

An open source approach to disseminate statistical data on the Web

Un Approccio Open Source per la Diffusione di Dati Statistici su Web

Giulio Barcaroli, Stefania Bergamasco, Stefano De Francisci, Leonardo Tininini

ISTAT – via Cesare Balbo, 16 – 00184 – Roma (Italy)

Abstract: *The Italian Institute of statistics has been working for several years on the development of generalised software, i.e. reusable in its numerous statistical production lines. The recent adoption of free/open source software has further fostered the reusability strategy, by making the technological platforms more independent from the usage context. Particularly, open source and reusable solutions were chosen in the dissemination context, by developing the Integrated Output Management System, aimed to support the integration of the several transformation phases in the statistical data life cycle, up to the various output products made available to final users. In this paper the main characteristics of this system are described, along with the numerous opportunities this kind of system can offer in the context of international cooperation.*

Keywords: open source software, statistical dissemination

1. Introduction

The statistical production activities of the Italian Institute of Statistics (ISTAT) are supported by a distributed organization. Each area of statistical production operates through its own subsystems that, as often as not, cover the full statistical data life cycle, from data collection up to dissemination. For this reason, the Institute is long since involved in the development of generalized software to harmonize and integrate the production processes and the common statistical contents of its own production areas. Lately, the adoption of open source software solutions has further enhanced the actions of generalization and integration, enabling ISTAT to carry out software packages reusable in external contexts too.

Also in the phase of statistical data dissemination, ISTAT has focused its own strategies on the development of standard and reusable software. In particular, generalised packages to manage the processes for statistics Web dissemination were developed. Such solutions are available to be used directly by the statistical production areas to deploy their specific instances of statistical data warehouses. Also in this case, the gradual adoption of free/open source software has promoted a large reuse of the dissemination systems in the field of international cooperation.

This paper presents the most important results achieved in ISTAT in the development of solution based on open source software. In particular, after a brief overview on the tools developed to support the main relevant phases of the statistical survey (Section 2), the Integrated Output Management System (ISTAR) will be illustrated in detail. The general strategy of the system is presented in Section 3, while the architectural framework and a brief history of the project is illustrated in Section 4. Section 5 describes a particular

ISTAR module, namely the data warehouse module to disseminate statistical aggregate data on the Web, while Section 6 concludes the paper.

2. ISTAT open source tools for statistical cooperation

ISTAT involvement in statistical cooperation projects has very often regarded support in methodological issues and also in related software, i.e. software implementing most advanced methods and techniques for any relevant phase of a typical statistical survey.

From a very general standpoint, the issue of the use of statistical software has an impact not only on the way each phase of the statistical production process is performed, but it has to do with the very sustainability of the statistical system that is supported and fostered through technical cooperation.

Donors often dedicate limited time and resources to cooperation activities, and once methodologies are transferred and acquired, the required objectives and results are obtained by supplying beneficiary institutions with the software applications used by the relevant partner: this is often either a commercial software, whose costly licenses expire, or software developed *ad hoc* by the partner institution, whose reusability is low or even null; training is also frequently provided, rendering that specific intervention acceptable, but limited, and with no given relation with the overall IT framework of the beneficiary institution. In these institutions, the situation is worsened by the high human resources turnover, especially of those expert in software development. The rapidly changing IT environment and knowledge creates an additional pressure for these institution to optimise their approach.

As a result, the managers of a statistical institution in development have to urgently face the issue with a comprehensive strategy, as it touches aspects like scientific independence, sustainability of development processes and human and financial resource management.

ISTAT has attempted to support the institutions involved in its cooperation programmes by fostering the use of generalised tools, based on open source software wherever possible, because this particular choice overcomes the drawbacks characterising *ad hoc* solutions. In fact, generalised software can be reused by definition, with limited or no need to write new code: so, applications developed under the support of the donor can be replicated by the beneficiary institution even when the project has ended, once an adequate training has been provided.

Moreover, if a given generalised system has been developed by using open technologies, one relevant benefit is the fact that it is portable on every platform with no cost for the user. And this is a very important feature in terms of the sustainability of the cooperation projects.

In achieving that, ISTAT has been facilitated by the internal policy it had adopted since long, consisting in *internally* developing a number of important generalised systems and tools, namely for editing and imputation, for dealing with sampling issues (design and estimation) and for disseminating statistical information: as a consequence, the choice regarding underlying technologies has always been an open option.

So, when at the beginning of the new millennium, the strategy of privileging open source technologies instead of proprietary ones was adopted, it was an almost straightforward task to migrate generalised software.

In particular, we refer to the three most important set of tools used inside ISTAT, namely:

- CONCORD (*CON*trollo e *COR*rezione dei Dati) for edit and imputation of data (Riccini, 2004);
- MAUSS (*Multivariate Allocation of Units in Sampling Surveys*) and GENESEES (*GEN*eralised *S*ampling *E*stimates and *E*rrors in *S*urveys) (2005a and 2005b) respectively for sampling design and estimation;
- ISTAR for statistical data warehousing.

The first to be migrated was CONCORD, whose first version was based on SAS. The occasion was given by a cooperation project (2004 Household Budget Survey, “HBS”, in Bosnia): to let the beneficiary autonomously use CONCORD without having to pay for SAS, a different version (Linux based) was developed and, from that one, was derived the current version, that is a Java one, running on all platforms (Linux and Windows).

Immediately after, a R version of MAUSS was developed, substituting the SAS one; at the same time, a new R package, “EVER” (Zardetto, 2008) was produced, enabling users to perform the functions (calibration and sampling variance estimation) till then ensured by GENESEES, based on, and requiring, SAS. Also in this case, Bosnia was the first to adopt the new tools, for the past 2007 HBS (EVER) and for the next 2010 HBS (MAUSS-R).

First versions of ISTAR made a prevalent use of open technologies (namely Apache, Tomcat, JSP), but data management was entirely based on ORACLE system. The cooperation project for the preparation of the next Kosovo Population Census was the occasion to migrate towards an open solution, where PL/SQL procedures have been substituted by JAVA ones, and the user can choose the dbms he/she wants (for instance, MySql). Once again, Bosnia is going to be the first test-stand for the new version, as the web data warehouse containing 2007 HBS will be developed by using the new ISTAR toolkit.

So far, we have mentioned the tools ISTAT developed on its own, tools that are best suitable for cooperation projects in that they are generalised and based on open technologies. But many other software tools are available, so to cover all the phases of a statistical survey. We cite here:

- the R package “sampling” (Tillè and Matei, 2007) for optimal sampling units selection;
- for data collection: CsPro and LimeSurvey, respectively for CAPI and web surveys;
- the R packages “yaImpute” and “mice” for (single and multiple) imputation of missing values;
- R, Adamsoft, Weka, Rattle, Knime, RapidMiner for data analysis and data mining;
- ARGUS of disclosure control while disseminating data at different levels of aggregation (micro and macro).

Even if they are not all “pure” open source software tools (as their source code is not available, as in the case of CsPro), they are all free software, commonly downloadable from the Internet.

3. The Integrated Output Management System of ISTAT

The INTEGRATED OUTPUT MANAGEMENT SYSTEM (ISTAR) of ISTAT is an information system oriented towards the integration of part of the statistical data life cycle, particularly, all the steps required to produce purposeful statistical outputs for end users. In particular, ISTAR has been developed to maintain, integrate and manage the data and metadata supplied by the statistical production areas of ISTAT after the validation processes.

The track ISTAT is following to integrate its own information systems is mostly based on the exploitation of new technologies and on a new way to organize and manage the knowledge. The experiences already completed have shown, on one hand, the chance of integrating and sharing knowledge existing in differently structured information (from legacy data base or data warehouse to textual documents, volumes, etc.) and on the other hand paying more and more attention to the information needs of the users. For this reason, ISTAR is based on a complex scenario of integration which includes not only data warehouses and metadata information systems, but also descriptive and textual information and diverse models of classification of reality. It combines the approaches for browsing dimensional data, typical, for example, of statistical data warehouses on the Web, with new models for the management and representation of knowledge and the information architecture.

The technical solutions developed take advantage of the construction of specific metadata layers and their strict associations to the objects managed in the database. At the same time, the system preserves all the features of the search engine relating to optimization of search, in order to improve the performances of the scanning operations.

ISTAR is based on two general principles: workflow and toolkit¹. The system supports the statistical data transformation workflow, by adopting ETL (Extract, Transform and Load) methodologies and technologies, enabling the automated and integrated transformation of input data into statistical output data for dissemination. These data are then loaded into a generalised database, independently of the statistical domain. This type of process organisation improves the timeliness and coherence of the dissemination process, both minimising the delay between data checks and publication – i.e. the moment that data are returned to the community in a usable form – and, thanks to the high degree of automation, reducing the probability of human error during calculation and storage of aggregate data in the dissemination database.

From the point of view of the typology of data handled, the structure of ISTAR is based on several levels and kinds of statistical data and metadata. With respect to the data layers, ISTAR is able to manage both elementary and aggregated data. The system offers a set of packages to extract data from statistical sources, transform them into manifold formats, load the data into statistical data warehouses or data banks and make the information available to many different users, by means of different types of dissemination channels and technologies. The metadata layers cover not only the description, the design and the reference of the contents, but are also oriented towards

¹ The term *toolkit* is commonly used in the computer programming domain to denote a collection of generalised tools and components, which can be used to implement a unified system, customized according to the user's specific needs and requirements. Toolkits are usually made available as libraries or application frameworks.

the management of the navigation, the finding, the interchange and the semantics of the data. In detail, the main typologies of ISTAR metadata are:

- Statistical metadata: they are centred on the description and management of meaning and role of the statistical data in the system. Through the use of generalized packages and databases, statistical metadata accompany the data along the statistical surveys life cycle, from the validated elementary data environment to the dissemination on the Web;
- Reference metadata: this component is handled by a specific module for the integration of documentation metadata. They are stored in the SURVEYS DOCUMENTATION INFORMATION SYSTEM (SIDI) and in the QUALITY SYSTEM (SIQUAL), strictly integrated with ISTAR;
- Controlled vocabularies and glossaries: these components are managed by a semantic thesaurus and a specialized glossary of statistical terms directly associated to the statistical information contents of ISTAR;
- Metadata for searching: this layer enables the retrieval of collections of information available in several formats and digital supports (electronic publications, press releases, spreadsheets, databases, and so on) through a specialized search engine.

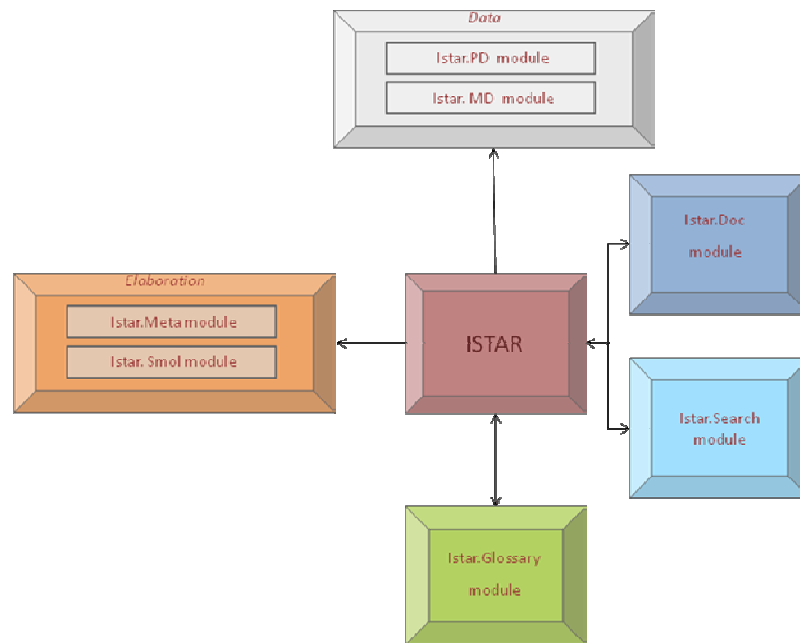
In order to manage all these kinds of data and metadata, a collection of tools strictly integrated has been developed. As mentioned above, the toolkits are specifically designed to support the statisticians in all the phases required to disseminate statistical aggregate data on the Web. From the functional point of view, the collection of toolkits is structured in two different kinds of packages: modelling tools and analysis and reporting tools. Modelling tools allow the user to design the semantic layers of the system, through the mapping of the structures of data sources (not easily understandable by the end users) into statistical outputs specifically oriented to describe the subject matter domains closer to the user language. From the application point of view, the modelling tools include both tools for managing on line interaction with designers and batch procedures for running ETL functionalities. Analysis and reporting tools provide navigation tools, in-house or publication on the Web of the data warehouse contents. In the next section we will present in detail the architectural framework of ISTAR.

4. The ISTAR framework

The project ISTAR was born in the year 2004 with the intent to deploy a whole software package allowing the statistic production sectors to elaborate and to spread the data on the Web, starting from the validated microdata.

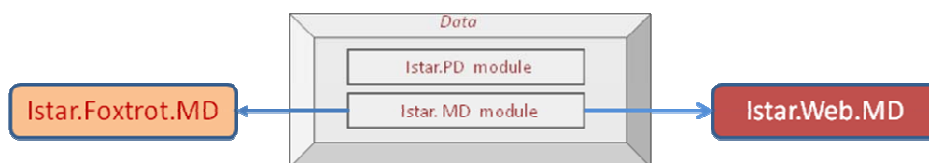
ISTAR, as represented in figure 1, is composed of seven modules: modules to elaborate the data (Elaboration - *Istar.Meta* e *Istar.Smol*); data module to publish the tables on the web (Data module - *Istar.PD*); data warehouse module (Data module - *Istar.MD*); glossary module (*Istar.Glossary module*); documentation module (*Istar.Doc module*); search engine module (*Istar.Search module*).

Figure 1: ISTAR framework



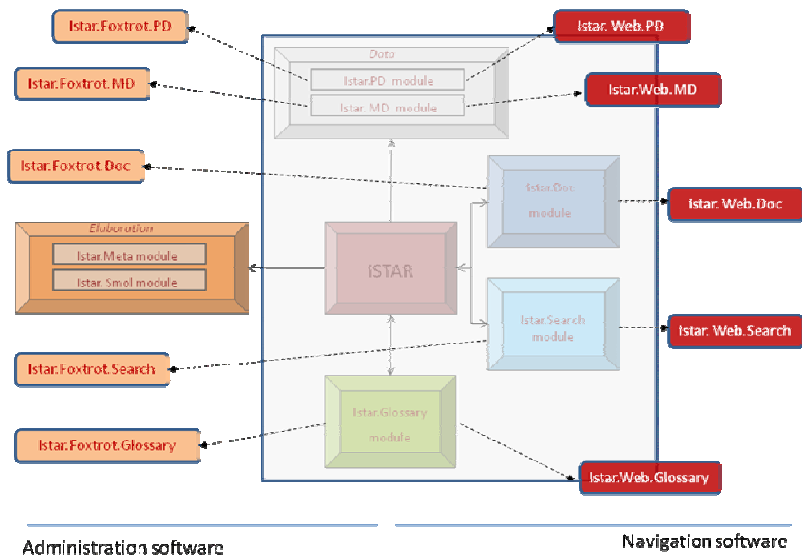
Every module, except the first one that is meant for exclusive use of the ISTAT production sectors, is composed of two software components: the administration component, named with *foxtrot*, to create the web site, and the software, denominated with the prefix *web*, to surf among the data or metadata or documentation. Accordingly, *Istar.MD* has, for instance, two components: *Istar.Foxtrot.MD* to design and feed a data warehouse and *Istar.Web.MD* to navigate its data on the Web (figure 2).

Figure 2: Istar.MD



The set of *Istar.Foxtrot* components represents the dashboard with which it is possible to create a complex web site, while, the set of *Istar.Web* components represents the different ways of navigation and retrieval the information publish on a web site. The ISTAR modular structure allows to choose which components to use and which on line services to offer to the external users. For instance if we want to create a data warehouse system, we have to use the *Istar.MD* module and its two software components (*Istar.Foxtrot.MD* and *Istar.Web.MD*). If we want to also equip such system with the glossary we have to use the *Istar.Glossary* component with the administration and consultation tools.

Figure 3: ISTAR framework detail



In the following a brief description of the different components is given:

Istar.Meta = it allows to pass from a sequential structure of a microdata file to a Data Mart organization. Such component is linked with the Centralized system of validated microdata;

Istar.Smol = it allows the elaboration of the data and the creation of tables (tables for the press, excel tables, html files, xml files);

Istar.PD = it allows to publish statistical tables (also in excel downloadable format) making them navigable according to their topics, years and territories;

Istar.MD = it allows the realization of a Web data warehouse environment;

Istar.Glossario = it allows the deployment of a glossary of the terms used on the Web site;

Istar.Ricerca = it allows to equip our own system with a search engine. Such component has been implemented through the use of Google Search Appliance (GSA) and, therefore, not completely free;

Istar.Doc = it allows to equip the system with a series of information related to the documents and to the surveys to which the data refer. Such software is the connection to the central documentation system of the ISTAT surveys (SIQUAL).

The ISTAR project originated from some previously developed software applications, which were evolved and integrated also with other ones. Such software applications are *Dawinci.MD* and *Web.PD*, respectively the first versions of *Istar.Web.MD* and *Istar.Web.PD* components. Such components were developed to publish the 2001 census data on the Web.

Starting from 2004, various systems were implemented by the Italian Institute of Statistics within the ISTAR framework. During this experience several different ISTAR versions were released but the adjustment of the systems already realized to the last consolidated version and the realization of new was planned for 2009 and 2010. In the

following a brief synthesis will be given of the deployed systems, available on Internet, and of in progress systems, currently on the Web.

Beginning from the experience of the 2001 Census the first realized system, in 2005, was an international collaboration with Bosnia-Herzegovina. The National Institute of Statistic, in fact, collaborated for the realization of the data warehouse related to household budget survey realized in Bosnia-Herzegovina during 2004. This system is available on the Web at the address <http://hbsdw.istat.it>. Such experience is, today, in phase of consolidation through a new step of collaboration. We are releasing the whole *Istar.MD* module to the national institute of Bosnia-Herzegovina, in order to allow the statistic users to realize their own data warehouse environment related to 2007 household budget survey.

During 2005 the ISTAT has respectively realized the systems related to the data on the law I.S. and Water Census I.S. The year 2006 saw the release of the systems about graduate employability (<http://dip.istat.it>, <http://lau.istat.it>). During this year the National Institute of Statistics began to experiment the realization of thematic multi source systems. These are systems into which various surveys publish altogether their own data about a topic. Particularly, in collaboration with the Ministry of the Social Politics, the INCIPIIT system was carried out (<http://incipit.istat.it>). The aim of this system is to offer a series of local information to support the local politics. The realization of such system required the evolution of the *Istar.MD* component, passing from the original system based on a single year and territorial aggregation type to a new one enabling multi-year and multiple territorial aggregation type management.

The positive experience of the multi-source system, as well as the external requirement for thematic systems, has brought the institute to reiterate it during the years 2007 and 2008 building further thematic environments. Particularly, it is now available on the internet the system related to the agriculture and zootechnical data (<http://agri.istat.it>) while further systems on the job market and on foreigners and immigrants are in the validation phase.

The modular organization of the ISTAR framework enables every project to use the software components which are more appropriate for its own context in terms of objectives to reach, time, resources as well as typology of data. More and more projects, however, are envisioned to use the framework in its entirety with the objective of creating a Web system that has contextually the table navigation component, the data warehouse, the glossary, the documentation, as well as the search engine. The above mentioned systems on the job market and on foreigners and immigrants have been carried out in this perspective.

Finally, the demand of software reusability in different linguistic contexts and for different DBMSs led the institute to an operation of maintenance of ISTAR modules. Accordingly, the *Istar.MD* module is currently available on both Oracle and MySQL databases, as illustrated in the following section.

5. *Istar.MD*

Istar.MD 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. In the following we illustrate *Istar.MD*'s main

characteristics, as well as the motivations underlying its development, based on open source technologies.

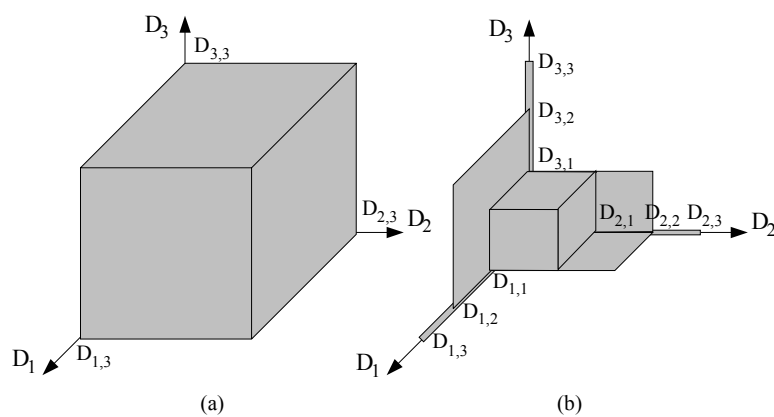
5.1. *Istar.MD* modelling basic concepts

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, in particular by modelling microdata using *star schemas* (Kimball, 1996) and by navigating the corresponding aggregates (*data cubes*) and classifications (*dimensions*) using commercial DWHs.

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 Tininini, 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 4, 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 4: *accessible dimension combinations in (a) a conventional data warehouse and (b) a Statistical Dissemination System*



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. In other words, a trade-off is

required between the characteristic freedom and flexibility of DWH multidimensional navigation and the constraints arising in the statistical dissemination context. This trade-off was achieved in *Istar.MD* by modelling aggregates in terms of object-classification(s) combinations (*basic-tables*) and *spatio-temporal instantiations* of these basic tables, as well as by precisely defining which instantiations have to be made accessible to the users (and conversely which should not):

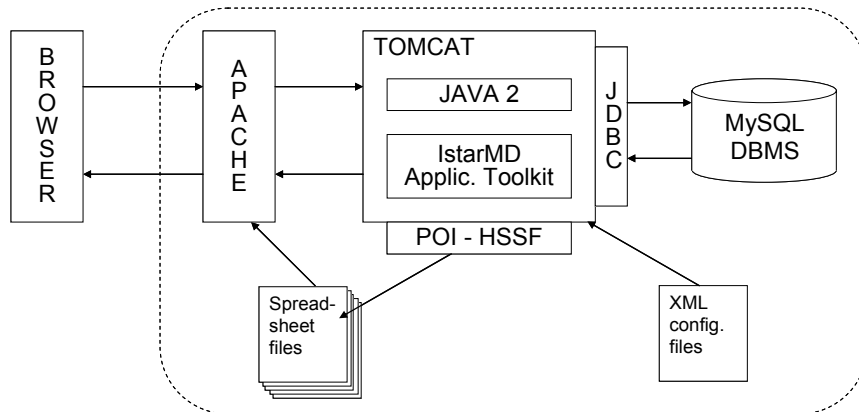
- *Istar.MD objects* basically correspond to measures in a conventional data warehouse, although there are some subtle differences, mainly related to the fact that an object may also incorporate some slicing operations on the data cube. For example in a conventional data warehouse “Resident population” and “Resident population aged 6 and over” refer to the same measure, whilst in *Istar.MD* they may be modelled by two distinct objects if they are related to different classifications, arising from filter questions in the questionnaire.
- *Istar.MD classifications* basically correspond to dimension levels in data cubes, although the structure of a classification can be more complex and articulated with respect to usual flat dimension levels. Examples of classifications of the object “Resident population” may be “sex”, “civil status”, “age by single year”, etc.
- A *basic table* (b-table for short) is nothing but the combination of an object with a (possibly empty) list of classifications. For example the combinations (“Resident population”; “sex”, “civil status”), (“Resident population”; “age by single year”) but also (“Resident population”;) are examples of b-tables.
- *Spatio-temporal instantiations*. In fact, specifying a b-table is not sufficient to precisely identify a collection of aggregate values, as two further components need to be specified: the territory and time to which the data refer to, e.g. European Union and year 2009. Hence we say that a single b-table can have many different *spatio-temporal instantiations* and only the combination of a b-table with a specific pair (territory, time) actually identifies a precise collection of aggregate values. For example, given the b-table BT=(“Resident population”; “sex”), its spatio-temporal instantiation (BT, European Union, 2008) identifies 2 aggregate values corresponding to the female and male resident population in the European Union in 2008.

All data that have to be disseminated by *Istar.MD* are expressed in terms of the above mentioned concepts. In particular the dissemination administrator can define the allowed combinations depicted in Fig. 4(b) in terms of a certain number of *maximum detail b-tables*, namely those representing the corners of the gray zone in the figure. In practice a maximum detail b-table is defined by the combination of a b-table with a certain year and a (maximum) territorial detail. Consequently, the same b-table may be disseminated with different maximum territorial details in different years, e.g. up to the regional territorial detail in 2008 and up to the municipality detail in 2009.

5.2. Open source software architecture

Both the navigation and administration component of *Istar.MD* are based on the same multi-tier (open source) software architecture, depicted in Figure 5.

Figure 5: *Istar.MD software architecture*



The *presentation tier* is the part of the application responsible of the interaction with the user and is based on a conventional Web browser (e.g. Internet Explorer, Firefox, etc.) and HTML. The browser interacts with the application (business logic) tier through a series of usual Web pages, producing HTTP requests and responses, i.e. the normal interaction between browsers and Web servers.

The *data tier* is where the application data and metadata are permanently stored, updated and retrieved. In the system's original deployment this was constituted by the Oracle proprietary DBMS. *Istar.MD* current deployment can support either Oracle or MySQL open source DBMS, interchangeably. The possibility of adopting other open source DBMSs (e.g. PostgreSQL or a light-weight and very portable DBMS like SQLite) is currently under study.

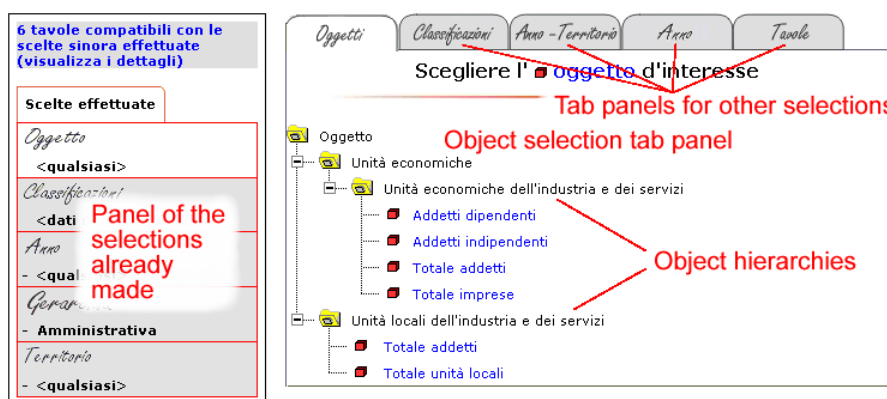
The *application tier* is where the application's functionality is implemented and is constituted by several modular/layered components:

- The *Java software platform* can be considered as the "glue" of all application components and is constituted by a number of software products (the Java programming language, libraries, compiler, run-time environment, etc.), enabling the user to develop application software and deploy it in a cross-platform environment. Java has always been free to be used, but was not originally open source. Its transition to truly open source software started in 2006 and was basically completed in 2008 thanks to the third-party project, IcedTea.
- The top-most component (i.e. closest to the presentation tier) is the *Apache HTTP Web Server*, released under an open source license and available for a wide variety of operating systems. Its fundamental task is to reply to requests received from browsers on the Web according to the HTTP protocol. In some cases the reply is simply constituted by (static) HTML pages (or more generally files), already available on the server, but in most cases the request involves *Istar.MD* dynamically generated HTML pages and is forwarded to the Tomcat servlet container (see below).
- *Apache Tomcat* is a JSP/servlet container, i.e. a server which is able to handle Java Server Pages and Java Servlets. These are particular forms of Java applications, which are able to reply to HTTP requests, by producing dynamically generated HTML pages in response.

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.

The initial selection of the table to be visualised can be performed by using the table selection page shown in Figure 7. This page enables the user to express the required table by selecting (without a predefined order and possibly only in part) the single components of the corresponding b-table instantiation, i.e. the object, classifications, territory and year of interest. In order to guide the user in selecting the required table, objects and classifications are organized into hierarchies, mainly based on generalization relationships, and the user can choose “generic” concepts, 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.

Figure 7: *WebMD statistical table selection page*



5.4. The Foxtrot.MD administration component

Foxtrot.MD is *Istar.MD* administration component, specifically designed for metadata management and aggregate data computation. By *Foxtrot.MD* the dissemination administrator can:

- manage the *objects* and *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 object or tables defined using a given combination of classifications). As mentioned above, objects and classifications can indeed be arranged into a hierarchical structure based on generalization.
- 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). *Foxtrot.MD* 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 *Foxtrot.MD* can aggregate the data and store them in the aggregate data table used during statistical table visualisation by *Web.MD*. The aggregation process is automatically performed at all levels of the territorial partitioning hierarchy specified by the administrator.

Figure 8: *Foxtrot.MD* user interface for ETL procedure management

Lista Gerarchie list of territorial hierarchies

Codice	Descrizione	Sigla Gerarchia	Dettagli
1	Amministrativa	cod1	
2	Sistemi locali del lavoro del 2001	cod2	

Lista Oggetti list of objects to be calculated

Codice	Descrizione	Stato ETL	Visualizza Errori	Reset	Esegui	Esegui a fasi	Gestione DWP	Gestione Qnt e Mr
mtr	Matrimoni	Mancano tavole						
mtrs	Matrimoni con almeno uno sposo straniero	Da eseguire	1					
...	2	3	4			
mtrsr	Matrimoni tra sposi stranieri di cui almeno uno residente	Eseguita Fase 1/7						

Figure 8 shows the user interface of the *Foxtrot.MD* for ETL procedure 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 objects 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 (*figure 8 – point 2*) if the warnings actually correspond to what expected and possibly reset the process execution (*figure 8 – point 3*) or enable its prosecution (*figure 8 – point 4*). In the latter case (*figure 8 – point 1*) 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 and Phase 2: in these phases the system verifies if the microdata territorial granularity (e.g. municipal, provincial, national, etc.) has been specified in the metadata and the microdata table structure. 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.

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. 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.

6. Conclusion

The Integrated Output Management System can be considered as the core of a global integration architecture of ISTAT, aiming at providing a seamless cooperation among: (i) local production systems, (ii) the set of reference and documentation metadata systems, (iii) the centralised repository of validated microdata and (iv) the environments for analysing and disseminating statistical data on the Web.

The recent adoption of open source software for managing the entire life cycle of dissemination, from the elementary data level to the aggregated data to be published on the Web, has encouraged the reuse of the implemented solutions also in the context of international cooperation. The current experiences in this direction have already shown that ISTAR is a valid alternative to commercial solutions in terms of architectural deployment. At the same time, it ensures a full control of all the steps needed to generate meaningful statistical output, providing a fundamental support from the methodological point of view.

References

- AA.VV. (2005a) – GENESEES v3.0 Funzione Riponderazione - ISTAT Tecniche e strumenti n.2-2005
- AA.VV. (2005b) – GENESEES v3.0 Funzione Stime ed Errori - ISTAT Tecniche e strumenti n.3 -2005
- Bergamasco S., Colasanti C., De Francisci S., Giacché P., Giacomini P. (2009) - An integrated approach to turn statistics into knowledge combining data warehouse, controlled vocabularies and advanced search engine. - NTTS Conference (New Techniques and Technologies for Statistics), Bruxelles 2009
- Gray, J., Bosworth, A., Layman, A., Pirahesh, H. (1996). Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Total. In *Proceedings of the International Conference on Data Engineering (ICDE'96)*. (pp. 152-159).
- Kimball, R. (1996). *The Data Warehouse Toolkit*. John Wiley & Sons.
- Lehner, W. (1998). Modelling Large Scale OLAP Scenarios. In *International Conference on Extending Database Technology (EDBT'98)*. (pp. 153-167).
- Malvestuto, F. M., Moscarini, M. (2003). Privacy in Multidimensional Databases. *Multidimensional Databases*, M. Rafanelli (editor). Idea Group Publishing. 310-360.
- Riccini E (2004) – CONCORD v1.0 Controllo e correzione dei dati - ISTAT Tecniche e strumenti n.1-2004
- Shoshani, A. (1997). OLAP and Statistical Databases: Similarities and Differences. In *Proceedings of the ACM Symposium on Principles of Database Systems (PODS 97)*, (pp. 185-196).
- Sindoni, G., Tininini, L. (2006) Statistical warehousing on the Web: navigating troubled waters. In *Proceedings of the International Conference on Internet and Web Applications and Services (ICIW'06)*, IEEE Press, 2006.
- Sindoni, G., Tininini, L. (2008) Statistical Dissemination Systems and the Web. *Handbook of Research on Public Information Technology*, G.D. Garson and M. Khosrow-Pour (editors). Information Science Reference, 578-591.
- Tillé Y, Matei A. (2007) sampling: Survey Sampling. R package version 0.8.
- Zardetto D. (2008) "EVER: Estimation of Variance by Efficient Replication". R package version 1.0, Istat, Italy.