**Standardization of Business Statistics Processes in Istat**

Mauro Bruno, Istat, [mbruno@istat.it](mbruno%40istat.it)

Orietta Luzi, Istat, luzi@istat.it

Giuseppina Ruocco, Istat, giruocco@istat.it

Monica Scannapieco, Istat, [scannapi@istat.it](scannapi%40istat.it)

**Abstract**

*Since the second half of 2014, Istat has launched a modernization program to improve the quality of statistics produced. In order to increase the effectiveness and efficiency of the production chain, the revision and the standardization of the business statistics processes have become of utmost importance. Currently, business surveys use highly customized methods and tools. Some processes are tied to specific persons, and built on their knowledge, presence and skills. This organization produces two major negative effects: duplicated work and limited reuse of tools and competencies.*

*The Generalized Process for Business Statistics (GPBS) project aims to identify and implement a general data model and architecture, to standardize similar steps of business surveys. According to the Generic Statistical Business Process Model (GSBPM) framework, the GPBS initiative has the main objective to standardise:*

* *metadata in order to harmonize concepts and data structures at domain level;*
* *methods and tools for the statistical phases following the data collection stage;*
* *the workflow, meant as the most appropriate combination of data and methods to be used in the statistical process.*

*The core of GPBS will be a collection of shared and generic corporate services for processing, storing and analysing statistical information. Such services will be designed and implemented enhancing the software currently available in the Istat’s generalized software repository. Adopting the conceptual model of the Statistical Production Reference Architecture (SPRA), the main components of the TO-BE integrated architecture, will enable statistical activities related to GSBPM phases and steps.*

**Keywords:** Process standardization, workflow, statistical service.

**1. Introduction**

Following the modernization programme principles, from the second half of 2014 Istat has launched a significant revision of business statistics. The main objective of the Generalized Process for Business Statistics (GPBS) project is to standardize the statistical production chain and to support the process documentation, thus increasing the effectiveness and efficiency of the overall activity. Starting from collected data, the GPBS aims to integrate in a single environment the different steps of data processing, overcoming the stovepipe organizational model (as shown in Figure 1).

**Figure 1. AS-IS versus TO-BE scenario**

****

The Generalized Process for Business Statistics project has the main objective to standardise:

* **metadata,** in order to harmonize concepts and data structures at domain level;
* **methods and tools** for data editing and imputation;
* **the workflow**, to manage the different steps of data processing and to connect the services used in the statistical production chain.

The GPBS will use a collection of shared and generic corporate services for processing, storing and analyzing statistical information. Such services will be designed and implemented enhancing the software currently available in the Istat’s generalized software repository[[1]](#footnote-1). In terms of Generic Statistical Business Process Model (GSBPM) [1], the initial focus of the project is on the ‘Process’ phase.

In this paper, we will focus on how the design of the GPBS system addresses some of the most relevant process quality issues for Official Statistics [2], namely:

* **process traceability**, i.e. the ability to know the detail of each single step that has been performed for data processing;
* **process reproducibility**, i.e. the ability to reproduce a process instance several times on the basis of user requirements;
* **process standardization**, i.e. the characteristic of the process to be defined in terms of standard services implementing shared statistical methods;
* **process efficiency**, i.e. the ability to reduce process instances execution times.

**2. GPBS architecture**

In order to describe the main components of the GPBS target architecture, we need to introduce the following concepts:

* **data service**: a software providing generalized functionalities to load/retrieve data to/from a data repository. The main data repositories are: the relational databases containing the raw data collected for each survey and the Statistical Registers, which instead consist of validated data;
* **statistical service**: a software providing one or more statistical business functions (extract sample, calculate weights, perform error checking, etc.) that can be invoked as a service;
* **process step**: to describe a statistical process, it is useful to subdivide the process in a limited number of steps. Each step is tied to one or more statistical services, according to their level of granularity. The description of a process must also include a specification of the sequence and the routing of the different process steps;
* **process orchestrator**: to manage the execution of a statistical process, modelled according to a defined workflow. The use of a process orchestrator based on shared services will ensure an easy replication of statistical production across domains and therefore minimize the cost of adjusting or expanding statistical production;
* **metadata component**: the inputs and outputs of the different components of the architecture (Data Services, Statistical Services) are described in terms of standardized metadata. According to the Generic Statistical Information Model GSIM [3], structural metadata will describe the meaning of the data used by the statistical services in the workflow and all parameters, rules and auxiliary data sets, necessary for the involved process steps.

**Figure 2- SPRA architecture components**



Adopting the conceptual model of the SPRA Architecture[[2]](#footnote-2), Figure 2 illustrates the main components of the TO-BE integrated architecture, where a statistical production process corresponds to a subset of GSBPM phases and steps. A step of the ‘Process’ phase, for instance ‘Edit&Impute’, will use one or more functionalities implemented in a statistical service. Services are invoked providing both input datasets and metadata managed by the process orchestrator.

The output of the service is stored in a data repository by a data service and/or used as input for the next service in the process chain. Services may provide a graphical user interface for manual inspection and manipulation.

**3. Quality issues in GPBS**

As mentioned in the introduction, relevant process quality dimensions are: process traceability, process reproducibility, process standardization and process efficiency. In the following, we will show how our design will address each of these dimensions.

*3.1. Process traceability, reproducibility and efficiency*

The main component of GPBS related to process traceability, process reproducibility and process efficiency is the Business Process Management System (BPMS) [6], that will manage the workflow. Figure 3 shows the main elements of the BPMS:

* **Process model repository**: the repository will contain a process model for each survey edition deployed in GPBS. The process model will specify the services to be invoked as well as the parameters to be configured for execution. Each model can result in several process instances to which configuration parameters are to be associated.
* **Execution engine**: the engine executes process instances on given parameters and manages the invocation of services as well designed human interactions.
* **Execution logs**: permit to concretely implement process traceability.

**Figure 3- Main components of the BPMS**



*3.2. Process standardization*

Process standardization is bound to: (i) process model (ii) service definition and (iii) methods implementing services.

Concerning the process model, GPBS is based on GSBPM. In particular it addresses all the phases of GSBPM, though focusing on the “Process” phase as an initial step.

Service definition implies the exact definition of services input, output and related parameters, described in terms of standardized metadata defined by domain experts. In GPBS metadata, data structures, rules and, more in general, all the parameters needed to run the statistical services are defined according to a shared model. As an example, Figure 4 shows a simple case of variable name standardization referred to the turnover.

**Figure 4- Variable name standardization (turnover)**



With respect to methods standardization, let us consider the example of the ‘Edit&Impute’ phase (E&I). For this phase, the methodological standard ‘Generic Statistical Data Editing Model’ (GSDEM) [7] has been considered as main reference [2]. Actually, an E&I process is viewed in GSDEM as composed of different steps (and their associated statistical functions). An E&I work flow is defined as a configuration where the generic E&I steps and functions are placed in tandem, parallel or iteration.

The configuration is specified in terms of the mapping from the input to the output of each E&I function, and the associated metadata including the relevant concepts, data structure, routing conditions, stopping rules, etc. In GSDEM a service-oriented perspective is adopted. Indeed, E&I functions are viewed as statistical services performed at a given step of the data processing.

The GSDEM E&I steps currently under implementation are: (i) deterministic editing; (ii) selective editing; (iii) interactive editing.

Deterministic editing aims at removing from data those errors which are easily detectable and treatable (obvious errors), and systematic errors. In the selective editing phase, influential errors are detected, based on the idea of looking for important errors in order to focus the most accurate treatment on the corresponding subset of units. In interactive editing, selected micro-data (e.g. the ones corresponding to influential errors) are manually checked and, if necessary, adjusted using expert judgment and/or re-interview.

**4. GPBS USE CASE**

This section describes the invocation of the statistical service for detection of units with influential suspicious values. This task relates to GSBPM sub-process [5.3] ‘Review & validate’.

**Current situation**

Within our current short-term statistical surveys, thematic experts run at local level procedures to detect outlier units to review. Let us remark that:

* Each outlier detection procedure is specific for each survey, i.e. no shared method is adopted;
* Each dataset has structural metadata that are not harmonized with the ones of the other surveys;
* Several steps are manual, even if automatable, hence not easily traceable;
* Each step is sequential, even if some tasks that compose the step are indeed parallelizable.

**GPBS enhancements**

In order to overcome the above process quality issues, in GPBS we performed the following design choices:

* A general outlier detection procedure is adopted for each survey, namely selective editing based on a contamination model[[3]](#footnote-3);
* Global structure metadata are defined for the while set of surveys (see e.g. turnover example in Figure 5);
* The outlier detection procedure is implemented by a software statistical service;
* The statistical service is invoked within a defined workflow that will manage (i) automatic executions, (ii) human interactions, (iii) parallel service executions and (iv) process logging for the purpose of traceability and reproducibility.

**Implementation**

The statistical service implementing the outlier detection procedure within GPBS results from the wrapping of the R-package Selemix.

**Architecture components**

* Metadata: survey structural metadata retrieved from the metadata repository;
* Statistical Service: influential values detection;
* Data Service: read/write data from/to the Data Repository.

**Figure 5 - Invocation of a statistical service in GPBS**



The invocation of a statistical service is shown in Figure 5. On the left, a user invokes the statistical service to identify units with influential errors. Such service, accessing metadata and input data (invoking a data service) runs the statistical procedure. After the processing, output data and metadata are stored in the repository and reports are displayed to the user.

**5. Conclusions**

GPBS is a system that enables the execution of standard processes defined in a service oriented way. Both data and statistical services are designed with a specific focus on harmonizing metadata and standardizing services and related methods. The system will permit to automate the execution of survey editions and to fully trace such execution. In addition, it will be highly configurable, allowing human interactions by means of a controlled workflow. Hence, GPBS is a system that improves the quality of statistical surveys in terms of the aforementioned quality process dimensions.

The current ongoing activities relate to the development of the system for an initial testing phase, on a subset of short term business statistics planned by the end of 2018.

**6. References**

[1] Generic Statistical Business Process Model (GSBPM) [https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0](https://statswiki.unece.org/display/GSBPM/GSBPM%2Bv5.0) (Accessed: 18 May 2018).

[2] Quality Assurance Framework of the European Statistical System <http://ec.europa.eu/eurostat/web/quality> (Accessed: 18 May 2018).

[3] Generic Statistical Information Model (GSIM) https://statswiki.unece.org/display/gsim/Generic+Statistical+Information+Model (Accessed: 18 May 2018).

[4] Statistical Production Reference Architecture (SPRA) https://ec.europa.eu/eurostat/cros/content/spra\_en (Accessed: 18 May 2018).

[5] Enterprise Architecture Reference Framework (EARF) <https://ec.europa.eu/eurostat/cros/content/ess-enterprise-architecture-reference-framework_en> (Accessed: 18 May 2018).

[6] Dumas M., La Rosa M., Mendling J., Reijers H. A.(2013). Fundamentals of Business Process Management, Springer

[7] Generic Statistical Data Editing Models (GSDEMs) <https://statswiki.unece.org/display/kbase/GSDEMs> (Accessed: 18 May 2018).

1. https://www4.istat.it/en/tools/methods-and-it-tools [↑](#footnote-ref-1)
2. The international standard Statistical Production Reference Architecture (SPRA) [4] is part of a more general framework on the Enterprise Architecture at the level of the European Statistical System [5]. SPRA models the statistical services to be implemented for GSBPM processes. [↑](#footnote-ref-2)
3. For more information about selective editing, see:

 <https://www.degruyter.com/view/j/jos.2013.29.issue-4/jos-2013-0039/jos-2013-0039.xml>. [↑](#footnote-ref-3)