Processing and managing statistical data: a National Central Bank experience

Fabio Di Giovanni, Daniele Piazza, Banca d'Italia[1]

Summary

In a central bank statistical information sustains many institutional functions and is a strategic resource for research and decision-making. To serve this purpose, new statistics often have to be created.
This complex and dynamic scenario demands a comprehensive and flexible IT solution which should support various kinds of processes and be promptly adapted to new requirements.
The solution should be founded on a rigorous and general information model that can describe all the characteristics of the statistical data: the meaning, the properties and the transformation rules to produce other data.
Compliance of this information model with international standards is important, in that statistical production requires a high level of co-operation among all the stakeholders (statistical agencies, users, etc.).

The Bank of Italy has been very active in this field for a long time. The need to conduct a constantly increasing number of statistical surveys, subject to frequent changes, has led the Bank to develop a theoretical vision of statistical processing and to develop software with an increasing degree of flexibility and coverage. In addition, the Bank actively contributes to international bodies that define international standards concerning statistics.

This article gives an overview of the IT system for statistical processing recently developed. The system (INFOSTAT) exploits a well founded information model (called "Matrix") and the best opportunities coming from the ICT technologies.
The first part of the article depicts the general characteristics of the information model. This is a model able to support all statistical activities, from collection to dissemination, and interoperable with other models and standards (e.g. SDMX, XBRL). Furthermore, the main principles concerning the architecture of INFOSTAT are described as well as the way they contribute to meet the users' requirements.
The second part of the article focuses on the capability of INFOSTAT to represent and perform calculations and consistency checks. The system makes ample use of the Matrix metadata, which describe the links between the input and the output data as well as the calculus expressions, using an independent IT language. These definitions can be administered directly by the information workers and are able to drive the IT services allowing, among other things, sophisticated validations on the data collected and support for the production of statistical data with effectiveness.
Finally, some typical cases of the way INFOSTAT is used are provided.

[1] Prepared by :

- Fabio Di Giovanni, Statistics Collection and Processing Department, fabiodigiovanni@bancaditalia.it ("Information model", "Representation of data transformations", "Practical cases of calculations and consistency checks").
- Daniele Piazza, Statistics Collection and Processing Department, daniele.piazza@bancaditalia.it, ("Context and general requirements", "IT Architecture", "Calculation services").

Authors thank all the colleagues who put their knowledge at their disposal, made suggestions and revised the document.

1. Context and general requirements

In the last decades the demand for statistical information has increased considerably, as have the requirements concerning its quality. To cope with the dynamism of real world phenomena, more and more significant, timely and accurate data are needed for analysis and decision making activities, so new statistics have to be generated more frequently or existing ones must change.

Statistical processing systems must therefore be flexible (evolving fast and cheaply), usable and manageable (capable of governing the growing complexity), and effective and efficient (able to process all the information needed in a suitable time).

In addition, as the same statistical information may support multiple decision makers, exchanges and sharing of information are also increasing, driven by phenomena such as the globalization of the world economy, the European integration process, etc. As a matter of fact, national statistical institutes, central banks, and other national and international institutions dialog through an articulated network of information flows.

As a consequence, integration of information exchanged with different sources turns out to be of greater and greater importance. Compliance with international standards can enhance the efficiency of these exchanges. Nevertheless, their relentless growth requires the extension of existing standards in order to cope with a larger set of objects (registers, micro and macro data, qualitative and quantitative data, textual data, graphs, etc.) and to represent the logical dependencies existing among data (e.g. calculation rules).

Even inside a single institution such as a central bank, the same statistical information may support a wide range of institutional functions. For instance, the Bank of Italy's statistical information system supports functions such as economic analysis, central banking, supervision, the payment system, the central credit register, the Financial Intelligence Unit (FIU).

As in the case of different institutions, organizational units inside the same institution can be viewed as nearly independent actors, sharing and/or exchanging information as desired. There is therefore strong demand to some extent in opposite directions: statistical systems must support scenarios in which users act in a coordinated way (harmonizing common data and functionalities) as well as scenarios in which they require independence from all other users.

The Bank of Italy's experience shows that a good response to such challenges requires both organizational[2], methodological and technological measures.

At a methodological level, a key success factor appears to be the adoption of a general model to represent statistical data, their properties (data quality, completeness, etc.) and the processing rules to be applied to them. To promote an integrated use of data, the model should be unique, therefore abstract enough to represent all the needed types of data (arrays and time series, micro and macro, qualitative and quantitative) and independent from the specific statistical subject (e.g. money and banking, payment system, FIU statistics). It should be well founded and formal, so that the information represented by it (metadata) can be managed and exploited by software. Finally, the model should provide a "user oriented" view so that metadata can effectively improve data understanding and be directly managed by metadata administrators.

At a technological level, the success of the system seems to increase provided that it is founded on the following principles. The IT solution should derive from a holistic view of business processes so that it can meet all requirements in a coherent way, avoiding functional redundancy or inconsistencies. The IT architecture should be modular so that the overall solution can be decomposed in smaller and more manageable parts and evolve easily by adding new functionalities or by composing the existing ones in new ways. Software should be fully aligned with the information model so that it can process all the different kinds of information represented. The IT system should actively use metadata (data

[2] Organizational measures are required to promote and sustain the cooperation among the various "actors", involved in statistical activities. However, these aspects are outside the scope of this paper.

characteristics and processing rules) so that the most common changes to system's behaviour should be done by means of metadata update. Finally, the IT solution should be compliant with the most common international standards in order to facilitate data exchange with statistical agencies, interoperability with external systems and integration with statistical packages.

Some of these principles have shown their effectiveness for years; moreover recent developments in the ICT field have allowed the Bank of Italy to build a new and much better processing system (named INFOSTAT). Part of it is already running and showing the expected benefits (the web collection system, which includes the data definition, validation and calculation services).

The very good results obtained with this first part of the implementation have confirmed the validity of the undertaken strategy. Therefore, a second part of the development programme is going to be launched in order to support further business activities (e.g. compilation, dissemination, registers handling) and to gradually dismiss the older software systems.

2. Foundations of the statistical information system (INFOSTAT)

2.1. Information Model

The aim of an information model is to offer a unified and generic vision of the statistical data and the relationships among them (e.g. logical dependencies, processing rules). [3]

Many are the models coming from the IT market and international standardization initiatives that deal with the representation and manipulation of the data, for example the "relational" model of the Data Base Management Systems, the "multidimensional" model of the OLAP tools, the time series models embedded in statistical packages like FAME, the models born for data exchanging (SDMX and XBRL standards) and so on. In building a statistical information system it is necessary to deal with many such models, in order to benefit from using IT tools and to cooperate with other statistical agencies.

Meeting the requirements described in the previous section would require a comprehensive "information model" able to render a homogeneous representation of all the aspects related to statistical data processing. No one such model, however, is able to play this role effectively, because none is equipped with all desired features.

Most of the standards and commercial models lack the representation of time changes (for instance it would be impossible to represent mergers, acquisitions, spin-offs, etc.)[4], do not provide a general solution for the description of the data availability in the information system[5], do not properly support some kind of activity (e.g. consistency checks, sophisticated calculations, estimations)). Furthermore, some models are not designed to give a uniform representation of different data categories (time series, cross sections, mixed structures, registers, etc.) so they cannot be used to describe, manage and integrate heterogeneous data, or lack a proper representation of the levels of detail (aggregates and their relationships, classifications).

It therefore happens that available models focus on aspects related to some specific activity (e.g. reporting, data sharing and exchange) but cannot depict an end-to-end view of complex processes involving a wide set of integrated activities (e.g. collection, compilation, dissemination).

---

[3] It is acquired long since that a good information system is complete and self-consistent, i.e. contains not only the statistical information but also the 'meta-information' (or 'meta-data') that describes the meaning of data, their structure, the rules for managing and administering the system for all categories of users (end users, administrators and software). The way the meta-information is represented within the information system, that is its model, is a key success factor, and must be properly faced in order to meet requirements and exploit all potential advantages (see [Sundgren, 1991],[Del Vecchio, 1997]).

[4] Time change representation is essential for documenting and managing important cases (e.g. codelist changes in time series, as when West and East Germany merged into Germany in 1989) and for properly defining and obtaining aggregated data (for instance where a new member becomes part of the EU from a certain date onwards)

[5] For the representation of the data availability ("knowledge" of the data) see [Del Vecchio 1997].

In this context, the Bank of Italy has designed a model (called Matrix[6]) endowed with all the features needed and easily interoperable with other models.

The Matrix model is derived from mathematical and statistical theory, is designed for use by experts in the subject (statisticians and information workers) and with a view to support all the processing activities regarding statistical information. The model has been developed over many years and has reached a well consolidated and mature structure that makes the information system info-logically complete and, with the latest IT Platform development, also procedurally complete[7].

The Matrix model has been designed with specific consideration given to the major international standards, to enable as much cooperation and integration as possible with other institutions and the international statistical community.

The Bank of Italy is so conscious of the importance of international standards that it actively participates in their design and evolution. In the recent past, for instance, it has made significant contributions to SDMX IM v2[8] and to some XBRL specifications, e.g. the XBRL Dimension. [9]

Matrix metadata can be easily transformed into SDMX and XBRL metadata.[10] Using this feature, all data can be internally represented using a unique model (i.e. Matrix) but other models (i.e. SDMX and XBRL) can be used for data exchange..[11]

Using the model interoperability features, the Bank of Italy also contributed to the COREP and FINREP initiatives, that adopted the Matrix graphic schema to present the XBRL data definitions to users, automatically converting XBRL taxonomies into the Matrix graphic schema.[12]

The Matrix model is able to handle information about both the conceptual meaning of data and their physical representation. In detail, Matrix can represent data stored using the most common IT data models (Relational, multidimensional, time series structures, etc.). This capability is actually used to integrate different software tools in the INFOSTAT, to convert data models during collection and/or dissemination of data and metadata in different formats and to integrate different statistical packages (see below).

Due to its features, the Matrix model (or in some cases parts of it) has recently been adopted also by other organizations.

---

[6] See [Del Vecchio, Di Giovanni, Pambianco, 2007]: “The Matrix Model: Unified Model for Statistical Data Representation and Processing“.

[7] An information system can be defined Infologically complete if it contains as an integral part all the metadata that its users use for correct interpretation of data (i.e. the information system must, in addition to the extensional form of the data, also include their intensional definition and the definition of the concepts used). Moreover, the information system is procedurally complete if it contains, as an integral part, all the metainformation that is used to support software artifacts and information system administration for operating and maintaining procedures. See [Sundgren, 1991].

[8] See [SDMX 2005].

[9] The Bank of Italy also also participated in the European Commission Information Society Technologies Programme, see [Del Vecchio, 2001], [Del Vecchio, 2002], [Del Vecchio V., Froeschl K.A, Grossmann W, 2003], [Del Vecchio 2003].

[10] It is possible, for example, to export/import data and metadata from/to the Matrix format to/from the SDMX format (structure message, different types of data messages) and to/from the XBRL format (taxonomy and instance messages). Using the Matrix model as a “pivot”, it is also possible to export/import an SDMX structure message into an XBRL taxonomy and vice-versa.

[11] For example, the XBRL format is adopted in the balance-of-payment surveys and the SDMX format in the data exchanges within the European System of Central Banks.

[12] See [Romanelli, 2006], [Romanelli, 2007] and Matrix schemas of XBRL taxonomies available at:
http://www.eurofiling.info/corepTaxonomy/descriptions/COREP%20Matrix%20schemas%201.2.4.zip
http://www.eurofiling.info/finrepTaxonomy/taxonomy/descriptions/FINREP%20Matrix%20schema%201.3.zip

2.2. IT Architecture

The Bank of Italy has been very active in the field of statistical processing for a long time. The need to support a high and constantly increasing number of statistical surveys, subject to frequent changes, has led the Bank of Italy to develop software with an increasing degree of flexibility and coverage, able to process different data on the basis of their definitions, so that adding new data or changing the existing ones hardly needs any software maintenance.
A first generation of software applications of this kind became available from the early 1980s onwards. They showed a good support for users and were constantly enriched, so their functionalities grew.

However, even within the same organization, different processes should coexist in order to support heterogeneous requirements. Processes may differ on the basis of several aspects, for example the collection methods (census vs. sample surveys), the type of survey (registers vs. quantitative), the production schedule (systematic vs. ad hoc), the dissemination approach (data bank, publications, etc.) and so on.
Partly because of the limitations of the existing technologies, partly because of the intrinsic difficulty of the problem, the more common response to this problem was to build "silo" applications, each one dedicated to automate a specific process and tailored to ”domain specific“ requirements.
Such a response has several drawbacks. First of all, it does not permit the exploitation of similarities among various processes, often leading to a large number of different user interfaces and operating models. In addition, the relentless pursuit of integration of different subject matters inevitably results in a dense network of interconnections among applications. Over time, especially in the case of strong growth of the statistical tasks, this approach cannot handle the complexity as much as needed and becomes an obstacle to the rapid alignment of the IT solution to the evolution of the business requirements and to the containment of the operating costs.

In recent years, new IT technologies have been introduced, especially suitable for solving problems of this kind. In the meanwhile, the Bank of Italy information model was constantly improved in the light of experience and theoretical results, so that the gap between the theoretical vision and the IT solution grew.

For all these reasons, in 2007 the Bank launched a development programme aimed at building a completely new IT solution for statistical processing. The new system, called INFOSTAT, exploits the most interesting opportunities coming from new technologies, methodologies and tools. The main features and benefits the INFOSTAT architecture is intended to achieve are described below.

2.2.1. Comprehensiveness

An effective and modern IT solution for statistical processing should be based on a holistic view of the working processes and their data, in order to support the full set of users’ requirements by means of a uniform approach, to foster the data integration and to avoid duplicated or redundant functionalities.

The most challenging goal in this scenario is to manage a large variety of data and processes using a unique IT solution. INFOSTAT is succeeding in this challenge using an approach which combines the full exploitation of its information model (Matrix) and the adoption of a Service Oriented Architecture.

INFOSTAT provides a unique logical environment (warehouse) where all data can be stored. The warehouse acts as a collaborative workplace because it provides the input data required for each activity involved in a statistical process and receives the related results.
The warehouse handles various kinds of information which are usually perceived as very different by the users: for example multi dimensional arrays, time series, register data.
The warehouse can include data repositories physically distributed and handled by different tools. INFOSTAT includes full-featured connectors which allow to operate on relational archives stored using the most common RDBMS. Other connectors allow to exchange data with statistical packages (e.g. FAME, SAS) exploiting functions exposed by their external interfaces. In such a way, data can be stored using the tool which best fits the users' needs.
All data handled by the warehouse are represented using the Matrix model, so that the data dictionary can provide an overall picture of all data available within the information system, together with their description and information about the physical location.

Registry services allow to discover where data are stored and how to access them. Specific software components use the information provided by the registry in order to search, retrieve and store data accordingly.

At an high level of abstraction, statistical processing can be described as a supply chain which includes various activities and involves several actors. A generic, end-to-end, statistical process typically includes the following macro-activities: data definition, collection, production and dissemination.
INFOSTAT is intended to provide a common framework for a wide variety of specific statistical processes which can include the activities indicated above as well as other domain specific activities. For example, it can handle typical processes involving the collection and processing of raw data to produce statistical outputs as well as maintenance of administrative registers (e.g. registers concerning monetary and financial institutions, central credit register, securities register).

Service Oriented Architecture plays a key role to cope with this large variety of business processes, avoiding unnecessary growth of software and allowing the complexity of the system to be governed. As a matter of fact, using such an architectural style, the overall IT solution is a combination of coarse grained and loosely coupled IT services, each of which implements a specific part of the business logic. Services are implemented by software components, which in turn can be developed as a composition of other lower grained parts. Components are designed in order to promote their re-usability as well as to facilitate their integration. In this way, the effort required to design and maintain a comprehensive and coherent IT system can be broken down into largely independent activities.

Using this approach any specific statistical process can be supported by means of integration of IT services. The list below contains some of the most valuable services envisaged in the architecture.

- Identity and access management (e.g. user registration, authentication, user profiling)
- Metadata definition (Matrix registry)
- On-line data-entry and data upload
- Validation checks and remarks handling
- Calculations
- Data and metadata import, export and exchange
- Events subscription and notification
- User environment for metadata prototyping and data production
- Reporting and publishing
- Search of information
- Inquiry and download of metadata and data
- End-to-end monitoring of business processes

It is worth noting that the list is open ended and may easily grow in connection with the needs of the processes to be automated. For example, at the moment, development of new services is under way to support business processes related to handling of monetary and financial institutions register.
It is also intended that INFOSTAT should support the processing of unstructured data (such as textual and reference data) and some emerging trends (e.g. collaborative work, interactive graphics).

2.2.2. Flexibility

As mentioned above, the ability to meet new users' requirements in a very short time is crucial. The same approach mentioned above for comprehensiveness allows INFOSTAT to provide flexibility too and to ensure that the cost of changes is as low as possible.

INFOSTAT actively uses metadata which describe most of the issues concerning data processing (e.g. data structures, transformations, presentation). Consequently, the majority of developments in the statistical processing can be carried out promptly by metadata administration, avoiding software maintenance. The changes which can be supported in such a way include, for example, the collection of a new survey, the production of new sets of statistics, the delivery of a new publication, etc. The conceptual level of the definitions and the advanced user interfaces make metadata administration easy and effective, so that administrative users can accomplish it directly, without intervention of technical staff.

Changes in users' requirements which imply the delivery of new features can also be effectively supported and the Service Oriented Architecture plays a key role in providing this kind of flexibility. In fact, a modular architecture based on loosely coupled components makes it possible to manage changes concerning specific functionalities, avoiding large-scale impact. In addition, service reusability can support new requirements exploiting the existing services, thus dramatically reducing development time and costs.

An environment with a large number of different business processes that must be frequently changed can gain substantial benefits from a "Business Process Management" (BPM) tool. Such a tool allows INFOSTAT to define workflows by using a declarative approach and to properly orchestrate services invocation. Changes in user needs that impact only on the workflow can be accomplished without software intervention. In addition, the visual representation of workflows enables developers as well as users to easily inspect the service integration solution and be confident that it reflects the underlying business process.

2.2.3. Independence and cooperation between users

The ability to customize the behaviour of the information system according to the organizational structure and the responsibilities is essential.
For example, specific users (either a person, or a group, or an organization) may require a full independence from all other users. In such a case, they should be able to define their own data and metadata and to carry on their own operational activities without interfere with any other users.
In others cases, users require to act in a coordinated way so that concepts are defined once and are shared by the whole group, enabling data integration. Concepts sharing does not prevent to handle operational activities using independent processes. For example, the balance-of-payment and the MFIs balance sheets, which share a common set of concepts, may be produced using parallel processes handled by different administrators.

INFOSTAT is designed to support both the scenarios described above as well as any other combinations of them. In detail, it can provide various isolated environments (so called "statistical communities") each of which hosts data, metadata and processes belonging to independent users' groups (for example, statistics concerning Central Bank and Financial Intelligence Unit can belong to different statistical communities). Within each statistical community, a wide range of cooperation schemas can be carried on. They can range from loosely dependent activities of individuals to strongly coordinated work of specific users' groups.

A specific section of Matrix represents the configuration of responsibilities on data, metadata and processes, in term of ownership, administration and use rights.[13]

Another characteristic of the statistical activity is coexistence and interaction between "systematic" tasks, which are typical of an organized work, and "non-systematic" tasks, usually freely performed by single users, analysts, researchers.
Systematic activities are performed on the basis of well-known, pre-defined workflows which allow a high level of automation, while free tasks depend on a large number of factors that cannot be defined in advance, so they must be carried out under the user's control. A free task, however, may eventually become procedural. As a matter of fact, some activities are undertaken as free and their workflow can be consolidated and automated only after a trial phase.
INFOSTAT provides a virtual personal environment where users can perform "free" data production activities using services as they desire. This environment also allows defining metadata and testing their functioning within a pre-defined workflow in a sort of metadata prototyping activity. For such purposes, an advanced interface allows users to search and access all available information as well as the related documentation, and suitable services allow transferring definitions and data from the "official" environment to the personal one and vice versa.

In the end, it is worth noting that cooperation among users is improved by services such as events subscription, events detection and events notification: for example a user can ask to be automatically

[13] For the modeling of "roles" and "competencies" and for the "administration model" see [Del Vecchio, 2002] and [Del Vecchio, Di Giovanni, Pambianco, 2007].

informed (e.g. via e-mail or instant message) of the change to certain data or metadata (for instance when new data are available or when a code-list is updated).

2.2.4. Openness to the statistical world

A statistical information system should be open to the external world in many ways. Interactions may regard many actors (users, tools for statisticians, external information systems) and involve various degrees of complexity. In this context, information sharing should be promoted within and outside an organization.

The basic interaction toward users is provided by an advanced user interface able to search and access all available information and the related documentation.

More complex scenarios involve import and export data from/to the most common statistical packages (SAS, STATA, MathLab, Excel, etc.). These scenarios can take great advantage of the SOA as well as of compliance with the most common international IT standards (e.g. Web Services, WS-I). These elements make it possible to expose services which can be accessed from any system at any location, thus facilitating interoperability with various counterparties. Dedicated INFOSTAT services provide an "application to application" interface: invoking them, for instance, statisticians can access statistical data from a statistical package, produce new statistics using such a package and make them available for other users.

Efficient data sharing and exchange with external agencies require alignment with internationally agreed guidelines. Accordingly, INFOSTAT is designed to support the most common standards concerning statistics (i.e. SDMX and XBRL).

In detail, interoperability among Matrix and SDMX allows to extend Matrix registry services in order to inquiry SDMX registries. In such a way, INFOSTAT can deliver to its users a unitary catalogue of information which include both internal warehouse and external SDMX sources.

Furthermore, in order to achieve interoperability and service sharing within the European System of Central Banks, the INFOSTAT architecture is aligned with the ESCB IT Reference Architecture.

2.2.5. Usability

Data analysis can be dramatically improved by means of advanced and dynamic interfaces. They should allow to search information stored within and outside the organization as well as to show data using different visual representations (pivot tables, graphs, territory maps etc.).

INFOSTAT provides various tools for searching data (e.g. browsing through hierarchical contents' catalogue, full-text search). Nevertheless, search effectiveness could be greatly enhanced exploiting academic researches as well as emerging technologies.
Search results could e.g. be improved by the adoption of a "semantic search". Such technique could take advantage of ontologies in order to represent the knowledge embedded into the rich collection of available metadata.
Furthermore, the development of a network of SDMX registries could allow to better search external contents located at various organizations.
In the end, other enhancements could be achieved by sending search requests to several different search engines and aggregating the results into a single list

INFOSTAT adopts a Rich Internet Application (RIA) paradigm to build highly usable and dynamic users' interfaces. This paradigm enhances the user-friendliness and dynamism of the interfaces by combining the richness of desktop applications with the ubiquity of Web applications. RIA interfaces can be faster, more enjoyable and much more usable than traditional HTML pages.

2.2.6. Software independence of underlying resources

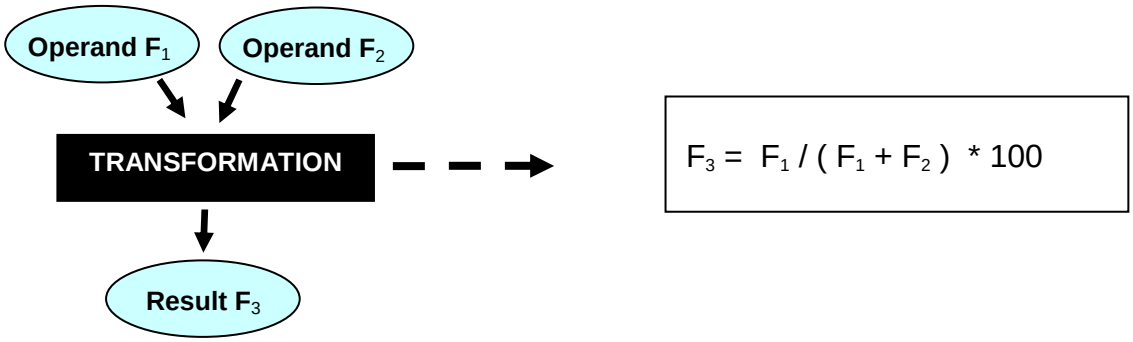
A software system should be, to some extent, independent of the underlying IT technologies and resources so that they can evolve smoothly to support new requirements (e.g. more users or processes) or exploit new opportunities (e.g. virtualization, cloud computing).

INFOSTAT is designed to be portable with respect to various technological environments. For example it can be installed on different hardware and can operate with several DBMS. At present, it runs on industry standard servers equipped with Linux operating systems. In order to enhance portability and interoperability, application software is mainly written in Java. The joint adoption of a modular architecture and widely used technologies (Linux, Java, Web Services) permits full exploitation of the opportunities coming from market tools as well as open source software. In detail, a wide set of INFOSTAT functionalities are developed by integrating custom software with open source software (e.g. ANTLR, System R, Jasper Report).

3. Representation of data transformations

The representation of the rules to obtain new data from the existing ones is an important feature of the Matrix model. The Matrix object committed to this end is the "Transformation". [14] It represents the algorithm by which a specific result can be derived from one or more operands. Results and operands are "Matrix cubes" so that they can represent the various kinds of data involved in statistical processing (multidimensional arrays, time series, register data). The algorithm is a sequence of calculation steps described in a formal syntax (expression).
The representation of Transformations has an "external" and an "internal" view. The former considers the Transformation as a black box, representing only the oriented links between the input and the output data. The latter sees the Transformation as a white box, specifying the algorithm that transforms the input into the output data, by means of a formal expression[15]. In the following example the statistical function F_3 is calculated as the percentage of F_1 on the sum of F_1 and F_2 using the operators +, *, /.

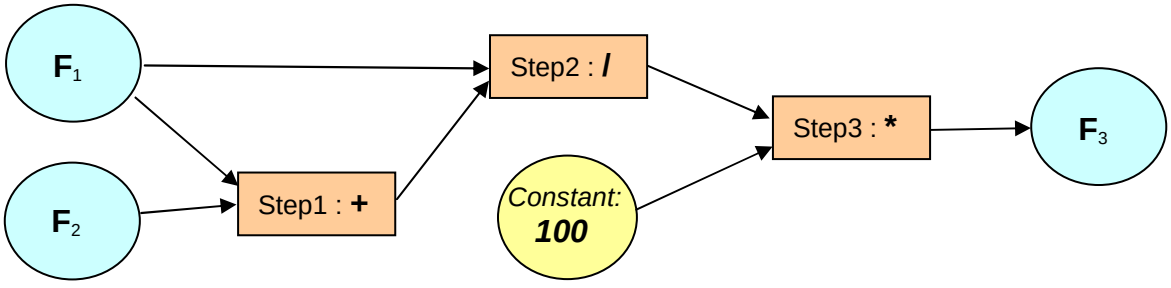
Figure 1 - The Transformation definition



Transformation: External view *Transformation: Internal view*

When the Transformation is executed, the software, using the Internal view, applies the operators to the operands in the sequence specified by the expression:

Figure 2 - The Transformation execution



Because the calculated data may be, in turn, the input of other Transformations, the external views form a "directed acyclic graph" tracing the calculus sequences of the processing system. As usual, this graph is available to users and the software, thus fully documenting the calculation system. In executing Transformations, it allows the software to determine the correct calculus sequence. Navigating the graph both in the "direct" and in the "reverse" directions it is possible to determine which outputs contribute to a given input and, vice versa, which inputs derive from a given output,

[14] See [Del Vecchio, Di Giovanni, Pambianco, 2007]

[15] The metamodel of Transformation's expressions is fully compliant with the "expression metamodel" of the CWM foundation packages (see [Object Management Group, 2000]) and in addition it is historical.

allowing impact analysis and both "input driven" and "output driven"[16] approaches for data production. For example it is easy to determine which calculated data should be refreshed if some collected data change.

Transformations also take into account the availability of the operands to determine the calculability of the result and to document the latter's availability, once calculated.

The calculus specifications may be completed with possible external conditions that can drive the execution of a given calculus, business rules to establish "what" has to be processed and "when", data to be processed together and coherently, and so on.

The Matrix model adopts a specific language called EXL to define expressions. EXL syntax is formally defined using the Backus-Naur Form notation. It includes a large set of operators and the rules for combining operators in expressions.
EXL has been designed recently, within the programme for building INFOSTAT. It takes into account several international initiatives[17] complementing the Bank of Italy's specific experiences and requirements.

The appearance of EXL expressions is very similar to the formulas of a spreadsheet, so it is quite intuitive to define them. As noted earlier, the overall structure of the Transformation expressions is also analogous to the spreadsheet structure (directed acyclic graph). However, while in a spreadsheet the operands and the results are single values, for the EXL they are whole statistical data sets (arrays, time series, multidimensional quantitative and qualitative data).

Furthermore, the language complies with all the characteristics and features of the data representation model. For instance, the language is "historicity sensitive", i.e. operators are aware of the historicity of concepts[18] and of their relationships[19] and work accordingly, making proper use of the representation of time changes (and so on) within calculations that involve data at different reference times (e.g., to manage cases of mergers, acquisitions, spin-offs, or changes in the composition of aggregates, hierarchical lists of codes, and so on, as mentioned earlier). The language also takes into consideration the other Matrix meta-information (like the definition of synonym codes, and so on) to be applied within calculation expressions.

The Matrix EXL is also designed to determine automatically not only the "extension" of the result (that is, the set of data occurrences) but also its "intension" (that is its definition according to the Matrix model).

Like the general approach of the Matrix model, the vocabulary of the EXL is at a conceptual level, suitable for use by subject matter experts and independent of any specific IT implementation. Therefore the EXL supports a very effective role distinction between Business and IT people within an organization: EXL is a tool for the information workers and fosters user autonomy, which is a critical success factor in meeting "time to market" and "cost" challenges[20].

The EXL is conceived as an extensible language, in order to support the great variety and variability of business requirements. The language includes the operators involved in the most common statistical processing; they provide both basic operations (for example +, -, *, /) and very complex calculations (for example, Herfindahl index, seasonally adjusted series).
In addition, language includes the operators required to validate and process register data (e.g. referential integrity constraints, strings manipulation).

---

[16] In an output driven process the objective is to identify which inputs are needed to calculate a given output; vice versa in an input driven process the objective is to identify which outputs can be calculated from given inputs.

[17] For example MathMl - www.w3.org/Math/; XBRL Formula - www.**xbrl**.org/; Microsoft MDX - http://technet.microsoft.com/it-it/library/ms145506.aspx.

[18] In the Matrix model terminology, concepts are all the components that form the data: variables, sets of values and values themselves.

[19] Examples of relationships between concepts are the hierarchical links inside hierarchical lists of codes.

[20] However, nothing prevents the EXL from being used by IT specialists, if desired.

EXL allows also to define templates for recurring formulas thus improving the language's simplicity and user-friendliness (templates are ordinary EXL expressions whose operands are placeholders instead of specific cubes).
In the end, new operators can be added according to processing needs.[21]

The expression language can be employed in various statistical activities (e.g. data collection, register handling, derived data production) and can be also used to trace the graph of operations across them. The calculation of derived data plays a crucial role also in data validation and in fact the very first use of the expression language was to make consistency checks.

The data validation phase is obviously crucial to the improvement of data quality. The first kind of validation is the 'structural validation' of the observations, to verify whether the data extensions conform with the definitions. This kind of validation is made possible by the existence of a dictionary (registry) where data structures are defined and does not require calculations. Structural validation, however, does not give a great deal of information about the plausibility and the quality of the data. The "spatial" and "temporal" coherence of the data are evaluated by means of consistency checks, based on EXL expressions. Simple examples are the discovery of unusual observations with respect to the likely behaviour (e.g. abnormal trends across time or of outliers within a population), the comparison between data having coherence constraints (e.g. "assets" and "liabilities" of the balance sheets)[22] and the detection of data which do not meet referential integrity constraints (e.g. wrong register data).

Consistency checks are treated as normal calculations: they are defined by means of Transformations using the EXL. The same conceptual language is therefore used to define all types of calculation, independent of the purpose. Consistency checks are thus defined and performed following the same criteria described earlier and with the same advantages. Inter alia, the graph of the consistency checks is part of the overall graph of the Transformations of the statistical information system. And consistency checks may also be applied in other phases of statistical activity (for example, the coherence of derived data can be checked against expectations).

The results of the consistency checks are "cubes" containing information about the data quality of other "cubes". Quality cubes can be used within EXL expressions as any other cube, in addition, they can be used to drive the actions to be performed on the data they refers to (e.g. discarding or accepting them, correcting or estimating them, making them available to some kind of users or not).

4. Calculation services

One of the most important activities performed by statisticians is the production of derived data by processing existing data.

INFOSTAT includes specific services to support both the definition of the algorithms based on EXL (i.e. metadata definition service) and their evaluation (i.e. calculation engine).

An appropriate user interface helps define EXL algorithms. It assists users in discovering the data they need (operands), in composing formally corrected expressions and in defining the characteristics of the results to be produced (e.g. storage properties). Algorithm validation is a complex process: each expression must be validated against the EXL syntax[23], then the coherence among data structures of the operands and the results must be checked, finally the full set of the expressions must be analyzed in order to avoid possible cyclical references. Once defined and checked, the expressions are stored as any other metadata (using the Matrix registry) so that they can be actively exploited by the INFOSTAT software components.

Calculation engine supports a large set of operators belonging to several categories. Using these operators, users can define a wide range of algorithms.

[21] The new operators must obviously comply with the Matrix and EXL general principles, in order not to make the system inconsistent.
[22] For example, data redundancies may be intentionally used for data quality controls.
[23] The open source ANTLR has been adopted as a base for developing the EXL parser component.

EXL operators require software components that implement their semantics. Because these components can be very complex (e.g. seasonal adjustments) but must also have a high degree of correctness and efficiency, most EXL operators can be fruitfully implemented by exploiting the calculation capabilities delivered by statistical packages. To this end, INFOSTAT translates the EXL expressions into the languages of the package to be exploited. In such a way INFOSTAT can combine the calculation capabilities of market tools, custom and open source software.
Currently, INFOSTAT embeds two different packages: an open source package (i.e. System R), which is used to evaluate most of the expressions, and a custom component specialized to perform multidimensional aggregations involving large amounts of data.

The main benefit of this modular architecture is flexibility. It is easy to implement new EXL operators (e.g. estimates) by exploiting the capabilities of the currently embedded packages or by integrating new packages (e.g. SAS). In addition, the implementation of existing operators may be changed (for instance to enhance performances) without impacting the EXL or the expressions already defined by users.
The architecture of the calculation engine makes it possible to reduce the complexity while improving the quality (e.g. correctness, efficiency) of the overall system.

The same approach is followed for the operators performing data access. They allow to retrieve and store several kinds of data using a user oriented syntax independent of technical details concerning the storage tools.
As described earlier, in performing these operations INFOSTAT retrieves by the “metadata registry” the information that describes where data reside and how to access them and translates the EXL data access operator according to the proper storage tool and location. This way EXL algorithms can use data extracted from the enterprise warehouse as well as from other systems. External systems can include RDBMS supporting operating procedures, SDMX-compliant systems, OLAP tools, etc. Data extraction must be done dynamically in order to avoid unnecessary duplication. INFOSTAT is equipped with several extractors which can handle data which differs in several respects: logical format (Matrix, SDMX, XBRL), data source (DMBS, file system, web services), etc.. INFOSTAT modular architecture allows the rapid addition of new extractors to access other systems.

These features make it possible to define EXL expressions which operate on data defined according to other information models (e.g. SDMX or XBRL) and located everywhere in the web or in an internal network. For example, a user can define an EXL expression which combines data extracted from an SDMX compliant source with data coming from Matrix compliant sources. In such a case, the calculation engine automatically retrieves SDMX data by using SDMX query services. The user which writes the expression does not need to deal with the implementation details concerning the access[24].

It is worth noting that calculations are characterized by “granularity”: although the algorithm is completely defined by means of its EXL expression, the execution may be granular because of input availability or the decision of the administrator. For instance, the results of a periodic survey are available only for the dates for which the survey has been executed and only for the interviewees that have actually responded.
INFOSTAT, according to the Matrix model, represents data availability as ordinary “cubes“ (status cubes) both for collected and calculated data, so the suitable and desired granularity can be chosen. The calculation system takes into account the availability of the input data “grains” to determine the calculability of the result “grains” and updates the “status cube” relevant to the result, once the calculation is done, to document the availability of the calculated grains. Like other cubes, status cubes are available to users and can be used for further processing, also as EXL operands in conjunction with other cubes.
Computational efficiency is one of the most critical requirement for the calculation component. However, the large amount of data to be treated as well as the complexity of algorithms to be performed, make it extremely challenging to achieve high performances. For these reasons, the engine is designed to perform several operations in parallel: it decomposes each expression in smaller chunks which can be performed independently and assigns their evaluation to different executors deployed on various servers.

[24] This way also the ECB statistical data warehouse will be accessed.

In the end, setting up a new algorithm often requires an iterative approach in which the expressions are gradually refined until the results meet the users' needs. INFOSTAT includes a specific service to support this activity (algorithm prototyping). Users can export a specific vintage of an algorithm in a dedicated environment[25] where the algorithm can be modified and then evaluated using the calculation service. Calculation can be performed using input data taken from the warehouse or provided by the user itself. At the end of the prototyping activity, the algorithm can be consolidated into the registry.

5. Practical cases of calculations and consistency checks

This section describes two scenarios taken from the Bank of Italy's experience in which the calculations and consistency checks play an important role.

5.3. A survey of small financial intermediaries

Recently, INFOSTAT was used for collecting and processing financial regulatory data submitted by approximately 1.200 small financial institutions. Financial data are submitted using a web interface and are validated by means of EXL algorithms immediately upon receipt. In case of validation failure, an e-mail notification is sent back to the submitter.
In this real example, two different survey micro-data (designated **3A_6413200** and **3A_6413400**) have the same meaning (loans) but different classification variables except for the date, so their total amount by date is expected to be the same. The check consists in calculating the discrepancy (as an indicator of lack of coherence) between the total amount by date of the two original data.

The steps to calculate such an indicator are the following (the EXL instructions are also shown):

- calculation of two aggregated data expected to be equal, starting from the two collected data:
  Step1a = get([3A_6413200],keep(DATE,AMOUNT),sum(AMOUNT))
  Step1b = get([3A_6413400],keep(DATE,AMOUNT),sum(AMOUNT))
  The results of these steps are the two total amounts by date.
- comparison of the two total amounts by date and storage of the difference/discrepancy in a new data (designated 3A_6432600):
  3A_6432600 = check(Step1a – Step1b)
  The result of this step is the difference by date[26], which is stored.

Check results, together with the collected data, are thus available for further processing to end users, to inquiry tools, to other statistical packages.
Besides this simple case, the migration of about 10,000 business validation rules, also very complex, from the older systems to INFOSTAT is now under way.

5.4. Production of the information for the ECB concerning the balance sheet of the monetary and financial institutions sector

The Bank of Italy produces statistical data for the ECB, according to the regulation 25/2009/EC concerning the balance sheet of Monetary and Financial Institutions (MFIs). Some of the time series to be produced concern information on the securities managed by the MFIs. The process involves data collected from MFIs as well as other information sources, such as the securities register.

This case is currently handled by the old processing system and will be migrated to INFOSTAT in the next future. The new solution consists in a unique graph of Transformations able to calculate the final result from the input data. Besides mathematical calculations on quantitative data, this same graph includes operations on qualitative information, such as those drawn from the securities register. The key point is that the security register too is represented in the information system as a "cube" and can be used as operand of Transformations.

The main steps of the process are described below.

[25] That is the virtual personal environment mentioned earlier.
[26] The "check" operator actually returns also other information besides the difference.

1. The Bank of Italy collects the required statistical information on securities issued and held by MFIs on a security-by-security basis.
2. The data are checked and certified at the desired quality level by means of structural and consistency checks, as in the previous use case.
3. Next, the data are integrated with other information relevant to the collected security codes taken from the securities register. In detail, using a proper EXL operator it is possible to enrich the data with the desired dimensions.
4. Then the data are aggregated along some dimensions (for example currency, sector of economic activity, and country). The EXL aggregation operator follows the hierarchical code lists (also called Classifications) defined by the user in the Matrix registry. It also takes care of the time validity of both the codes and their hierarchical links.
5. Finally, estimates of missing observations and other derived data are produced, always by means of EXL expressions.

This example gives a glance at the system's ability of supporting a fully automated process involving highly heterogeneous data.

6. References

1. Del Vecchio V.: *Statistical Data and Concepts Representation*, Banca D'Italia, September 1997, http://www.bancaditalia.it/statistiche/quadro_norma_metodo/modell_SIS/StatisticalDataAndConceptsRepresentation.pdf
2. Del Vecchio V.: *The Banca d'Italia's Active Statistical Meta-Information System*, European Commission Information Society Technologies Programme, Proceedings of 1st MetaNet Conference, 2-4 April 2001, http://www.epros.ed.ac.uk/metanet/deliverables/D3/D3Proceedings.zip http://www.bancaditalia.it/statistiche/quadro_norma_metodo/modell_SIS/The_BI_Active.pdf
3. Del Vecchio V., *Modelling Levels in the Statistical Information System of the Bank of Italy*, Banca D'Italia, Oct. 2002, http://www.bancaditalia.it/statistiche/quadro_norma_metodo/modell_SIS/ModellingLevels.pdf
4. Del Vecchio V., Froeschl K.A, Grossmann W.: *The Concept of Statistical Metadata*, European Commission Information Society Technologies Programme, MetaNet Project Deliverables, February 2003, http://www.epros.ed.ac.uk/metanet/deliverables/D5/IST-1999-29093_D5.doc
5. Del Vecchio V., *The Banca d' Italia experience: hierarchical modeling statistical information systems*, European Commission Information Society Technologies Programme, Proceedings of the final MetaNet Conference, 7th-9th May 2003, http://www.epros.ed.ac.uk/metanet/deliverables/D8/D8_FinalConference_Proceedings_Final.pdf
6. Del Vecchio V., Di Giovanni F., Pambianco S.: *The Matrix Model – Unified model for statistical data representation and processing*, Banca d'Italia, May 2007, http://www.bancaditalia.it/statistiche/quadro_norma_metodo/modell_SIS/matrixmod.pdf
7. Object Management Group (OMG): *Common Warehouse Metamodel (CWM) Specification*, February 2000
8. Romanelli M.: *Extension of BdI Statistical Information System to manage XML-based data formats,* VI CEBS XBRL Workshop, 4th-5th October 2006, http://www.eurofiling.info/6th_workshop/VI_CEBS_Workshop_4-10-06.pdf
9. Romanelli M.: *Matrix schemas for COREP and FINREP taxonomies,* Committee of European Banking Supervisors (CEBS), 22th November 2007, http://www.eurofiling.info/finrepTaxonomy/taxonomy/descriptions/readme_matrix_schema.pdf
10. Statistical Data and Metadata Exchange Initiative: *SDMX Information Model: UML Conceptual Design (version 2.0),* november 2005, http://www.sdmx.org/docs/2_0/SDMX_2_0%20SECTION_02_InformationModel.pdf
11. Sundgren B.: *Statistical Metainformation and Metainformation Systems*, Statistic Sweden R&D Report, november 1991.