Distr.
GENERAL

Working Paper No.2
26 April 2007

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2007)
(Geneva, 8-10 May 2007)

Topic (i): Governance and management of statistical information systems

AN INTEGRATION APPROACH FOR THE STATISTICAL INFORMATION SYSTEM OF ISTAT USING SDMX STANDARDS

Invited Paper

Prepared by Francesco Rizzo and Stefano De Francisci, Istat, Italy

I. INTRODUCTION

1. The statistical production activities of Istat are supported by a distributed architecture. Several production Directorates operate through local subsystems that, independently one from the other, cover the full life cycle of statistical data, from data collection to the dissemination. For this reason, the need to harmonize and integrate these systems has become mandatory, with respect to both the production process and the common statistical contents of the existing systems.

2. In order to reach this goal, Istat is currently involved in the development of the INTEGRATED OUTPUT MANAGEMENT SYSTEM, an Information System oriented towards the integration of part of the statistical data life cycle, particularly all the steps needed to produce purposeful statistical outputs for end users[1].

3. The aim of this paper is to show the main features of the System, the most important results already achieved and the future perspectives of the project, overall regarding the impact of the new approach on the management of; production, interchange and integration processes of the statistical data in Istat.

II. OVERVIEW OF THE INTEGRATED OUTPUT MANAGEMENT SYSTEM OF ISTAT

A. General requirements and foundations of the system

4. The Integrated Output Management System has been developed to maintain, integrate and manage the data and metadata supplied by the statistical production areas of Istat after the validation processes. Particularly, this System is able to extract both elementary and aggregated data from the sources, transform them into

[1] Following the term "information" refers to both data and metadata.

multidimensional format, load the data in statistical data warehouses and make the information available to many different users, by means of different types of dissemination channels and technologies.

5. The project is based on some experiences already carried out in the last few years, particularly represented by: metadata information systems; generalized environments that perform OLAP functions on the Web; thematic databases.

6. The high level of application and technological heterogeneity of the involved systems has precluded a full integration of the component mentioned above. For this reason, the system has been configured as a multilevel and multiservice integration environment. It is supplied with flexible and multiple mechanisms for interchanging, sharing and integrating data towards a corporate vision. Such mechanisms are focused on:

- Service Oriented Architectures
- Cooperation applications
- Generalized packages
- Adoption of international standard for data and metadata exchange (compatible with the SDMX standards)
- Corporate data warehouse.

7. To reach these goals we have had to make the following essential choices:

- High flexibility of the system with respect to the evolution of requirements;
- Generalization of procedures and applications;
- Technological independency;
- Adoption of a Workflow mechanism to manage application processes;
- Separation between environments that manage pre-aggregated data from dynamic aggregations.

8. One of the main features of the system is its own position in the Istat scenario. The Integrated Output Management System has been placed in the middle of a global integration architecture, represented by local production systems, the whole set of reference and documentation metadata systems, the centralized repository of validated microdata and the environments for analysing and disseminating statistical data. The main components of the system are:

- Management, harmonization, sharing of the statistical data and metadata of the surveys
- Interchange of data between several production pipelines
- Data extraction and transformation in outputs
- Development of generalized packages to enable internal users to build statistical data warehouses
- Carrying out navigational and querying environments, such as:
  - *thematic databases*,
  - *OLAP applications*,
  - *Retrieving data integrated services*.

B. System architecture

9. From the architectural perspective Istat is carrying out an integration *framework*, which focus on statistical outputs management, that will be extended to cover the statistical production phases and will also incorporate the other centralized repositories.

10. Regarding data, the system architecture has been structured on:

- Documentation metadata database
- A specialized layer (an interface among all systems that manage metadata) for integrating metadata
- Statistical Data Mart oriented to specific subject matter domains
- Primary Data Warehouse of microdata
- Web Warehouse of aggregated data (in multidimensional structure)

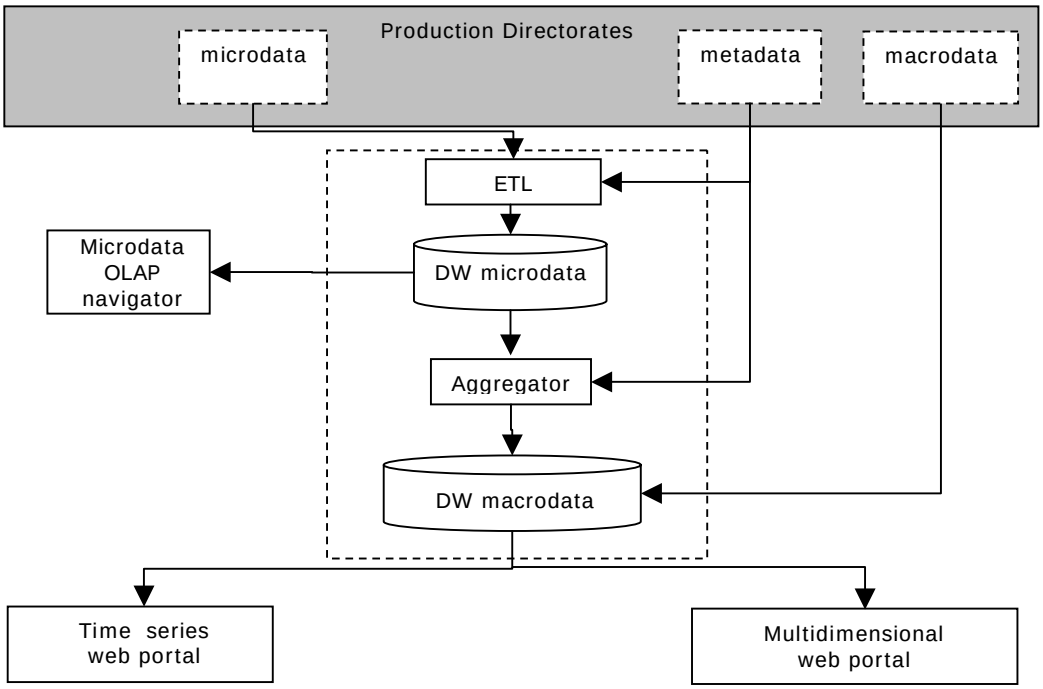
- Web Warehouse of time series data
- Web Warehouse of predefined statistical tables
- Spatio-temporal database for integrated territory managing (with GIS application).

11. Regarding statistical contents, the system deals with:
    - Validated microdata
    - Aggregated data in time series structure
    - Aggregated data in territorial cross-sectional structure
    - Aggregated data in multidimensional structure
    - Ready-made statistical tables expressed in homogeneous structures.

12. Regarding applications, the System is structured as following:
    - Metadata management and integration functions for building up the semantic layers[2] to be adopted during the phases of on-line analysis
    - Extracting data from outside sources
    - Transforming data to fit business needs compliant with the semantic layers
    - Loading data into the data warehouses
    - On-line aggregation and analysis processing
    - Access, navigation and querying through end-user interfaces
    - Maintenance, customization and set-up functions.

13. Technological profile is characterized by:
    - Use of international standards (SDMX) and experiences in similar projects (SODI)
    - Web Services

14. The following diagram shows the workflow of the entire life cycle managed by the System.



[2] A SEMANTIC LAYER is a business-oriented representation of the data with respect to a specific subject matter area. It helps end-users of data warehouses to retrieve information by resorting to a set of common terms adopted by the community interested in the specific domain.

III. NEW TECHNOLOGIES AND INTERNATIONAL STANDARDS TO FACILITATE THE INTEGRATION PROCESS IN ISTAT

15. In any organization, the system and the process integration is one of the most complicated operations that, sooner or later, the informatics manager must deal with. In the last few years we've had solutions that resulted in being "proprietary" and very complicated to implement. Often changes in the existing system were so substantial that many times the entire operation was too expensive to build.

16. For some time, technologies like XML and Web services have been aiding to simplify integration processes. Moreover, in the statistical information systems area, the SDMX initiative is active and deals with data and metadata sharing that depicts a big contribution to solve the integration systems problems.

17. Istat is following the evolution of the SDMX initiative with interest, taking into consideration these two ways:

- tactically, through the participation in the EUROSTAT SODI initiative;
- strategically, through the building of a working group whose main activity is to analyze and verify the use of SDMX in the internal architecture of the Istat Information System.

18. The experience that we have been gaining by participating in the SODI project will allow us to support the strategic interest of SDMX in Istat. In order to facilitate this aim we are achieving a framework consisting of various compatible SDMX software modules. The SDMX Istat Framework can be used entirely, from the reporting phase to the dissemination phase, or alternatively using the modules separately, integrating them into one information system.

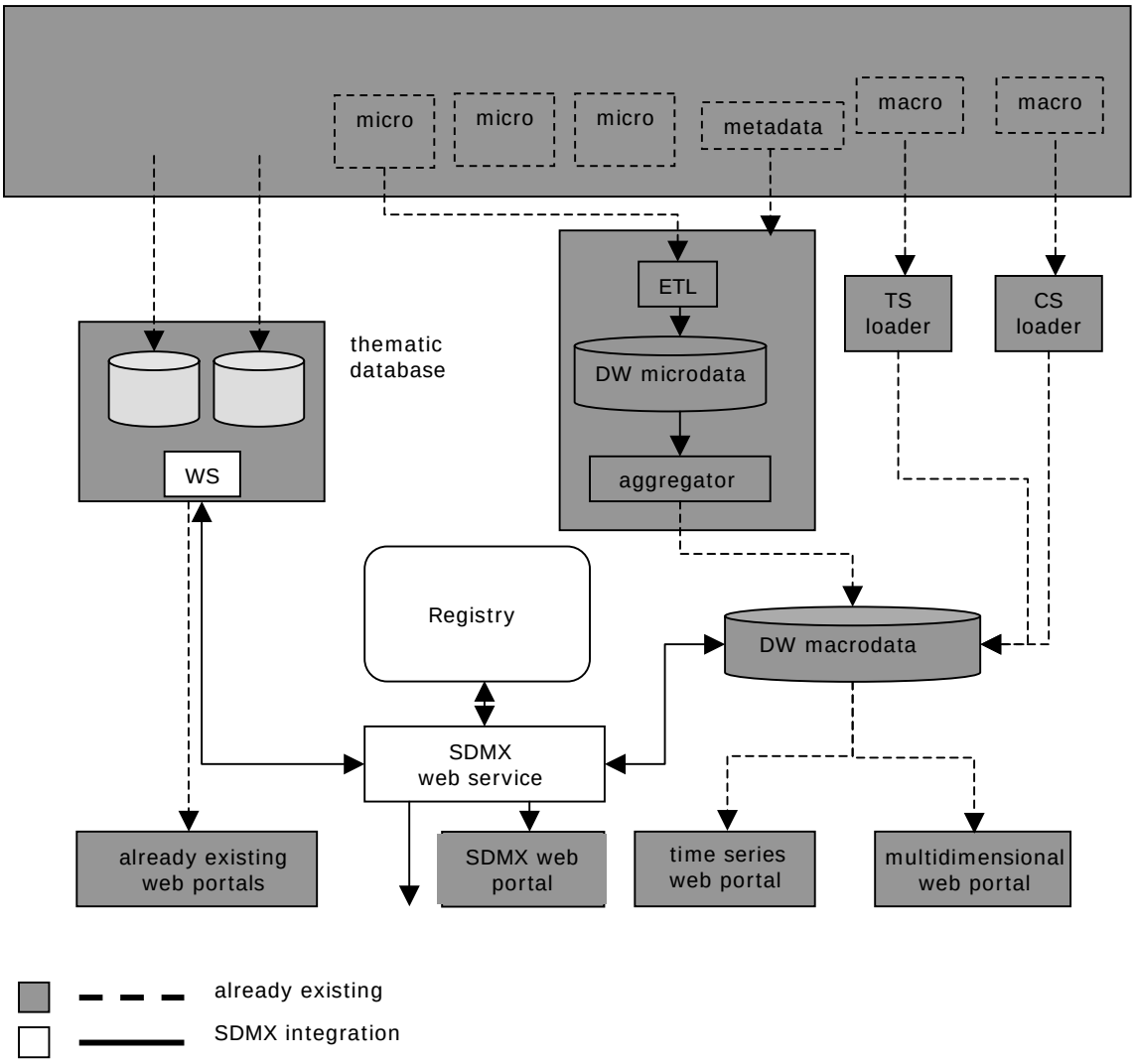
A. SDMX Istat Framework

19. SDMX Istat Framework was developed extending and adapting the existing Istat time-series database to the SODI project needs. Particularly the SDMX Istat Framework is made up of modules that perform the following functions:

- collect data through the Check/Loader module that accepts a fixed length records file format or GESMES file format;
- store data using Data Structure Definitions (DSD) supplied by Eurostat;
- publish data through a web service that:
  - accepts an SDMX query;
  - parsers and interprets an SDMX query and converts it in an SQL query;
  - queries the database using the DSD;
  - responds with an SDMX compact
- publish a RSS file that informs when new data is loaded or updated and which then specifies the URL where to find the files containing the SDMX Query which then locates the new or updated data;
- publish one o more SDMX queries.

B. Integration process architecture and governance

20. The portion of the Istat information System that will be interested by the integration process, described in the following, concerns the thematic databases, the centralized time-series database and multidimensional database. The integrating operation will have a limited impact on the existing sub-systems that will continue to work like they did before the intervention, but will allow the access of data in a standardized uniform way using a unique interface. Furthermore, in this way the international and national organization needs, to access data using SDMX technologies, will be satisfied. The following diagram shows where the integrating operation will take place inside the already existing Istat information system.



WS web service; TS time-series; CS cross-sectional; DW data warehouse

21. It is necessary to remember that the portion of the Istat information system under this type of integration is that related to “macrodata”. In the diagram above “Registry”, “SDMX web service”, “thematic database web services” and “SDMX web portal”, all use some of the SDMX Istat Framework modules.

22. Following a processes and information flows schema with the annotation that all the traditional existing access channel can still be used:

- data, after the integrating operation, will be accessed by means of the “SDMX web service”. An authorized client as a third part information system (ex. the “Eurostat Pull Requestor”) or as the “SDMX web portal” from Istat Information System, can utilize the “SDMX web service”. “The SDMX web portal” allows one to navigate data through a web browser using SDMX statistical concepts.
- any request for data must be sent to the “SDMX web service” using a SDMX Query;
- SDMX web service queries the Registry to know where required data are stored;
- in the case of requested data stored in the “DW macrodata”, the “SDMX web service” directly makes a query to the database. Then it formats data in SDMX Compact (time-series) or SDMX cross-sectional, depending on the nature of the data;
- in the case of requested data stored in a “thematic database”, the “SDMX web service” acts as a client sending a request to the correspondent web service placed on the “thematic database”.

The last one sends the result data back to the caller. The “SDMX web service” formats data in SDMX Compact (time-series) or SDMX cross-sectional, depending on the type of data;

- the “SDMX web service” supplies a method to directly query the Registry to know what data has been registered. This method can be used both by “SDMX web portal” or by external client systems;
- each database (“thematic databases” and “DW macrodata”) supplies an RSS file with the aim to inform the Registry that new data is loaded or updated in the specific database. A RSS reader, that is part of the Registry, checks the RSS file and consequently updates the registry.

23. The integrating process is going on for subsequent phases, so changes can be managed in an optimized way reducing the effect on the existing sub-systems. Follow a schematic enumeration of the subsequent phases:

- introduction of SDMX concepts and integration culture inside production Directorates;
- prototyping of integration processes, involving the “DW macrodata”, using both time-series and cross-sectional formats;
- participating in the SODI project with the main objective to acquire a good SDMX knowledge;
- extending the feasibility analysis to the entire integration process;
- implementing communication between “DW macrodata” and “SDMX web service”;
- implementing “SDMX web portal” module;
- implementing the Registry;
- implementing one after the other the WS on each thematic database.

24. At the present moment only phases 1 and 2 are completed. Phase 3 is under construction. We hope to start other phases over the next few months.
