

An integrated approach to turn statistics into knowledge combining data warehouse, controlled vocabularies and advanced search engine

Stefania Bergamasco, Cecilia Colasanti, Stefano De Francisci, Paola Giacché, Paolo Giacomini

Istat, e-mail: {bergamas, cecolasa, defranci, giacche, giacomini}@istat.it

Abstract

The aim of this paper is to illustrate:

- 1) the main components of the Integrated Output Management System of ISTAT (data module, doc module, glossary module, GSA module). In particular, the integration between the components for managing data and metadata will be focused;
- 2) the solution adopted to integrate the statistical data warehouse and the new ways to organize and retrieve the information on the Web;
- 3) the use of the principles of controlled vocabularies to manage the glossary of the system;
- 4) the technical solution adopted to optimize the search engine to scan the dynamic Web pages generated by the information system.

Keywords: Integration, Knowledge organization, Search engine

1. Introduction

The integration of the typical dissemination information system of statistical data with the documentation information systems, the semantic thesauri, the specialized search engine and the Web 2.0 technologies makes easier the retrieving process of the statistical data and supports the development of a collaborative environment in which the final users play an active role in the statistical information systems.

The Italian National Institute of Statistics (ISTAT) is involved in developing an integrated information system for the production of the statistical output (Istar) which combines the approaches of browsing the multi-dimensional data, typical of statistical data warehouse on the Web, with new models for the management and the representation of the knowledge and the information architecture.

The information system Istar has a twofold target: on one hand it allows the management of semi- or non-structured information as resources of the information system, on the other hand it makes statistical data accessible through searching functionalities. In the first case, the textual information can be associated to statistical data and retrieved by means of the multidimensional navigation, in the second case the structured data can be retrieved just like any other textual document.

This paper aims at illustrating the most important features of this approach, describing both the main functionalities of Istar and the issues of the integration between statistical data warehouse and the new ways to organize the information on the Web.

The paper is structured as follows: the section 2 describes the general structure of the Integrated Output Management System of Istat; the section 3 shows the main features of the search engine, while in the section 4 some final remarks and further issues are illustrated.

2. ISTAR: the Integrated Output Management System of ISTAT

The integrated system is based on the construction of several metadata layers. They cover not only the description, the design and the reference of the contents, but are also oriented towards the management of the navigation, the finding, the interchange and the semantics of the data.

The main components of Istar are:

- Data module: component to manage the life cycle of the multisource statistical data, starting from the validated elementary data environment to the dissemination on the Web, through the use of generalized packages and databases;
- Doc module: documentation environment, represented by the module for the integration of reference and documentation metadata stored in the Surveys Documentation Information System (SIDI) and in the Quality System (SIQual) with Istar;
- Glossary module: component to handle a semantic thesaurus and a specialized glossary of the statistical terms directly associated to the statistical information contents of Istar;
- GSA module: specialized search engine, dedicated to retrieve collections of information available in several formats and digital supports (electronic publications, press releases, spreadsheets, databases, and so on).

2.1. Data Module

The Data module of Istar is a collection of tools 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 is structured in two different kinds of toolkits: modelling tools and analysis and reporting tools. Modelling tools allow 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.

The application architecture of Data module of Istar is layered as follows:

- (a) OLAP engine: it makes available to the end-users specific functionalities for multidimensional navigation and dissemination on the Web. In particular, the user can start by displaying the data corresponding to a certain combination of measure (object) and dimension levels (classifications) and then navigate the 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 which 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.
- (b) Statistical data warehouse. The system is based on the definition of the maximum detail dimensional combinations of each global cube, basically corresponding to all permitted subcubes.

- (c) Administration module. It is the component specifically designed for metadata management and aggregate data computation (Istar.Foxtrot toolkit). Through this module the system designer can:
 - define and manage the structure of the statistical tables to be disseminated,
 - compute and store the aggregate data to be disseminated. By using the specified rules, the ETL component can aggregate the data and store them in the aggregate data table used meanwhile users visualize the statistical tables on the Web. The aggregation process is automatically performed at all levels of the territorial partitioning hierarchy specified by the administrator.
- (d) Metadata layers. They are represented by the collections of auxiliary data to perform both design and ETL steps and the navigation ones. Furthermore, all the descriptive information referred to statistical, technological, reference aspects are included in the metadata layers. They allow both mapping and description of the statistical data disseminated by the system.

2.2. Doc module

The Doc module manages the non-structured reference data and documents linked to the subject matter areas within the system. This module allows the connection and the interchange between Istar and the centralised system for surveys documentation (SIDI) and, in particular, its component dedicated to the Web dissemination (SIQual). The integration is possible through two navigation paths: on one hand the link to statistical source which feeds the system is enabled, on the other hand the information resources linked to the domain of interest are directly provided.

The documentation section makes available to the users a set of documentation materials of the specific domain of interest. It is possible to access to the documentation with a direct link to the information system of quality (SIQual). So, the feeding of the documentation section is guaranteed by the update of SIQual contents. There are two ways to update documents: direct way, updating the specific occurrence related to the domain of interest of the integrated information system, and indirect way, updating data of every source which feeds the whole system.

2.3. Glossary module

The glossary module allows the management of the descriptive information about the statistical contents of the system. It allows the building up of specific thematic glossaries, related to the corresponding domains of interest (foreigners and immigrants, labour market, agriculture and so on).

The glossary has been implemented as a controlled vocabulary and it is based on a linguistic structure which allows the association of the terms according to the semantic links. These links are based on the rationale of the thesauri such as equivalence, hierarchy, association.

In a relationship of equivalence among several terms each term is regarded as referring to the same concept. Such a relationship covers some basic types, as synonymy, near synonymy and lexical variants. Hierarchical relationships are based on levels of superordination (Broader Term) and subordination (Narrower Term), while more sophisticated approaches can include systematic presentations such as tree structures. Associative relationships are characterized by statistical links among the elements of the domain of interest like, for examples, analysis units, classifications, variables and context information.

The glossary runs through an ad-hoc application component which provides a set of functionalities for searching and connecting to the documents and to other information resources on line.


The levels of search provided are :

- **Alphabetic search:** by clicking on one of the available letter, the list of the terms starting from the selected letter will be visualized. Then, from the list of the available terms, it is possible to access the contents related to the diverse terms, by clicking on the related item. Each item of the glossary is identified by a set of predefined information, such as: *Statistics-semantics definition; Category; Definition source; Semantics correlated terms; Correlated surveys; Other useful information.* When provided, it is possible to access the further contents in the page by clicking on the related item.
- **Search by term:** this functionality offers two choices: *all* the terms and *each* term. The first choice allows to identify, from the selected string, only the terms which include in the same time all the terms, not dependent upon the rank or the syntactic structure of the diverse selected tokens. The second choice provides all the terms which include at least one of the terms selected in the search string.
- **The whole list:** it provides all the items of the glossary offering for each item only the official definitions used.

The most innovative issue of the glossary, namely to allow the direct link between the documents and the functionalities for accessing to the data, is the *table glossary*. This component provides the users with the definition of the most important terms used in each table of the system.

The screenshot shows the Istat.it website interface. At the top right, the title 'Stranieri e immigrati' is visible. Below the navigation bar, there is a section for 'Tutte le tavole dello stesso tema, territorio e anno:'. A blue box highlights a button labeled 'Table glossary button' which points to a small icon representing a glossary. Below this, a table titled 'Tavola SP01 - Perenni per motivo della presenza, sesso e principali cittadinanze al 1° gennaio. Dettaglio per continente - Anno 2007' is displayed. The table has columns for 'Continenti', 'Motivi della presenza', and 'Tavole (maschi) / Tavole (femmine)'. The 'Motivi della presenza' section includes categories like 'Lavoro', 'Famiglia', 'Refugiato', 'Residenza estera', 'Studio', 'Asilo', 'Richiesta asilo', 'Umanitari', and 'Altro'. The 'Tavole' section includes 'Totale', 'Totale (maschi)', and 'Totale (femmine)'. The table lists data for Europe, Africa, Asia, America, Oceania, and Apoliti, with a final 'Totale Paesi' row.

Continenti	Motivi della presenza										Tavole		
	Lavoro	Famiglia	Refugiato	Residenza estera	Studio	Asilo	Richiesta asilo	Umanitari	Altro	Totale	Totale (maschi)	Totale (femmine)	
Europa	707.984	371.004	9.279	38.095	24.077	2.769	4.021	2.101	14.843	1.174.173	511.530	662.643	
Africa	369.048	170.467	4.762	2.139	6.254	3.691	2.305	9.892	5.351	670.799	371.005	199.194	
Asia	272.248	116.448	9.687	1.367	12.554	1.990	1.030	1.179	2.813	419.964	231.407	118.607	
America	115.830	104.706	8.153	2.962	8.541	191	90	282	6.902	247.640	82.896	164.744	
Oceania	361	1.037	204	340	196	0	0	2	61	2.101	831	1.270	
Apoliti	84	84	6	44	3	12	0	11	51	265	153	132	
Totale Paesi	1.463.058	763.744	32.061	44.847	51.625	8.613	7.466	13.447	30.091	2.414.972	1.198.452	1.216.520	

By clicking on the button , the page of the glossary of the associated table can be accessed, as shown in the next figure:

The screenshot shows the glossary page for the 'Stranieri e immigrati' table. The page title is 'Guida alla lettura'. It contains the following sections:

- Informazioni sulla fonte dei dati:** 'Elaborazione Istat su dati del Ministero dell'Interno'. A blue box labeled 'Link to the data source' points to this text.
- Principali termini utilizzati nella tavola:**
 - Unità di analisi:** 'Permessi di soggiorno'. A blue box labeled 'Link to the glossary items' points to this term.
 - Variabili di classificazione:** 'Cittadini', 'Motivo della presenza', 'Area geografica'. A blue box labeled 'Link to the glossary items' points to these terms.
- Altre informazioni utili:** A section for additional information.

 Annotations include:

- A blue box labeled 'Specification of the statistical kind of the items' points to the 'Principali termini utilizzati nella tavola' section.
- A blue box labeled 'Analysis unit' points to 'Permessi di soggiorno'.
- A blue box labeled 'Statistical Classifications' points to 'Cittadini', 'Motivo della presenza', and 'Area geografica'.

Once visualized the page of the table glossary, the set of hyperlinks to the corresponding terms existing in the thematic glossary is at users' disposal and, consequently, all the information linked to each term is accessible. Such an approach allows to integrate the navigation among structured data, terms of controlled vocabulary and any textual documents referred to the statistical data of the table.



2.4. GSA module

Another innovative component of the system is represented by the search engine Google Search Appliance. Such a component enriches the system with the new Web paradigms, providing the development of functionalities able to combine the plainness of a textual research with the richness of metadata and diverse kinds of classification models. This meaningful enhancement takes place by means of an advanced and optimized scanning mechanism of the data warehouse component. Within a general reorganizational process of the data it handles, Istat has chosen Google products and services to match the search needs of both its internal and external users.

Google Mini was selected as the intranet search engine, while Google Search Appliance (GSA) is used to provide a query interface for all internet-accessible data. As the well-known Google Search Engine google.com, the GSA provides universal search across a variety of internal and external sources – including file shares, intranets, databases, applications and content management systems.

The GSA provides access to all of the information which users need through a single easy-to-use search box. With the latest version of GSA users can also customize according to their specific needs, or having the system administrator define pre-configured sets of search profiles, i.e. based on groups or users' organizational unit. Administrators also have the capability to bias results based on content metadata, while registered users may also subscribe to email alerts on specific documents or topics of interest.

GSA fully integrates with existing security and access control systems at the single document level: users are able to view search results only if they have access to the original content, thus ensuring that company data are always protected from unauthorized access. Several authentication and single-sign-on (SSO) protocols are supported, including Kerberos, LDAP, NTLM and base authentication, PKI

authentication using X.509 certificates and Windows Integrated Authentication. Physically GSA is packaged in an appliance including both hardware and software which is quickly deployed: by removing the need to configure hardware and OS specifics, it becomes immediately operational and can be easily managed by a single administrator.

3. The main features of the search engine

The search engine takes advantage of the three classification models: taxonomies, faceted classifications and, in our plans, folksonomies.

A simple taxonomy is adopted to perform the basic search and it allows the management of the collections of several kinds of information resources. Before to launch the search query, the user can choose one or more kinds of collections in which the search will operate. The engine will return all the documents and the data stored in the selected collections only.

Going into more technical details, we configured the product taking into account four disjoint collections: electronic documents, press releases, statistical tables in electronic sheet format and databases. It's also possible to choose whether to show or not the Metadata.



By searching, for example, “Population” into Publication and ticking off the Metadata item, you obtain the following snippet (Figure below):



The scheme will be enhanced with a further branching of the classification, according

to the principles of a hierarchical-enumerative approach: for example, the item “Publications” could be split into more specific disjointed items, like Yearbook, Subjects, Methods and Rules, Statistical Annals, Statistical Indicators, Essays and so on.

A different kind of classification is adopted for the advanced search mode. In this case, the engine is based on a multidimensional classification instead of a hierarchical-enumerative one. This kind of classification, in accordance with Ranganathan’s Colon Classification, is compliant with the PMEST model with some restrictions. The original model uses five primary categories, or *facets*, to further specify the sorting of a semantic description of a concept. Such categories are: Personality, Matter, Energy, Space, Time. In the case of Istar, in particular, each document of interest has been described by a set of elementary properties, each explaining a different topic (facet) of the subject, relevant for the scope of the search by GSA module. In other words, the advanced modality enables the selection of one or more focuses of the facets, combining one with each others and with the hierarchical-enumerative classification defined for the collections. In this way it is possible to perform a multidimensional approach for metadata too.

With regard to the PMEST model, in the first prototype of the system we are going to adopt the following facets:

- *subject matter areas as personality,*
- *statistical sources as energy,*
- *territories (both Italian administrative hierarchy and world geographical areas) as space*
- *year as time.*
- at the moment, the *matter* facet has not been defined in the model.

In the current phase, the system does not make use of a specific user interface to visualize the content of the database by means of a faceted classification tool, but it provides the users with the multidimensional facets directly via advanced search mode of GSA module. In this way it is possible to combine one another facet through the selection and to perform custom-cut searches.

By choosing “Advanced Search”, it is possible also to associate an item of the taxonomy with one or more facets (thematic area, statistical source, year, territory).

The image shows a screenshot of the Google Advanced Search interface. At the top left is the Google logo. Below it, the title "Ricerca avanzata" is displayed. The main search area contains several input fields and options:

- Trova risultati** section:
 - Input: "popolazione" (with "10 risultati" dropdown)
 - Input: " " (with " " dropdown)
 - Input: " " (with " " dropdown)
 - Input: " " (with " " dropdown)
- che non contengano** section:
 - Input: " " (with " " dropdown)
- solo in** section:
 - Yellow box containing: "Publication", "Press Release", "Data Base", "Data Table"
 - Checkboxes:
 - Pubblicazioni
 - Comunicati: Stampa
 - Banche dati
 - Tavole di dati
- Facets** section (yellow box):
 - Area tematica
 - Fonte statistica
 - Anno: 2001
 - Territorio
- Additional options** (bottom):
 - Lingua: tutte le lingue
 - Formato file: qualsiasi formato
 - Cerca in: in una qualsiasi parte nella pagina

and to obtain the following result:

Cerca Risultati 1 - 2 su circa 13 per **popolazione**. Durata della ricerca: 0.01 secondi.

La ricerca attuale è filtrata con i seguenti criteri:

- Pubblicazioni

[Ordina per data](#) / [Ordina per importanza](#)

Scheda volume
Popolazione e movimento anagrafico dei comuni. Anno 2000 Settori: **Popolazione**
 Periodo dei dati: Anno 2000 Collana: Annuari, n. 13 ...
 Anno: 2001
 Titolo: Popolazione e movimento anagrafico dei comuni
 Tipologia: Pubblicazione

Volume card
 Population and personal data...
 Year: 2001
 Title: Population and personal data
 Taxonomy: Publication

www.istat.it/dati/catalogo/schedavolume.php?ID=172-4k - Copia cache

In this case, we used for the snippet the feature of “Dynamic Page Summaries” that shows in the same page both the user query and the metadata.

A further innovative feature of GSA module is represented by the scanning mechanism of the web pages and the consequent control of the crawling.

We faced with the following problems:

1. how to exploit the indexing of documents by the engine search of GSA through web pages which allow drill down, roll up and slice & dice functionalities, namely within dynamic web pages of a Data Warehouse.
Each web page with these features, in fact, provides a wide number of links for the navigation through the metadata and a little number of *link for linking to service pages*, so that a lot of records are generated in the scan database of the search engine provoking rapidly the saturation;
2. how to exploit the concepts of taxonomy and facet enabling the user to retrieve the information, independently from their own repository;
3. how to provide search result with the related metadata within the snippet.

We adopted a solution based on three concepts:

1. we have associated to each “object” stored in the system all the metadata useful for managing the taxonomy, the facets and the snippet;
2. the potentialities of the search engine have been exploited not in reference of the scanned web pages but for the scanned database;
3. we tagged as “non crawling “ all the visited Web pages.

Going more in details: within each system we built a relational table called `search_engine`. This table is populated through specific procedures invoked by “insert, modify, cancel” events. The table has the following fields structure (Clob):

1. typology (*e.g.*: publications, press releases, databases, etc.)
2. theme
3. data source
4. territory
5. time (*i.e.*: the year to which the data are referred)
6. title
7. other (*e. g.* the modes of the classifications associated to the table)
8. url of the indexed object

The field *territory* contains the details of the spatial references, in particular territorial level (*e.g.*: Regions or Provinces) and geographical denominations (*e.g.*: Rome and Milan, or Lazio and Sicilia).

With regard to the objects “database”, in order to enable the user to get the table with the territorial detail required, each table contains the records referred to all the regions and all the geographical areas, as shown in the following figure:

RIPARTIZIONI GEOGRAFICHE ▲ ▼	N° alberghi		REGIONI ▲ ▼	N° alberghi
Italia Nord-Occidentale	6.538	Drill down	Piemonte	1.514
Italia Nord-Orientale	14.550		Valle d'Aosta	492
Italia Centrale	6.324		Lombardia	2.898
Italia Meridionale	4.270		Liguria	1.634
Italia Insulare	1.845		Trentino-Alto Adige	5.944
Italia	33.527		Veneto	3.079
			Friuli-Venezia Giulia	736
			Emilia-Romagna	4.791
			Toscana	3.002
			Umbria	554
			Marche	967
			Lazio	1.801
			Abruzzo	806
			Molise	106
			Campania	1.536
			Puglia	831
			Basilicata	224
			Calabria	767
			Sicilia	1.068
			Sardegna	777
			Italia	33.527

We also customized the search engine of GSA and the search interfaces.

For the first operation a specific functionality to enlist the databases and the tables to be scanned and indexed has been performed. Furthermore the URL of the related objects has been enlisted.

In this way we obtained that the search engine does not scan each web dynamic page of the system, but it scans the contents of fields in the database, with a considerable savings in terms of number of records the search engine is building inside the system and avoiding its saturation. This solution has avoided the great amount of the scanned page due to both the dynamic navigation links and the service ones, e.g. “Useful links”, “Contacts” and “Glossary”.

With regard to the second operation, the search interfaces have been customized using the field *typology* as hierarchical-enumerative classification, the fields *theme*, *data source*, *territory*, *time* for the faceted classification and the fields *time*, *title* for the snippet.

Through searching metadata the system allows the users to move with both controlled browsing functions and evolving search which allow to identify, step by step, the information needed in the field of the search, such as an evolution of *berrypicking* process (Bates, 1989).

Finally, through the tools offered by the search engine, it is also planned the development of a collaborative environment according to folksonomy models, in order to enrich the controlled vocabulary with the common terms used for the searches by the users. The adoption of textual mining techniques will allow to organize the information needs of the different stakeholders (users, experts, statistical operators, etc) and to use the results in order to improve the information framework of the system.

4. Conclusions

The strategy adopted by Istat to increase the value of the information at its own disposal 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.

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. Furthermore, the system combines the new opportunities offered by the Web 2.0 technologies with the capability of dynamic Web Warehouses.

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.

References

- ANSI/NISO (2005) *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*, Z39.19-2005 ISBN: 1-880124-65-3.
<http://www.slis.kent.edu/~mzeng/Z3919/index.htm>
- Bates, M. J. (1989) The design of browsing and berrypicking techniques for the on-line interface. *OnlineReview* 13 (5) 407-424
<http://www.gseis.ucla.edu/faculty/bates/berrypicking.html>
- De Francisci, S., Sindoni, G., Tininini, L. (2006) Multidimensional Statistical Data Dissemination on the Web, *Seminar on the Management of Statistical Information Systems (MSIS)*, Sofia, Bulgaria, 21-23 June 2006
- Gnoli, C., Marino, V., Rosati, L. (2006) *Organizzare la conoscenza: dalle biblioteche all'architettura dell'informazione per il Web*, Milano: Hops-Tecniche nuove
- Negrini, Gigliola (2003) *Principi filosofici per classificare: una teoria per la scienza*. [Journal Article (On-line/Unpaginated)] <http://eprints.rclis.org/10944/>
- Rizzo, F., De Francisci, S. (2007) An integration approach for the Statistical Information System of Istat using SDMX standards, *Meeting on the Management of Statistical Information Systems (MSIS 2007)*, Geneva, 8-10 May 2007
- 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