Disseminating data from the Business Registers: an Integrated Web System

Massimiliano Amarone\textsuperscript{1}, Stefania Bergamasco\textsuperscript{1}, Monica Consalvi\textsuperscript{1}, Stefano De Francisci\textsuperscript{1}, Leonardo Tininini\textsuperscript{2}

\textsuperscript{1}Istat, e-mail: \{amarone, bergamas, moconsal, defranci\}@istat.it
\textsuperscript{2}CNR, e-mail: leonardo.tininini@iasi.cnr.it

Abstract

The aim of this paper is to describe the integration of two important projects carried out in Istat in the last years: the Statistical Register of Active Enterprises (ASIA) and the Integrated Output Management System (Istar). ASIA has been developed in order to achieve the statistical integration of different administrative sources and covers the entire population of active enterprises of industry and services. Istar has been developed in order to integrate and manage the statistical data supplied and validated by the statistical production areas of Istat and produce purposeful statistical outputs for end users. The paper points out the main steps which led Istat to perform a full integrated workflow to disseminate multidimensional data on the Web by Istar starting from the ASIA legacy database.

Keywords: Integration of business registers, Dissemination of statistics, Web Warehouse

1. Introduction

During the last years, Istat has centered more and more its ICT strategies on the development of standard and reusable software, particularly in the field of data warehousing. Generalised packages to manage the processes for statistics dissemination on the Web have been developed. Such solutions are available to be used directly by the statistical production areas to deploy their specific instances of statistical data warehouses. This strategy has been preferred to a corporate monolithic environment, because of the organizational models adopted by the Institute, based on a distributed architecture. Several production Directorates operate through local subsystems that, independently from each other, cover the most part of the life cycle of statistical data, often adopting generalised software and technical solutions. For this reason, harmonizing and integrating also the dissemination processes has become mandatory, with respect to both the software adopted and the statistical outputs produced by the systems. In order to reach these goals, Istat has developed the Integrated Output Management System (Istar). This system is able to extract both elementary and aggregated data from manifold and heterogeneous sources, transform them into multidimensional format, load the data in statistical data warehouses and finally disseminate the information to many different users, by means of different types of outputs and technologies.

The basic principle of Istar consists in providing a set of standard toolkits to maintain, integrate and manage the data supplied by the statistical production areas of Istat after the validation processes. Also in the case of business registers the use of Istar for disseminating data has taken advantage of this underlying characteristic, making
available a generalised toolbox for navigating aggregated data on the Web starting from the Statistical Register of Active Enterprises.

2. The Statistical Register of Active Enterprises (ASIA)

In ten years, from 1998 to 2008, Istat has completely reformed the production of statistics on enterprises of industry and services. In 1993, the European Commission required the member States to realise business registers based on administrative data, to be used for the yearly production of harmonised official statistics on the whole population of non-agricultural enterprises. The evolution process led from the traditional enumeration of economic activities to an integrated system of statistical production, and can be defined as a continuous Census since it provides every year statistical information on the territorial distribution of economic activities and the employment, so far available through the Census every decade. In 1996 Istat started the project of the Italian Business Register (BR), named Statistical Register of Active Enterprises and the acronym ASIA was adopted (in Italian, Archivio Statistico delle Imprese Attive).

To explore the technical feasibility to set up the statistical business register, even to comply with the requirements of the EU (Council Regulation (EEC) N. 2186/93 of 22 July 1993 on Community coordination in drawing up business registers for statistical purposes), the main steps were:

a. Definition of a metadata framework;

b. Study of the main available administrative archives (definition of units and characters, classifications used, coverage, maintenance and updating procedures);

c. Development of a “metadata translator”, i.e. a set of rules to convert the administrative data into statistical information, by identifying the statistically-relevant units among the legally-relevant ones;

d. Set up of a robust methodology to estimate/validate the characters of the identified statistical units.

Specific statistical methodologies have been developed to update ASIA. The main problem to solve in producing statistical information from administrative sources is to establish correspondences between the administrative rules and laws that define a legal picture of the observed universe and the concepts defining a statistical picture of the same universe.

ASIA has been developed through the statistical integration of different administrative sources, covering the entire population of active enterprises of industry and services, other minor archives available (covering particular sectors), and structural business statistics currently produced by Istat. The updating procedure, with reference to a generic year \( t \), consists of three macro-phases. With reference to the year \( t \), the set-up process starts in the last quarter of the year \( t + 1 \), when the yearly data supplies from the main sources are available. After a process of normalization and standardization, which converts the administrative units and variables in statistical ones, the data are integrated. The output is a set of statistical units, which is the ASIA release for the reference year \( t \). The main structural and identification variables for each integrated unit are then estimated. The attribution of economic activity sector, legal form and some identifying characters is done only for units presenting disagreements between different sources. For units that do not show changes in the input sources referring to the year \( t \), the characters are inherited from the \( t - 1 \) release. Besides, the activity status and all variables measuring the employment are estimated for all the units. This procedure leads to define a set of enterprises active in the year \( t \) together with their
characters. Any information obtained will be subject to a process of quality control, whose final round is the updating of the ASIA Information System, a relational database that contains historical information and changes regarding each statistical unit over the years since 1996 until today.

The new European Regulation on Business Registers (EC N.177/2008 of the European Parliament and of the Council of 20 February 2008 establishing a common framework for business registers for statistical purposes and repealing EEC N.2186/93) requires, for the multi-location enterprises, the registration and update of the information on all their local units, i.e. the realization, within the BR, of an additional information level, able to produce – just like the Census – territorial data. Therefore, in 2004 Istat started to implement a statistical register of enterprises’ local units (ASIA-UL).

Since 2006, ASIA-UL provides every year (with a two-year delay) the information on local units, that before was available only from the Census, that was normally taken every ten years. Just like the BR of enterprises, ASIA-UL is the result of the integration of administrative and statistical sources, in part already used for the construction of the BR. But the administrative archives available in Italy don’t provide reliable and complete information on local units, especially with regards to the territorial distribution of employees. Therefore, to fill this gap has been necessary to organise a new direct survey on local units of multi-location enterprises of large dimensions (IULGI) in order to verify on the field the state of activity and the other characteristics of the local units. The yearly survey on the local units of large enterprises, together with the administrative sources, allows the construction and the update of the local units register.

The BR is considered the reference universe and the official source for statistical information on the structure and the demography of the business population. Istat produces since 1999 a set of business demography indicators (birth, death and survival rates) by sector of activity and by region. This production became compulsory for all EU countries only in 2008, due to a modification of the SBS Regulation (Regulation (EC No295/2008 of 11 March 2008). The methodology used to identify the “real” enterprises’ births and deaths among all the administrative events recorded by the BR has been developed by a Working Group in which Istat played a prominent role.

Other four central registers, not all completely developed at the moment, are included in the ASIA system: ASIA Groups, ASIA-Institutions, ASIA-Agriculture and the Farm Register. The first one records information on financial links between legal units and enterprise groups. The first issue of ASIA-Groups refers to 2002. ASIA-Institutions covers the Public Administration and private non-profit organizations. At the moment, Istat completed only the part referring to the public administrations included in the sector S-13 of the SNA 1995 classification. ASIA-Agriculture includes all commercial enterprises carrying out their main activity in agriculture, fishing and forestry; its first release is foreseen by the end of 2008. The Farm register is currently under development, and its first release will be tested through the Agricultural Census of 2010.

The BR is more and more used as a dissemination tool, i.e. a source for statistical analysis of the economic system, especially with regards to the territorial dimensions of the economic dynamics. This process started in 2004 and it has important consequences, regarding in particular the production of time series from the BR, since the BR is a “live entity”, that is continuously updated.
Obviously, a direct access to the BR for information extraction is restricted to specific categories of users, due to privacy and sensitive information disclosure issues. Furthermore, a meaningful extraction and aggregation of the data in the BR requires an expertise, which is generally not owned by a conventional non-expert user. As a consequence, a specifically designed information system was required, enabling the user to freely browse the data of the ASIA system, although requiring a minimum amount of technical expertise and available by means of a conventional Web browser. The effectiveness of the Web front-end to browse the available data was obviously an important factor in the choice of the system to be adopted. However, due to the large amount of data to be disseminated, as well as the frequency of new data releases, the system should also support very efficiently (and as far as possible automatically) the several workflow activities required to disseminate new information, starting from the data in the BR. The comparative evaluation of some available systems finally led to adopt the IstarMD toolbox.

3. The ETL process

ETL is an English acronym that refers to the process of Extraction, Transformation and Loading of data into a data warehouse or OLAP system. In order to simplify the management, the ETL process has been split into three phases. In the first phase the classification variables have been defined and set up. In the second phase the classification values have been created, extracted or modified and, finally, in the third phase the microdata on enterprises, local units and enterprise groups have been extracted, processed and loaded. One of the main difficulties encountered was finding the required information from the source system (transactional database). These problems were due to the fairly high complexity of metadata (the structures that contain information regarding the microdata).

For uniformity reasons the entire ETL process has been implemented in SQL and PL/SQL languages as both the transactional (OLTP) and the analysis (OLAP) systems adopt Oracle as RDBMS. In particular, the data extraction process has been simplified by using virtual tables (SQL views), which embed PL/SQL functions to calculate totals and averages, whenever necessary. Furthermore, special grouping instructions (Case When) have been used to classify the data according to the defined variable modalities.

The loading of the classification variables and data sources (microdata) into the analysis system's primary data warehouse (PDW) have been implemented by PL/SQL procedures, collected within dedicated containers (packages). Classification variables have been loaded from specifically created text files, while microdata loading involved several steps: (i) a logical link (dblink) has been created between the two (OLTP and OLAP) systems; (ii) the PDW table has been created; and finally (iii) the microdata loading has been carried out by using a buffered selection (1000 lines at a time) from the views defined in the previous phase. The loading process has also included operations of data harmonization.

The PDW microdata tables include the key attributes used to define the mappings between the two systems, as well as the numeric attributes (measures) used by the aggregation functions, and the classification attributes (dimensions). The key attributes are fundamental to propagate any update from the source data to the PDW. The dimensions were defined so as to allow for the insertion of new attributes, without having to recreate the table and repeat the above mentioned steps. About 5 millions of records are added each year to the data on enterprises and local units. As a
consequence, the PDW microdata tables have been partitioned by year, to improve the performances of the data aggregation process.

4. IstarMD

IstarMD is a collection of tools specifically designed to support the statisticians in the several phases required to disseminate statistical aggregate data on the Web starting from a collection of validated data. Two of the main components of the IstarMD toolbox are WebMD (the component for multidimensional “data-warehouse-like” navigation and dissemination on the Web) and FoxtrotMD (the “administration” component for metadata management and aggregate data computation). In order to better illustrate the main features of IstarMD components, it is important to preliminarily outline some (common and distinguishing) characteristics of Statistical Dissemination Systems and Data Warehouses.

The strict correspondence between statistical dissemination systems (SDSs, sometimes called also statistical databases), and data warehouses (DWHs), also known as On-Line Analytical Processing (OLAP) systems, was pointed out a few years ago by Shoshani (1997). Consequently, as DWHs have well-established methodologies and techniques, as well as powerful and user-friendly tools supporting the design, storage and multidimensional navigation of data, one may think to straightforwardly extend their use to the interactive dissemination of statistical data. However, despite the evident similarities, SDSs have several peculiarities that require conventional DWH techniques to be extended with more specific models and structures (Sindoni and Tininni, 2008). These are mainly related to sample surveys, issues of privacy disclosure (Malvestuto and Moscarini, 2003), microdata unavailability, filter questions and heterogeneous classification hierarchies (Lehner, 1998).

The differences between multidimensional navigation in a conventional DWH and an SDS are depicted in Figure 1, where the dimension levels are represented with an increasing level of detail on the dimension axes (e.g., if D2 is an area dimension, D2,1, D2,2 and D2,3 may correspond to the national, regional and municipality level) and the grey areas represent the dimension level combinations which can be accessed by users.

![Figure 1. Accessible dimension combinations in (a) a conventional data warehouse and (b) in a Statistical Dissemination System](image)

In conventional DWHs (a) the user is free to drill-down and roll-up along any dimensional hierarchy of the data cube (Gray et al, 1996), independently of the detail
level of the other dimensions. In contrast, drill-down on a dimension in an SDS (b) can only be performed starting from certain combinations of the other dimensions and conversely, rolling-up on a dimension increases the number of possible explorations (drill-down) on other dimensions. This has obvious severe consequences on the conventional multidimensional navigation paradigm and in the following we analyse how the related issues have been dealt with in the IstarMD Web dissemination system, achieving a good trade-off between the characteristic freedom and flexibility of DWH multidimensional navigation and the constraints arising in the statistical dissemination context.

WebMD is the IstarMD component for multidimensional navigation and dissemination on the Web. WebMD originates from the DaWinciMD dissemination system (Sindoni and Tininini, 2006), initially developed to disseminate aggregate data from the 2001 Italian Population and Housing Census¹ and more recently used to disseminate data from:

- the graduate education and employment Italian survey²;
- different surveys for setting up a system about “The framework for integrated territorial policies”³;
- the household budget survey of the Bosnia and Herzegovina Agency of Statistics⁴.

Three further issues are currently under development to disseminate the data about foreigners and immigrants, labour market and industry. The system is based on the definition of the maximum detail dimensional combinations of each global cube, basically corresponding to all permitted subcubes (e.g., the gray zones in Figure 1). The user can start by displaying the data corresponding to a certain combination of measure (object) and dimension levels (classifications) and then navigate to other subcubes through roll-up and drill-down, without ever violating the dissemination constraints or returning to the data cube selection page. It is the system itself that proposes, on each visualisation page, all and only the dimension levels compatible with the measure and dimension levels already selected, thereby always leading the user to a permitted dimensional combination.

Figure 2. The table visualisation page of WebMD

---

¹ http://dawinci.istat.it/MD
² http://dip.istat.it; http://lau.istat.it
³ http://incipit.istat.it/
⁴ http://hbsdw.istat.it/dawincibosnia
Figure 2 shows the table visualisation page of WebMD with its two main sections: the control panel in the upper part of the page that contains the access mechanisms to all navigation functions; and the statistical data visualisation section in the lower part that contains the table with statistical data, or one of the pages that compose the table if the number of classifications is too large to be displayed in a single page. Preliminary cube selection is based on the interdependent selection of the object and classifications of interest. Figure 3 shows the table selection page of WebMD enabling the user to express the required table by selecting (without a predefined order and possibly only in part) the object, classifications, territory and year of interest.

Figure 3. The table selection page of WebMD

The concept of object in IstarMD basically corresponds to that of measure in a conventional data warehouse, although an object may also incorporate some slicing operations on the data cube. In order to guide the user in selecting the required cube, objects are organized into hierarchies, mainly based on generalization relationships, and the user can choose “generic” objects, i.e. those located in the higher levels of the hierarchy. The system is able to combine the generic user choices and map them to the actual object-classification combinations specified by the metadata. WebMD classifications basically correspond to specific dimension levels of the data cubes, although a classification’s structure can be more complex than usual flat dimension levels.

Classifications can be shared by several cubes, enabling a user to perform classification-based navigations: for example, the user may select a combination of classifications and the system will show all available statistical aggregates (cubes) classified in that way, independently of the measure (object). As with objects, classifications are organized into hierarchies, to enable the user to express generic queries and consequently facilitate access to data.

WebMD fully supports multi-lingual dissemination of statistical data: in any phase of the table selection or visualisation process the user can indeed switch from one language to the other, always maintaining the current selection and visualisation context. The system currently supports both Oracle and MySQL Database Management Systems for back-end storage.

FoxtrotMD is IstarMD administration component, specifically designed for metadata management and aggregate data computation. By FoxtrotMD the dissemination administrator can:
• manage the **objects** of interest for the statistical tables to be disseminated, in particular their descriptions in the two languages chosen for publication, the related statistical tables (i.e. tables defined using a given object), as well as the parent relationships between them. As mentioned above, objects can indeed be arranged into a hierarchical structure based on generalization.

• manage the **classifications** of interest for the statistical tables to be disseminated, in particular their descriptions in the two languages chosen for publication, the corresponding modalities in both languages, the related statistical tables (i.e. tables defined using a given combination of classifications), as well as the parent relationships between them. Similar to objects, classifications can indeed be arranged into a hierarchical structure based on generalization and level of detail.

• manage the **statistical tables** to be disseminated, defined by the combination of an object with a certain number of classifications. Each table will have its own multi-language descriptions, object and classification components and possibly multiple **spatio-temporal instantiations**, i.e. combinations of territories and years for which data are available (and have to be disseminated). FoxtrotMD also enables the dissemination administrator to define the rules to extract and aggregate the data to be disseminated, starting from one or more tables of (validated) microdata.

• compute and store the **aggregate data** to be disseminated. By using the specified rules, the **ETL component** of FoxtrotMD can aggregate the data and store them in the aggregate data table used during statistical table visualisation by WebMD. The aggregation process is automatically performed at all levels of the territorial partitioning hierarchy specified by the administrator.

Figure 4 shows the user interface of the FoxtrotMD component for statistical table management. The system allows the user to manage the entire workflow, by driving (and partially constraining) his/her activities in a series of consecutive and interdependent steps. For example, only classifications that are not currently related to statistical tables can be modified and the data can not be modified after a statistical table has been published (unless the whole process is restarted from scratch).
The process of aggregate computation is organised in several phases aiming at verifying the compliance of microdata structure and contents to what specified in the metadata. Alerts and blocking errors can be issued during the various phases. In the former case the user can check if the warnings actually correspond to what expected and possibly enable the process prosecution. In the latter case some errors in the data or metadata prevent the system to complete the process, a correction activity is required and the process will have to be restarted from the first phase.

In more detail, the ETL component functionalities are divided into seven phases, each of which has a specific purpose described below.

**Phase 1:** In this phase the system verifies if the microdata territorial granularity (e.g. municipal, provincial, national, etc.) has been specified in the metadata. If this is the case, then the user is simply alerted of the indicated territorial granularity, otherwise a blocking error is issued and the user is prompted to insert the required metadata.

**Phase 2:** In this phase several checks on the microdata table structure are performed. In particular the system checks the presence of two columns containing the year of reference and the territorial codes, the absence of "null" values in these columns, the existence of at least one record for the year of interest, the correct correspondence between the territorial codes in the microdata table and the ones extracted from the reference territorial database.

Further checks depend on the type of each classification in the considered microdata table, in particular if it has been declared either *quantitative* or *qualitative*. Broadly speaking, if a classification is quantitative in a given microdata table, then the table will have a column for each classification modality and the single column will contain (at least partially) aggregated values. In contrast, if a classification is qualitative in a given microdata table, the corresponding column (or, in some cases, columns) will only contain modality codes, that will be used to group the records during aggregation.

Note that the same classification, e.g. 'sex', may be qualitative in a certain microdata table (and in this case there will be a single column containing either the female or male modality code) and quantitative in a different microdata table (and in this case there will be two columns containing aggregate values, corresponding for example to the number of female and male employed in an enterprise). Obviously the classification type has a relevant influence on the rules that has to be specified in the metadata to compute the aggregate values.

Other checks are related to other specific characteristics of the classification, namely if it has a hierarchical structure (e.g. single years with subtotals for ten years age groups), if the corresponding variable is multiresponse in the microdata, if the classification has one or more "of which" modalities, etc.

**Phase 3:** In this phase the data checked in the previous phase are partially reorganised and stored in auxiliary tables to speed up the following phases, especially that of aggregation.

**Phase 4:** In this phase the contents of the single columns identified in phase 2 are checked, also by exploiting the reorganised data stored in auxiliary tables during phase 3. As already mentioned above, the checks strictly depends on the type (quantitative vs. qualitative) of each classification in the specific microdata table. If the classification is qualitative the modality codes in the microdata table column(s) must correspond to those stored in the metadata repository. In particular, if a classification corresponds to a multiresponse variable in the microdata, the microdata table will have as many columns as the number of classification modalities and a specific check will be performed on each column. The values found in the microdata
columns are compared with those expected, according to what stored in both the metadata repository and the auxiliary tables, generated in the previous phase. Depending on the analysis results, a simple alert may be provided to the user or a blocking error be issued, prompting the user to fix the inconsistencies found.

**Phase 5:** In this phase the data in the microdata table are aggregated, by applying the rules specified in the metadata and exploiting the auxiliary tables generated in phase 3. The obtained data are stored in the aggregate data repository. For many reasons, some of the possible modality combinations may not have a correspondence in the computed data (e.g., there may be no individual corresponding to a certain combination of professional activity and level of education modalities, say 'lawyer' with 'secondary school'). These missing combinations will be inserted in the following phase.

**Phase 6:** In this phase the aggregate data repository is completed with values corresponding to the missing combinations determined by the aggregation process of the previous phase. This completion is required to increase the system’s performances during multidimensional navigation.

**Phase 7:** In this phase some supporting files are generated, which are mainly used to increase the system’s performances in case of massive download operations.

5. Conclusions

In this paper we have illustrated the integration of the Statistical Register of Active Enterprises (ASIA) with the Integrated Output Management System (Istar). This experience has been carried out in accordance with the strategy for an integrated approach to statistical data life cycle management and, in particular, is oriented to produce meaningful statistical outputs for end users. It is part of the Integrated Output Management System development initiative that can be considered as the core of a global integration architecture, aiming at providing a seamless cooperation among: (i) local production systems, (ii) the set of reference and documentation metadata systems, (iii) the centralized repository of validated microdata and (iv) the environments for analysing and disseminating statistical data on the Web.

References


