**Modernizing Data Integration Systems at Istat**

Piero Demetrio Falorsi, Istat Rome Italy, [falorsi@istat.it](falorsi%40istat.it)

Roberta Radini, Istat Rome Italy, radini@istat.it

Monica Scannapieco, Istat Rome Italy, scannapi@istat.it

Laura Tosco, Istat Rome Italy, tosco@istat.it

**Abstract**

*Istat has engaged a modernization programme that includes a significant revision of the statistical production. The principal concept underlying such an important change is the usage of a system of integrated statistical registers as a base for all the production surveys; this system will be in the following referred to as the Italian Integrated System of Statistical Registers (ISSR). The ongoing work for building such a system required a big investment on architectural and methodological aspects to guide the enterprise-level design of the system. In this respect, there are two major activities that have been engaged: (i) the design of the data architecture of the ISSR and (ii) the design of the processes to populate the ISSR. The data architecture of the ISSR has been conceived according to modern semantic integration approaches, with a strong emphasis on active metadata guiding the access to the system. The design of the processes has been performed by relying on existing standards, like GSBPM and GAMSO, and frameworks, like ESS Enterprise Architecture Reference Framework. In this paper, we will describe both strands of work, focusing on how quality aspects have been taken into account and specifically on the role played by official statistics and world-wide technical standards to ensure the quality of data, processes and underlying information systems.*

**Keywords:** Data integration, process modelling, data architecture

**1. Introduction**

Istat has engaged a modernization programme that includes a significant revision of the statistical production. The principal concept underlying such an important change is the usage of a system of integrated statistical registers as a base for all the production surveys; this system will be in the following referred to as the Italian Integrated System of Statistical Registers (ISSR).

A system of statistical registers consists of a number of registers that can be linked to each other. The ISSR has been conceptualized as consisting of:

* Base Statistical Registers, composed by a collection of statistics units belonging to populations relevant for official statistics. The variables characterizing such units are “core” variables, meaning that they are (i) highly identifiable and (ii) quite stable in time. In particular, the base registers identified for the ISSR: (i) Individuals, Families and cohabitations; (ii) Economic Units; (iii) Places; (iv) Activities.
* Extended Statistical Registers, which extend the information available for a population of a specific base register with other variables.
* Thematic Statistical Registers, the units of which are not bound to populations specifications, but rather have the objective of supporting statistics referred to more statistical populations.

The ongoing work for building such a system required a big investment on architectural and methodological aspects to guide the enterprise-level design of the system. This paper will describe two major design efforts related to ISSR, namely: (i) the design of the data architecture of the ISSR and (ii) the design of the processes to populate the ISSR.The data architecture of the ISSR has been conceived according to modern semantic integration approaches, with a strong emphasis on active metadata guiding the access to the system. The design of the processes has been performed by relying on existing standards such as GSBPM. In this latter case, a specification of activities complementary to GSBPM has been proposed in order to take into account activities that are specific for the ISSR.

The rest of the paper is organized as follows. Section 2 will detail the design of the ISSR, specifying both the data architecture and the process architecture. Section 3 will deal with quality issues in the ISSR. Finally, Section 4 will draw some concluding remarks.

**2. Design of the ISSR**

In the ISSR the management of metadata results to be more complex than the traditional metadata systems used to handle survey data. In particular, since the initial design phases, there is the need to integrate concepts belonging to different thematic areas. To manage these aspects we adopted a particularly promising approach called Ontology-Based Data Management (OBDM) (Lenzerini, 2011).

This approach has enormous advantages in terms of:

* industrialization of processes;
* data harmonization;
* data quality control;
* flexibility in adding new sources or modifying existing ones.

*2.1. Data Architecture*

The aim of the ISSR is to integrate the data of all the registers, to this end the units of the registers are uniquely identified within the system by a specific code. This information is used to connect each unit in a register to other units in the various registers of the system.

The ISSR data architecture (see Figure 1) can be divided into two layers: data virtualization and data consumption. Data virtualization is the new data integration approach, which offers flexibility and agility to quickly retrieve data from all registers on the fly. Data virtualization overcome in this way the limitations of older approaches such as ETL and data replication. This layer is composed by *Physical Integration* that combines data, using the unit identification link (the code), and by *Integrated Micro Layer* that integrates and transforms data, then exposing them as views.

The *Macrodata Layer* positions itself upon the Integrated Micro Layer and provides aggregated data to users to conduct analyses or business intelligence activities. The semantic level provides access to data via OBDM to micro and macro data.



Figure 1: Data architecture of ISSR

The Semantic layer is implemented via the OBDM in which three different components are identified:

* **Ontology**: formal, shared and explicit representation of the conceptualization of the domain of interest expressed through the formal language which makes it "machine-actionable".
* **Data sources**: consisting of sources heterogeneous both semantically and technologically.
* **Mapping**: set of rules expressing the correspondence between data and concepts/attributes of the domain of interest defined in the ontology.

This approach allows, not only to document the domain of interest, but also to realize and/or manage the semantical integration of the data in the ISSR.

The ontology design process is iterative and is based on the cooperation between domain experts and ontology designers. The ontology modeling phase is iterated until the ontology does not represent completely and correctly the domain. After the ontology’s validation, the mapping rules are defined allowing an integrated, ontology driven, access to data.

Within the statistical production process, there is a data stack corresponding to different stages of data processing. These are:

* Raw Data: area in which data coming from different sources (administrative, surveys, big data and others) are collected. Procedures for analyzing and evaluating the completeness and consistency of the sources are performed on these data.
* Working Data: in this layer data are integrated and processed. Specifically, the following steps are performed: (i) selection of data of interest, (ii) deduplication intra-source, (iii) standardization of the information content with respect to official classifications, (iv) reconciliation and analysis of the consistency of each record, (v) deduplication between different sources, (vi) recovery from other sources of missing information or assignment of information from one source to another basing on defined priority criteria.
* Validated Data: this layer corresponds to the registers in which validated data, output of processes performed in the working area, are stored.

*2.2. Process Architecture*

Statistical production processes are usually described by the Generic Statistical Business Process Model (GSBPM) (GSBPM, 2013); however, this model is not sufficient to describe the specific phases of the process for the management of the Integrated Register System. To take into account the peculiarity of this process, a profiling of the GSBPM model called Generic Statistical Integrated Registers Model (GSIRM) has been proposed. According to this model, the process for creating and managing the ISSR has some specific **new phases** to take into account, for instance:

* *Populate*: phase in which the register is populated or updated with collected sources (survey and administrative data).

To the already existing phases of GSBPM, some **new** **sub-processes** are added to take into account the specific need of ISSR. For instance:

* The *stock/flow harmonization* sub-process is added to the *Process* phase

In Figure 2, new phases and subprocesses are shown as starred.



Figure : Generic Statistical Integrated Registers Model (GSIRM)

Let us remark that the relationships between GSBPM and GSIRM are of two types, namely complementarity and mutual dependency. Some examples of dependencies between the two models are shown in Figure 3, where Figure 3.a shows how GSIRM can depend from GSBPM, while Figure 3.b shows an example of the opposite dependency. In particular, in Figure 3.a the sub-process “Access Collected Sources” of the “Populate” phase of GSIRM depends form GSBPM phases related to the actual collection of data that are indeed only accessed in the population phase of registers. In Figure 3.b, the sub-process “Create Frame and Sample” of the “Collect” phase of GSBPM depends form GSIRM phases related to the build of the registers that can be used as a basis for frames and samples creation.



Figure 3: Dependencies between GSIRM and GSBPM

**3. Quality of the ISSR**

The ISSR has been designed with quality of data as a priority. Specifically, the system design is based on the following basic principles:

* **Usage of International Standards both in the process architecture and in the data architecture.** Designing the process architecture we referred to different statistical international standards such as the Generic Statistical Information Model (GSIM) (GSIM, 2013) to describe variables, unit types, populations etc. and the GSBPM to describe the specific process of registers production. We highlight that we needed to complement GSBPM, in order to take into account the tasks performed for ISSR implementation. With respect to the data architecture, we invested on adopting standards from both the technological perspective, e.g. W3C standards, and the design perspective, e.g. Eurostat’s EARF (EARF, 2015).
* **Usage of the OBDM approach.** Ontologies permit to represent metadata “coupled” with data, so they are not only limited to a “documentation” role but they do permit to govern the data integration step by ensuring the high quality of integrated data. Moreover, being the ontology a formal description of concepts, relationships and constraints typical of the domain of interest, it is possible to check that data stored in the system are conform to it. For example, if the *family unit* concept is modeled in the ontology as always composed of either a couple or at least one parent and an unmarried child, data in the system of families with a nucleus of less than two people are revealed as a violation of the ontology’s constraints.
* **Generalized production process for all registers within ISSR.** A single generalized process is defined for the registers and their integration in the ISSR.
* **Evaluation of quality indicators during all phases of the processes involving the ISSR.** The architecture of the processes and their internal and external interfaces allows for a systematic analysis, monitoring and intervention with improvement actions. Therefore the definition of monitoring indicators are not only oriented to measure the final product, but also to trace the various stages of the process;
* **Timeliness in the ISSR.** The timeliness of data in a register of the ISSR is a time period, defined as the difference between
	+ Time in which an event occurs
	+ Time in which the event is consolidated in the register

In this definition, timeliness is related to the consolidated version of the register. However, it could be referred also to initial and current versions of the register.

* **Accuracy of the ISSR.** The ISSR is created by massive integration of administrative archives and survey data. At microdata level different statistical techniques are adopted e.g.: Record linkage, Statistical Matching, Projection macrodata estimators, Model predictions for single units. At  macrodata level other techniques are used of e.g.: Calibrated Estimates, Small Area Estimation, Bayesian Models, etc.. As result, we have an increase of the available information with respect to that provided by each source taken individually. However, it is relevant to state that the register values are the output of statistical processes subject to uncertainty with respect to both units and variables. Thus, an aspect to be dealt is that of the accuracy of the statistics obtained as aggregation of the register values and the strategies and techniques  useful to inform on the accuracy the different stakeholders of the ISSR.

**4. Concluding Remarks**

In this paper, we described our design efforts to build the Integrated System of Statistical Registers. We highlighted the data architecture and related quality principles by detailing the important role played by the ontology based data management and the data virtualization approaches. In addition, we described the process architecture and the main tasks involved in the realization of the ISSR with respect to GSBPM. We also detailed the result of complementing GSBPM with a specific set of tasks for the ISSR, namely what we called the Generic Statistical Integrated Registers Model (GSIRM). Finally, we highlighted the main quality aspects taken into account in the ISSR.

Our steps in the next future are mainly related to continue the development of the ISSR according to the design and the quality characteristics described in the paper.

**4. References**

Lenzerini M. (2011), Ontology-based data management. In Proc. of the 20th Int. Conf. on Information and Knowledge Management (CIKM 2011), pp 5–6.

GSBPM – Generic Statistical Business Process (2013), Available at: https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0

GSIM C Generic Statistical Information Model (2013), Available at:

[https://statswiki.unece.org/display/gsim/GSIM+Specification](https://statswiki.unece.org/display/gsim/GSIM%2BSpecification)

EARF Enterprise Architecture Reference Framework (2015), Available at: https://ec.europa.eu/eurostat/cros/content/ess-ea-rf\_en