CONFERENCE OF EUROPEAN STATISTICIANS

ModernStats World Workshop 2022

27-29 June 2022, Belgrade, Serbia 13 June 2022

Modelling a Methods and Tools Catalogue compliant with Official Statistics standards

Francescangeli Paolo, Bruno Mauro, Amato Francesco, Ruocco Giuseppina (Istat, Italy)

pafrance@istat.it

Abstract

In order to foster process standardization, a general model for process design was conceived, based on the main concepts of Generic Statistical Information Model (GSIM) and Generic Statistical Business Process Model (GSBPM). More in detail, a conceptual architecture was built to connect the process design outlined in the Business Group of GSIM standard to the main phases and sub-processes of the GSBPM model. This architecture is also compliant with CSPA principles, allowing the development of metadata driven Statistical Services.

In the following article, we will describe Istat's experience in the development of a Service and Methodological Tools Catalogue (MTC), compliant with Official Statistics Standards.

1 Introduction

The design of a conceptual architecture for Process Design and Management, compliant with official statistical standards, allows achieving several goals, such as:

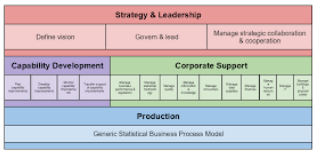
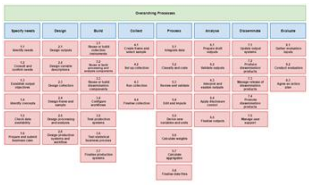
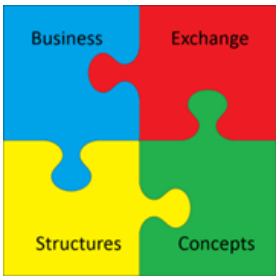
- Reuse of available Methodological Tools and software solutions
- Standardization and reproducibility of statistical process
- Harmonization of statistical output, both within and between National Statistical Institutes (NSIs)
- Increase of interoperability, in terms of data and metadata management and sharing.

The following paragraphs provide:

- A brief summary of official Statistics standards
- The Core Ontology as a bridge for Standard Models Interoperability
- A focus on the Business process, the Business Service and the methodological tool in the context of Standard Models
- An overview of Methods and Tools Catalogue, implemented according to official Statistics standards.

2 *Statistical Standard Models Overview*

Official Statistics standards have been conceived as high-level models, describing several aspects of statistical organizations. More in detail:

	**The Generic Activity Model for Statistical Organizations (GAMSO)** *describes and defines the activities that take place within a statistical organization. It extends and complements the Generic Statistical Business Process Model (GSBPM), including additional activities needed to support statistical production.*[1]
	**The Generic Statistical Business Process Model (GSBPM)** *Is a means to describe statistics production in a general and process-oriented way.*[2]
	**The Generic Statistical Information Model (GSIM)** *Is a reference framework of internationally agreed definitions, attributes and relationships that describe the pieces of information that are used in the production of official statistics. This framework enables generic descriptions of the definition, management and use of data and metadata throughout the statistical production process.*[3]
	**Common Statistical Production Architecture (CSPA)** *Is a reference architecture for the statistical industry.CSPA aims at documenting statistical services in a standard way, in order to ease their exchange and reuse between statistical institutes.*[4]

[1]GAMSO: https://statswiki.unece.org/display/GAMSO
[2] GSBPM: https://statswiki.unece.org/display/GSBPM
[3]GSIM: https://statswiki.unece.org/display/GSIMclick/Clickable+GSIM
[4] CSPA: https://statswiki.unece.org/display/CSPA/Common+Statistical+Production+Architecture

3 Core Ontology as a bridge for Standard Models Interoperability

The Core Ontology supporting group, in an effort to build a standardized framework of model integration and interoperability, aims at tying official statistical standards together by an ontological description. There have been issues regarding the ambiguity in the definitions of several similar concepts between different standards which have been addressed. The main difficulty is to find a correlation between similar objects in several models that appear to map to a single concept, but whose definition depends on the context of its own model description.

The first attempt was to organize several related concepts in a relation chain that could detail and describe the model from an abstract point of view and then add detail to it. Agreement at the most general level of abstraction should be assigned to the concepts from GAMSO model, that define an organization strategy, field of influence, areas of operation and so on.

GAMSO describes Activities as general high-level tasks, while GSBPM Business Process description fits very well in the Statistical Activity definition, restricted to the activities that produce statistical data.

GSBPM Process is organized in Phases and Sub processes, consisting in a list of admissible task or high-level objectives that the GSIM concept Business Function can perform.

GSIM also possesses high-level concepts like statistical program that has relation with both GAMSO Activity and GSBPM Business Process.

Therefore, a chain of relations was modeled starting from the generic concept of GAMSO Activity to the GSBPM Business Process and then connecting the related GSIM concept of Function, Process and Step, as shown in the figure below. Other concepts from the associated statistical models can be linked on top of this conceptualization.

CSPA related concepts can be linked through the GSIM Business Service. A Business Service can also act as an interface for objects which are not actually services but can perform useful tasks for the process itself. From this perspective, a business service may serve as the functional interface for Methodological Tools, encompassing a wide range of tools which are not actually described by any standard. Thus, this work can also be useful to attempt a tool description and standardization too.

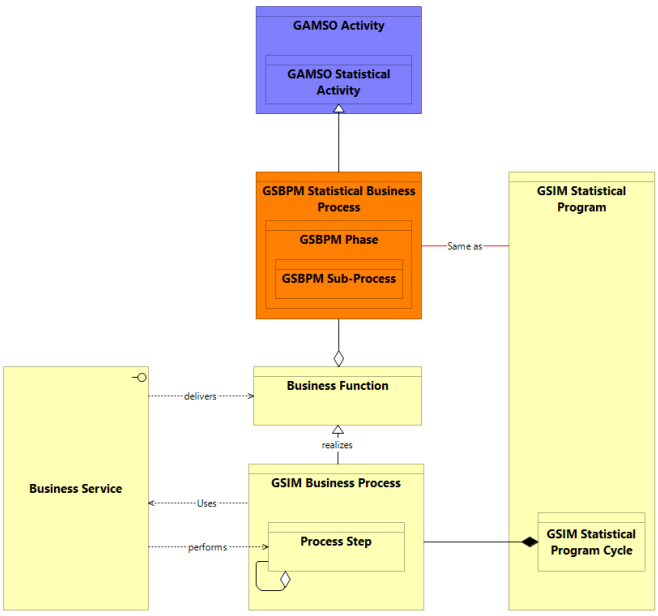


Figure 1: Relation between Official Statistics Standards and Core ontology

4 *The Business Process in the context of Standard Models*

Starting from the central concept of "Business Process", according to GSBPM, Business Process is conceived as a "means to describe statistics production in a general and process-oriented way". This is a way to classify and simplify the complexity of the actual process, by cutting it into several well-established process phases of common tasks. All the complexity, feedback loop and conditional structures are hidden at this level.

GSBPM considers the "process" as a whole and cuts it into pieces that classify objective or desiderata completely. Abstracting from GSBPM, we can relate it to Statistical Activity in GAMSO model, or Statistical Program at GSIM level. GSIM Statistical Program matches with GSBPM Business Process, except for process support concepts which are not directly related to the statistical production. According to the definition of Statistical Program from clickable GSIM: "A set of activities, which may be repeated, that describes the purpose and context of a set of Business Process within the context of the relevant Statistical Program Cycles".

Finally, we can model the "Activity" concept from the three standards and link them all together, by the mean of "same-as" equivalence:

1. GAMSO Statistical Activity
2. GSBPM Business Process
3. GSIM Statistical Program.

There is a clear difference in significance and scope between similar terms in GSBPM and GSIM, but "Business Process" is a central concept for all of the models to tie. Process is a very wide term and as a matter of fact, it relates to every aspect of the production activity, from the very general to the little detail. So, having GSIM a variety of concepts describing the "process" at any specification level, we can relate them with the "GSBPM" process and its derivative phases and sub-process. More in detail, we could tie the most general concepts of GSIM Statistical Program to GSBPM Business Process and leave the specification of phases and sub-processes to a GSIM business function which, in turn is realized by a set of GSIM business processes.

The key GSIM concepts are listed here:

- Business function, identifying the objective that needs to be performed
- Business Process, specifying the component performing the objective
- Process Step, performing an operation, a command, or an instruction
- Process Step Instance, specifying the instruction being actually executed
- Business Service, identifying a set of tools able to realize business objectives, not limited to software or automated tools only
- Process Design, specifying the application of a process step to a set of information objects within a process method
- Process Pattern lists the related information objects
- Process Control: "A set of decision points which determine the flow between the *Process Steps*"
- Process Method: "A *Process Method* describes a particular method for performing a *Process Step*"

Excerpts of GSIM definitions are shown only for the purpose of being used in the following analyses, to focus on what really matters, for efficient implementation. Whenever we need to perform some kind of tasks to realize an objective, we can implement some tools compliant with statistical standards. A set of tools providing some

functionalities composes a Business Service. Moreover, these functionalities can be expressed as commands provided with I/O specifications, to realize an objective (Business Function).

The purpose of the described framework is to build an application executing a set of steps contained in a process, through a service, with a certain method. Everything here is abstracted from the details regarding the implementation of any software or programming language. The focus is on the task, and on what the system should do to perform it.

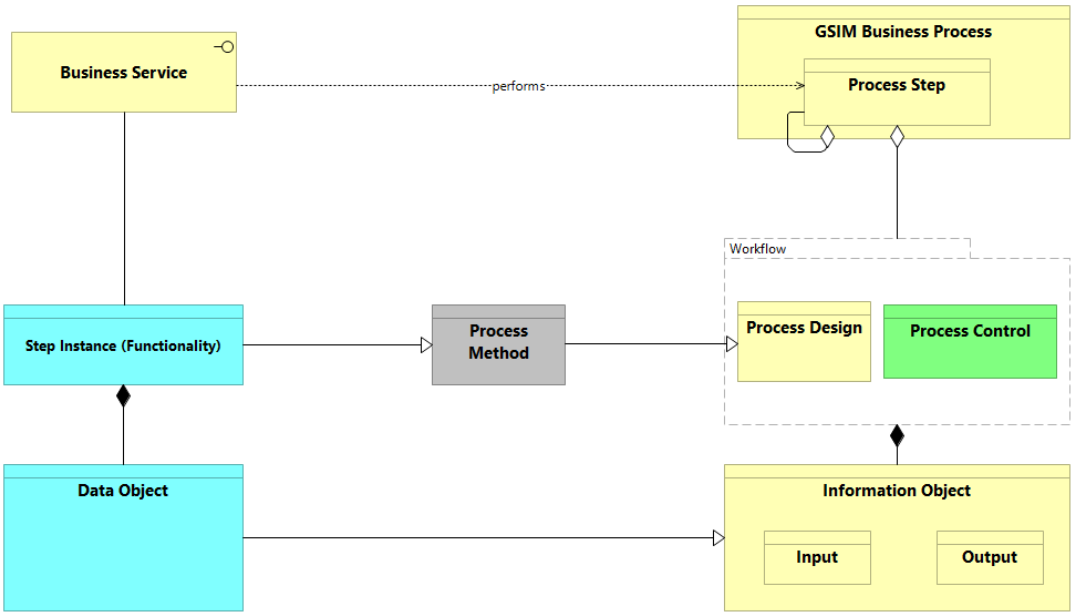


Figure 2: Business Process and services with added detail from GSIM

5 *Business Service implementation*

In an effort to standardise and define the detailed structure of the process, it is possible to separate the information about **what** a process does, its objectives and its conceptual details, from the specification about the functional implementation and mode of use (**how**).

At the Business level (yellow in the figure below), we are interestad on what a process does and its effect on the desired outcomes more than how the process is actually performed. At the Application level (cyan in the figure below), instead, the focus in on actual step activation and execution.

We can use several other constructs from GSIM to describe the step activation process, decoupling into two separate mirror layers.

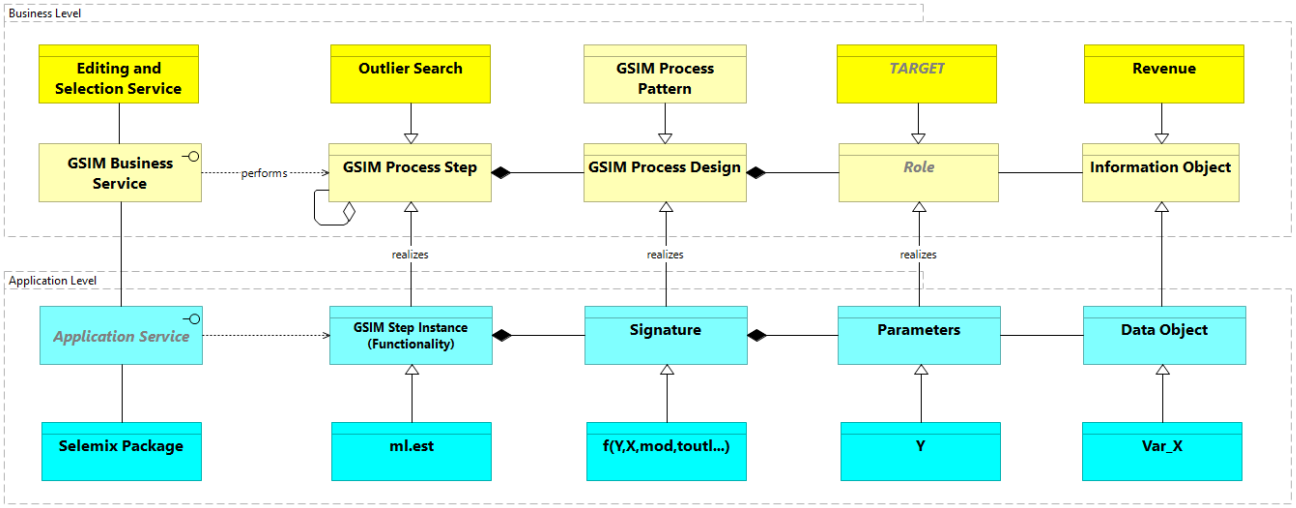


Figure 3: Process Design outline with example

In order to implement a *Business Service*, we need to split the business from the application layer, to provide information about the working environment, thus creating the concept of *Application Service*.

Step Instance object describes the details of the command issue.

Process Design lists I/O interaction, but, in order to better describe the relation between information object and process steps, the concept “Role” was created. It explains what purpose the information object has in the Process Design (i.e. a threshold, a dependent variable, a working parameter, the target of analysis, weights etc.). Role is dual to the static parameter in the signature of a function.

Decoupling the conceptual and application levels gives also degrees of freedom in changing implementation functions for the same conceptual tasks, although the two implementations must match the same design pattern to some degree. In practical terms, functionalities that perform the same tasks should share the same information interface. On the other hand, it means that we standardize the I/O interface once, and by matching the given pattern, we can use several implementations of a conceptual task, regardless of the actual working environment. As an example, if we had two different functions for outlier detection, we could make available them by wrapping them seamlessly. Thus, the output may differ because of the chosen process method, but the system would manage both regardless. A Catalogue of Methods and Tools, based on this schema was built, complete with a proprietary data management system.

6 *Extending the model from Statistical Service to the Methodological Tool*

Conceptual decoupling makes easier to describe functionalities and processes with a level of abstraction, saving the implementation details for later. It helps standardize and organize the information from a functional point of view. We are interested in the available functionalities, and how we can use them in the context of process standardization. But functionalities themselves are not designed according to standard models and most of the times, they are available in a set of different tools and formats.

A new concept of "Methodological tool" needs to be introduced. Regardless of the structure and technical details, a tool is something that offers functionalities, and a methodological tool has functionalities tied to the statistical methodology. So, repeating the process of concept abstraction done with Business Services, a business layer focused on Tool functionality, and an application layer describing the structure and working environment were defined.

In this analysis, a Methodological tool takes the form of a service provider for the business process. So, in our schema, a Methodological tool is always associated with a Business service whose task is to make the tool functionalities available to the Business Process. That is to say that, alternatively, the Business Service, sketched in the following figure, is the business interface for the Methodological tool.

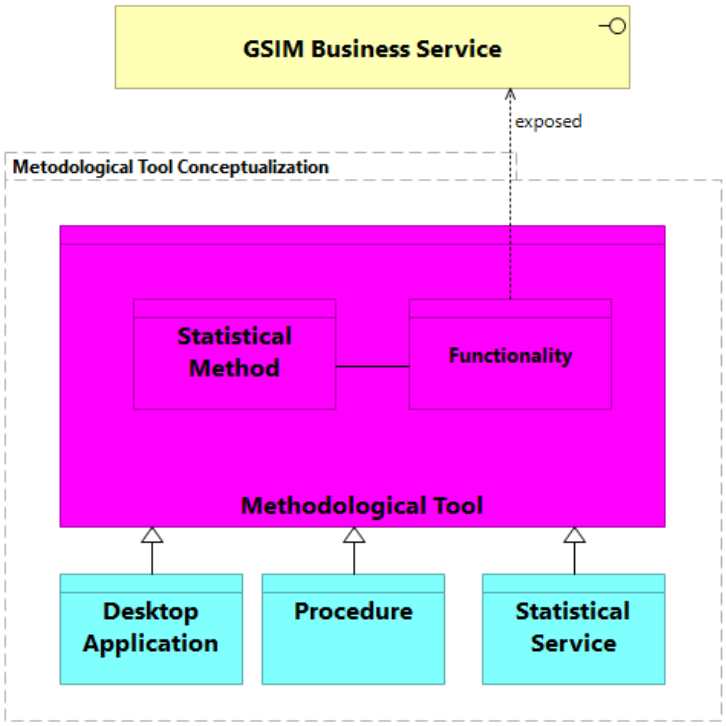


Figure 4: Methodological Tool conceptualization
(yellow: business layer – cyan: application layer)

The details about the tools specification are given to a set of objects hierarchically tied to the tool concept, containing the general information can be found here. Details on the working environment are postponed to the specialization tools. These can also be used to wrap actual code and software. The figure below shows the interaction between the process and the methodological tool.

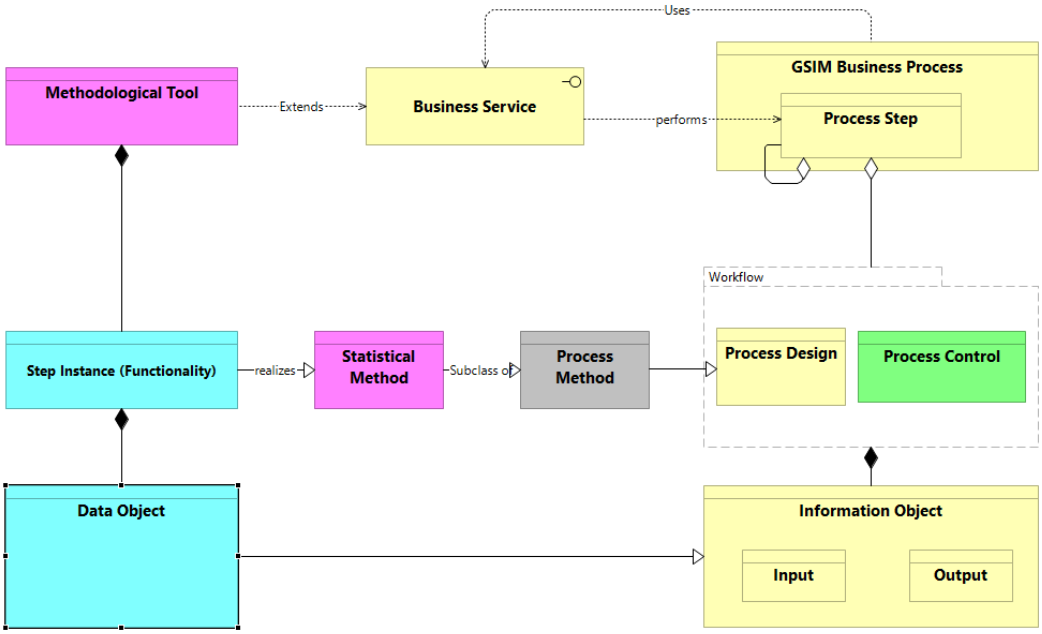


Figure 5: Business Process with Methodological Tool extensions
(yellow: business layer – cyan: application layer – Magenta: Methodological Tool)

The figure above describes how a Catalog of tools acts and interacts in the context of the standard models as a service provider. The available functions are many and diverse. This is a conceptual framework that can foster the development of many applications. A lot of care has been given not to avoid binding the system with too many constraints, but there are many paths connecting the concepts, so that they can be described and put in relation from more than one point of view and wrap up any information later.

This schema is useful for methodological studies, because it gives a standard model for process description, it can be used both by technicians, to develop software in standardized fashion, and by software reengineering. It can be connected with existing systems for statistical and survey standardization, like metadata-oriented data management.

7 Working Example

The following is an example of how to model a tool in the context of the Catalogue: R package Selemix performs outlier and influential errors detection.

The diagram explains how the Methodological Tool "Selemix" depicted in magenta in the lower part has functionalities that can be used to realize processes and their steps.

Functionalities and their processes are color coded with the same shade of blue, to highlight what process is realized by a functionality. Particularly, *ml.est* realizes a full estimate, prediction and outlier detection process, while *sel.edit* only realizes influential error search process.

Each process realizes one or more objectives (business function) which in turn are belonging to a GSBPM Area or Sub Process. "5.3 - Review and Validate" in the example above. GSBPM process context is set up this way.

The graph shows how each functionality maps to the process steps and how the steps are composed in processes that perform a Business Function. Notice how there could be processes performing one or more than one objective and functionalities that can realize one or more steps.

Process flow and I/O Information Objects are also highlighted in the graph: All of these objects can be described by the Methodological Tool conceptual framework.

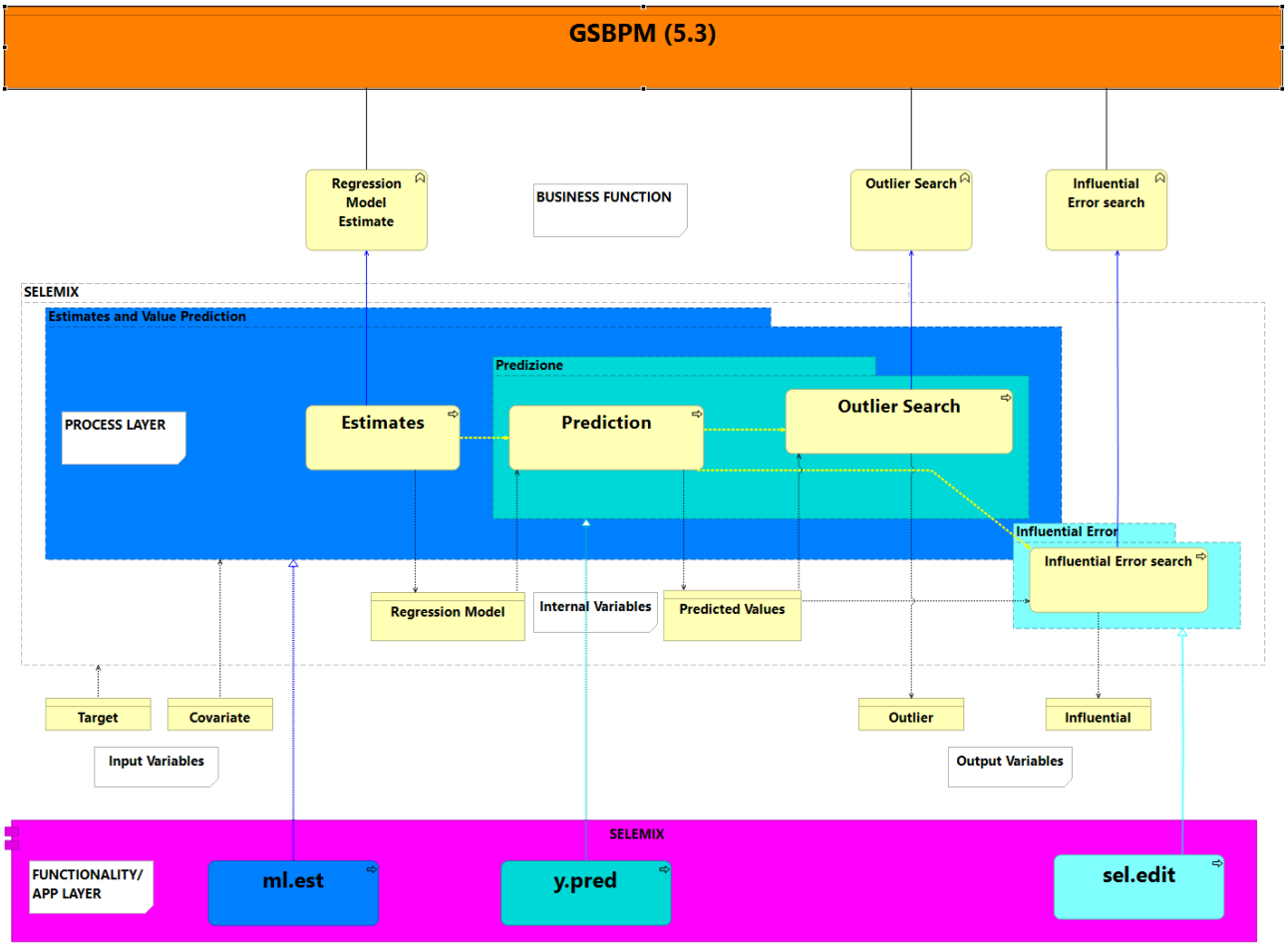


Figure 6: Process modelling example (Package Selemix)

8 *Common framework implementation*

A database containing metadata information for the Methods and Tools Catalog was developed based on the Statistical Service described above. The process management is very similar and the Methodological Tool was added as a service extension. Some objects are being maintained with redundancy for Statistical Service compatibility. It means that, if the right information is supplied, the system is able to actually run different services too.

Multi pathway for information intake has been maintained. It means that the data must be locally coherent to a degree, and that structural coherency must be ensured by a set of rules and consistency maintenance must be performed by software. This is necessary because of the wide range of information that cannot be collected completely and accurately in one fell swoop.

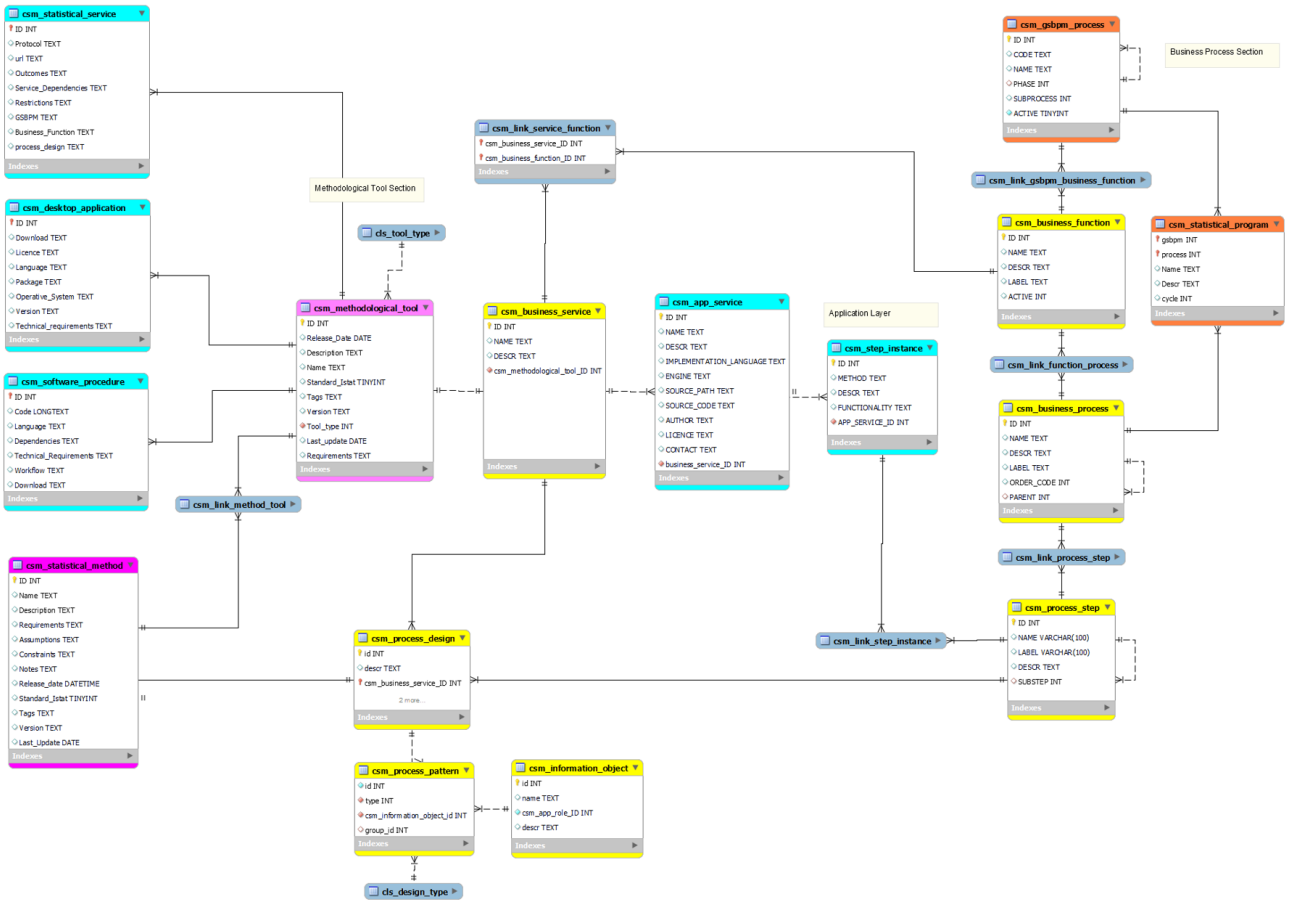


Figure 7: Data Base implementation

This schema shows the system integration at business level. Layers of functionality execution, workflow, data layer and user and session management are not depicted here. The system is complemented with the set of metadata tables and with GSBPM classifications used by the tools described in the Catalog. Process design and statistical methods implemented by the tools are also outlined here.

9 Conclusion

The MTC has the task of collecting information about the available tools and methodologies using the conceptual structure that was built from standard models integration. Its main task is to collect and standardize both the tool archive, thus supporting the process standardization using the common language of the statistical models. Based on the compliance with the statistical standards, the Methods and Tools Catalog can:

- Foster process restructuring, to avoid silos oriented design
- Standardize tool description and integration
- Help collecting information to transform tools into statistical services
- Be integrated with statistical services for direct service utilization.

The ongoing activities concern mainly data acquisition, upload and management. Future developments will focus on the improvement of the user interface, to meet the requirements of different stakeholders accessing the Catalog.

References

Generic Statistical Business Process Model (GSBPM), version 5.1, available from: https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1

Generic Statistical Information Model (GSIM), version 1.1, available from: https://statswiki.unece.org/display/gsim/Generic+Statistical+Information+Model

Common Statistical Production Architecture (CSPA), available from: https://statswiki.unece.org/display/CSPA/Common+Statistical+Production+Architecture

Bruno Mauro et al.: Using ModernStats standards to facilitate the sharing of statistical services, available from: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.58/2020/mtg2/MWW2020_Italy_Bruno_Paper.pdf