Integration of different micro-databases – a significant value added for statisticians

Francesca Monacelli[1]

Some preliminary concepts

Micro-data are defined as the information describing the characteristics of the units of a population. With regard to the quantitative data collected by the Bank of Italy, such population can correspond to the reporting entities when the reporter supplies information about itself (for example in supervisory reports or in non financial enterprises direct reporting). Alternatively, it can refer to third parties when the reporter is a commercial data provider/other Authority or when data are derived from operational platforms (*e.g.* TARGET). There is also a third instance in which the reporter produces in the same data flow information about itself and about its related parties as a kind of personal attribute. This is the case, for example, of borrowers' data supplied with Central credit register reports.

Another family of micro-data we deal with are the Registers, which hold the qualitative information of the units of a population. The most relevant are the Subjects Register, which contains the identification variables of individuals, public entities and enterprises, and the Securities Register, which contains the descriptive variables of the securities issued in Italy and of foreign securities held in the banks' portfolios. Both Registers are fed by multiple sources.

Although these categories of micro-data are logically different to one another, our system considers them simply as different specific occurrences of a reported data. In fact our methodological approach and inquiry tools are generic so that we are able to treat them in the same uniform and unique way. For this reason, we indistinguishably call them Reported data.

The same approach and tools are also applied to aggregated/compiled data (*i.e.* macro-data) and, in our experience, uniform treatment of information represents a significant value added for statisticians.

The pursuit in the integration of micro data-bases

So far we have managed to build a single company-wide statistical Data Warehouse for multidimensional data. We are currently in the process of merging into the Data Warehouse also the time series data-base with macroeconomic statistics used in the Economic research Area so that, in the end, all users will be able to compute complex statistics and aggregations, by putting together multidimensional quantitative and qualitative data and time series, using the same tools and Data Dictionary.

[1] Statistics Collection and Processing Department, Banca d'Italia, francesca.monacelli@bancaditalia.it.

The Data Warehouse has been designed to be used for multiple purposes across the Bank (*i.e.* research, supervision, payment systems surveillance, statistical publications, statistical flows for external entities) and for any category of information such as quantitative, qualitative reported data as well as data quality management indicators.

Its distinguishing factors are that it is governed by a unique Data Dictionary with harmonised concepts, it is hosted on the same technological platform and it can be inquired with the same tools regardless of the category data. More in detail, with the same metadata managing software and by means of the transformation rules present in the Data Dictionary we are able to compute any complex statistical multidimensional or time series output on the basis of the different Reported data. The output can be additional Data Warehouse tables with ready to use statistics, a statistical flow for external parties or a publication. The common element of these 3 types of output is that they are generated from the same elementary data.

With regard to ready-made statistical tables which enrich the Data Warehouse, it is worth specifying that they represent an intermediate step between the Reported data and the final statistical production in order to facilitate the user computational tasks. In particular, ready to use statistics enhance the efficiency of the statistical analysis as they save computation time and space. On the other, the existence of data compiled following agreed and fixed definitions ensures that the same concept is homogeneously adopted by all of the Bank's users and in the external dissemination, thus reducing ambiguity.

Furthermore, when different groups of users require that one concept is shown with different facets (*e.g.* with regard to the level of detail/aggregation or because they are authorised to view the data only after a certain time lag) we create additional tables in the same section in order to accommodate these needs. Therefore it can happen that the same concept is shown across the tables of the same section according to different criteria.

It is important to specify that access rights to the Data Warehouse are organised according to strict internal rules based on the "need to know principle" whereby users are grouped according to the institutional area they operate in and are consequently allowed to view all or only some of the Data Warehouse tables.

The Data Warehouse is organised in different logical thematic sections (databases): Supervision, Financial instruments, Central credit register, Monetary policy operations, Money market, Payment systems, Registers etc. Each section comprises tables with the relevant Reported data which can be used in a flexible and personalised way; this represents the foundation layer of the Data Warehouse. In what follows I will describe the solution that we have put in place to create the Central Credit Register [CCR] section which, for its complexity, is a good example of integration of the different elementary data as a way to offer a service tailored to the specific and diversified needs of researchers and supervisors. I will also outline the solution found to integrate these data with other individual information stored in non harmonised Registers.

The internal dissemination of Central Credit Register’s data

The CCR’s section has been built with the aim of managing the following issues:

- efficient elaboration of large volumes of Reported data
- sensitivity of borrowers personal data in terms of privacy protection law
- ensuring that statistics on credit default rates and other commonly used concepts are compiled in the same way by all Banks’ users
- identification variables of the borrowers are separately stored in the Subjects Register
- exploiting as much as possible the information concerning the borrowers characteristics and the relationship lender-borrower.

All of the above has determined the implementation of a rather articulated section based on the CCR reported data which are processed in different ways with a view of satisfying different user need and constraints. It is composed by the following sub-sections, the first 2 focusing on the borrower and the 3rd focusing on the lender:

1. CCR-N. The basic layer contains the data reported to the Central Credit Register; the borrowers are identified by the Subjects Register code. The individual data are also enriched with some ready made indicators. By means of the borrower’s code, this information can be freely integrated by users with the Subjects Register or it can be cross related to other quantitative individual information collected with supervisory reports. Since it is possible to easily identify the borrowers, we maintain a detailed log of every single access to this data. Authorised users are supervisors, researchers and also applications (such as the one producing the Credit register personalised return flows of information, banking supervision models, statistical publications).
2. CCR-A It. is an exact copy of CCR-N, however the borrowers code is an alias and can be associated to a reduced version of the Subjects Register with more limited information which does not allow the identification of any subject specifically. As a consequence no access log is needed.. Since these data are anonymous, in addition to supervisors and researchers, authorised users are also restricted groups operating in other areas which have an institutional interest to analyse these data.
3. CCR-S. It contains several ready to use statistics based on a joint use of the CCR-N section and the Subjects Register focusing on the single lenders or on the credit system as a whole. The browsing tools and the authorisation level of this section are the same as for CCR-A.

Frequently users need to put together information related to the same group of subjects coming from the data-bases present in the Data Warehouse and from other data-bases with individual data acquired from commercial providers and therefore not harmonised. An example of the latter is the Central balance sheet data-base. The problem they face is to identify the same subject in the 2 data-bases when the 2 sources have different metadata and can describe the subjects with slightly different values. This activity requires specific skills and may be very time consuming if applied to large number of entities.

To solve this problem one can decide either to harmonise the metadata of external sources, so as to allow an integrated use with the rest of the Data Warehouse, or to leave the integration to the moment when the data are actually used by supplying the users with an appropriate way to reconcile different sets of metadata. We opted for the second solution by developing a general application that is able to compare the identification variables of any pair of Registers returning the pairs of codes that correspond to the same subject. The comparison is based on the same algorithm used in the Subjects register to assign a code to a new entity. The reconciliation of the unmatched subjects is performed manually by specialised staff.

The integration of the time series data-base

We are currently undergoing a major technical transformation with the aim to achieve a wider integration of data-bases by merging into a unique company-wide platform the 2 Data Warehouses currently available to internal users: one being the one described so far and the other containing macroeconomic information in the form of time series. In particular, this data-base is composed by the time series built on the basis of the Supervisory reports, the ones acquired by national and international institutions (National Statistical Office, OECD, IMF, ECB, etc.) and others created by the researchers as a result of statistical and econometrical analysis.

The project entails two fundamental preconditions. The first is the complete harmonisation of the Data Dictionary although we have decided not to harmonise the metadata derived from external sources, instead, for this data we have created specific tables to link the internal and the external codes. The second is the definition of a comprehensive Information model that not only it is capable to manage different categories of data, their rules and logical dependencies but also can cover the statistical production process as a set of integrated activities (collection, compilation, dissemination). Once the merge will be finalised the production of time series from Reported data will be more efficient, thus reducing the process overheads, and they will be updated at the same time as the elementary data. In addition it will be easier for researchers to analyse time series and other cross section data together.

Final remarks

The unification of different data-bases is not an effortless task but we are convinced that it will give a greater value added to our internal and external users. By means of the same Data Dictionary, same Inquiry tools and same data representation model we will be able to exploit all the micro and macro information held in the future integrated statistical platform to produce reports/analysis in a more efficient manner and, for the benefit of our external users, we will also be able use the same metadata in the statistical outputs, regardless of the category of data they are based upon.