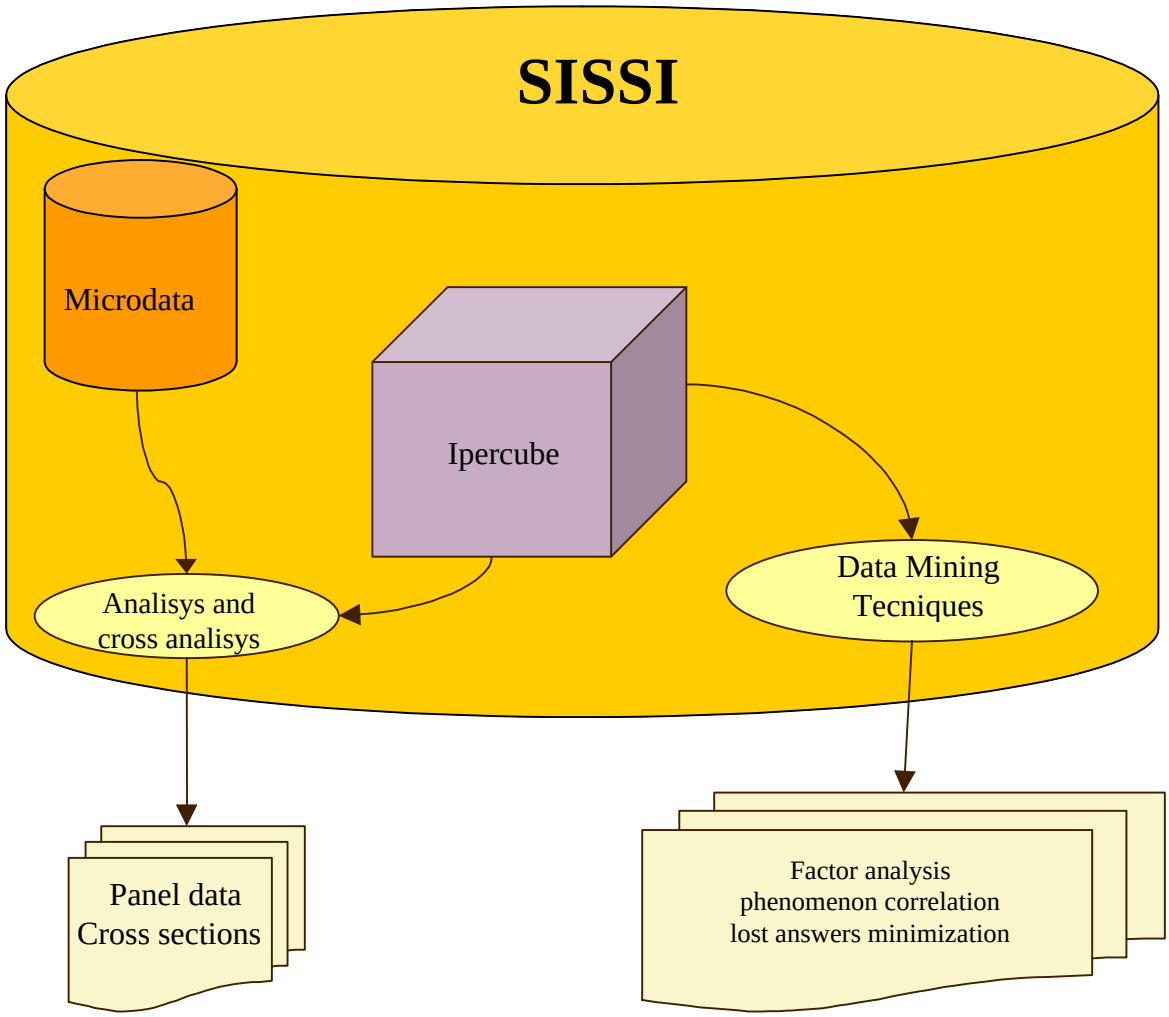

Fig. 4 *- Analysis of data collected by SISSI*

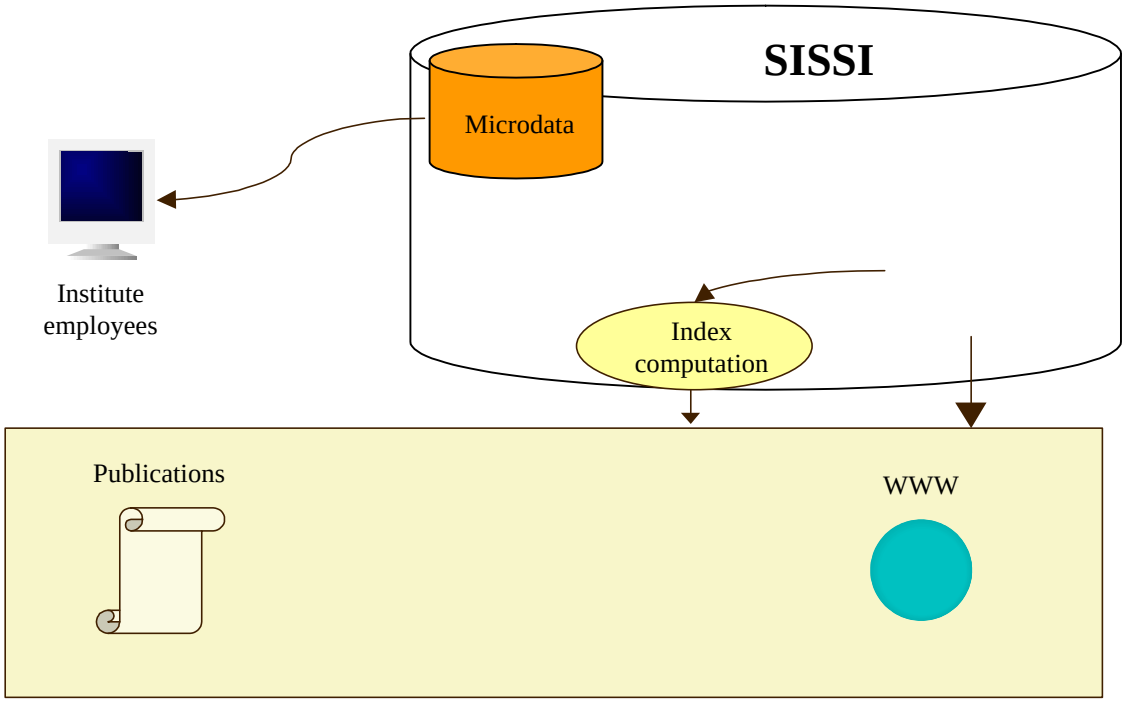

Fig. 5 *- Data dissemination by new technologies*