On the Design and Implementation of a Generalized Process for Business Statistics

M. Bruno, D. Infante, G. Ruocco, M. Scannapieco

1. INTRODUCTION

Since the second half of 2014, Istat has been involved in a modernization programme that includes a significant revision of the statistical production. The main objective of this project is to enrich the supply and quality of the information produced, while improving the effectiveness and efficiency of the overall activity.

Within this context, the standardization of the statistical production chain is of utmost importance.

The Generalized Process for Business Statistics (GPBS) has the main objective of designing and implementing a standardized system to support business statistics. The GPBS aims to integrate in a single environment the different steps of business surveys, leading to a more effective and efficient statistical processes. 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 generalized software repository.

2. BASIC CONCEPTS

2.1. The standard framework for modelling statistical processes

The reference standard for statistical business process modelling is the Generic Statistical Business Process Model (GSBPM). It describes and defines the set of business processes needed to produce official statistics. It provides a standard framework and harmonized terminology to help statistical organizations to share methods and components, within and between different surveys.

GSBPM identifies the possible steps in the statistical business process, and the inter-dependencies between them. Although the presentation of the GSBPM follows the logical sequence of steps in most statistical business processes, the elements of the model may occur in different orders, depending on specific circumstances.

GSBPM is designed to be a matrix, through which there are many possible paths. In this way, the GSBPM aims to be sufficiently generic to be widely applicable, and to encourage a standard view of the statistical business process, without becoming either too restrictive or too abstract and theoretical.

The GSBPM comprises three levels (displayed in Figure 1):

- **Level 0:** the overall statistical business process, including quality and metadata management;
- **Level 1:** the eight phases of the statistical business process;
- **Level 2:** the sub-processes within each phase.

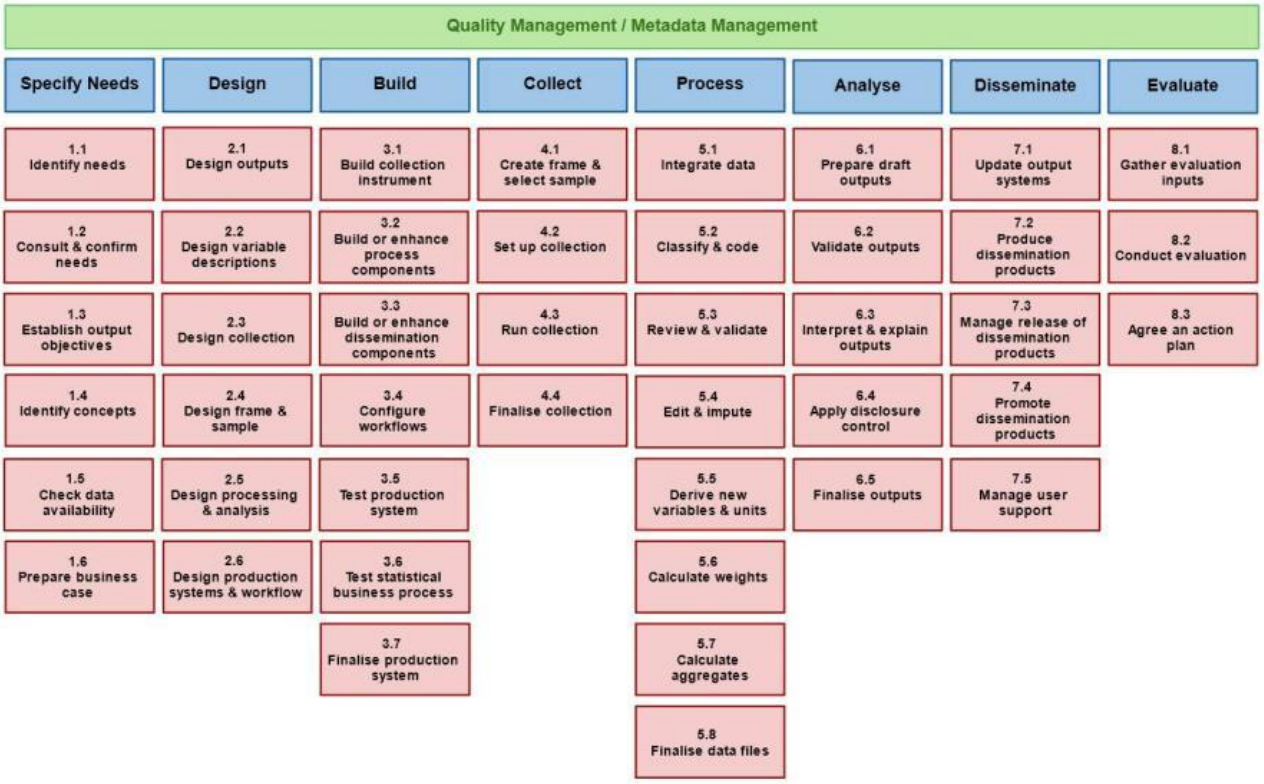

Figure 1 – GSBPM phases and sub-processes

2.2. Applying GSBPM to business surveys in Istat

According to the GSBPM framework, the Generalized Process for Business Statistics project has the main objective to standardise:

- **methods and tools** for some phases of the statistical production process;
- **the workflow** to manage the different phases of the production process (the workflow provides several functionalities to connect the services used in the statistical production chain).

To achieve these objectives, a deep review of the statistical production process is needed.

Currently our AS-IS situation is characterized by business lines with highly customized methods and tools. Some processes are tied to specific persons, and based on their knowledge, presence and skills. The strong connection between people and processes is a barrier towards sharing and standardization.

The major negative effects of this organization are, respectively: i) **duplicated work and inefficiencies** in several phases of the production process; ii) **no reuse of tools and competencies**.

To overcome the current scenario, we need to model statistical surveys based on processes more than vertical production lines.

Being GSBPM very broad, as a first step we have selected a subset of GSBPM sub-processes, focusing mainly on the Process and Analyze phases and their interdependencies (the list of sub-processes is shown in Figure 2).

Then, we have selected a subset of Business Statistics surveys (five short term statistics and three structural business statistics) to start the modelling phase. The selected list of the business surveys is shown in Table 1.

Business Statistics Surveys	***Periodicity***
Short Term Statistics	
Industrial turnover and orders	Monthly
Industrial Production	Monthly
Industrial import prices	Monthly
Industrial producer prices, external & total	Monthly
Industrial producer prices, domestic	Monthly
Structural Business Statistics	
Production of manufactured goods (Prodcom)	Annual
Information and communication technology (ICT)	Annual
Enterprises economic performances	Annual

Table 1 - List of GPBS pilot surveys

For each of the selected surveys, we have:

I. Interviewed the referring person, to obtain a clear AS-IS description of the current processes[1].
II. Modelled and designed the TO-BE integrated system built with generalized tools, supporting the different process steps.

[1] Attached, an example of AS-IS process modeled.

As a preliminary result, the AS-IS description has confirmed the need to harmonize GSBPM sub-processes (mainly 'Review & validation' and 'Edit & impute'), in order to reduce duplications, primarily between surveys referring to the same domains, regulation and statistical framework.

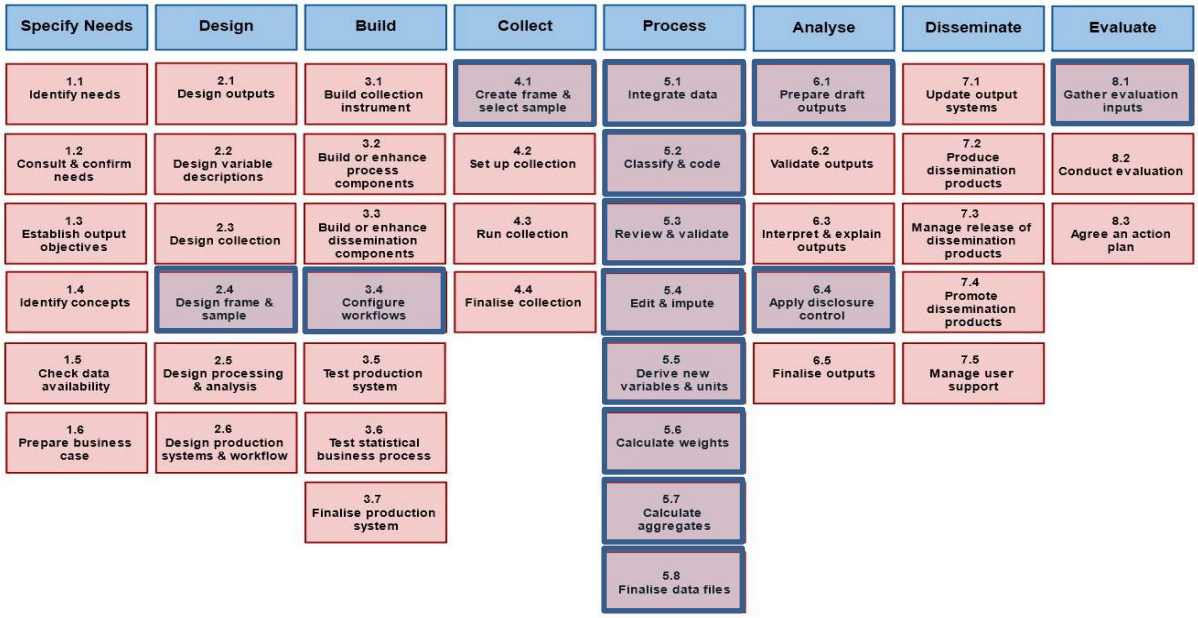

Figure 2 – Scope of GPBS project in terms of GSBPM sub-processes

3. GENERALIZED PROCESS FOR BUSINESS STATISTICS ARCHITECTURE

In order to describe the main characteristics of the GPBS, we should 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.
- **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.
- **Orchestrator**: the Process Orchestrator manages the execution of the Statistical Services performing a statistical process. The use of a Process Orchestrator based on a standard notation such as BPMN and using shared Services will ensure an easy replication of statistical production across domains and therefore minimize the cost of adjusting or expanding statistical production.
- **Metadata Management:** the inputs and outputs of the different components of the architecture (Data Services, Statistical Services) are described in terms of standardized

metadata. Structural metadata describe the meaning of the data used including the definition of a data element and a data set, variable names, variable types, unit identifier, classification identifier, etc. Referential metadata are the information objects necessary to run the process. Such metadata contain the process flow and all parameters, rules and auxiliary data sets, necessary for the involved process steps.

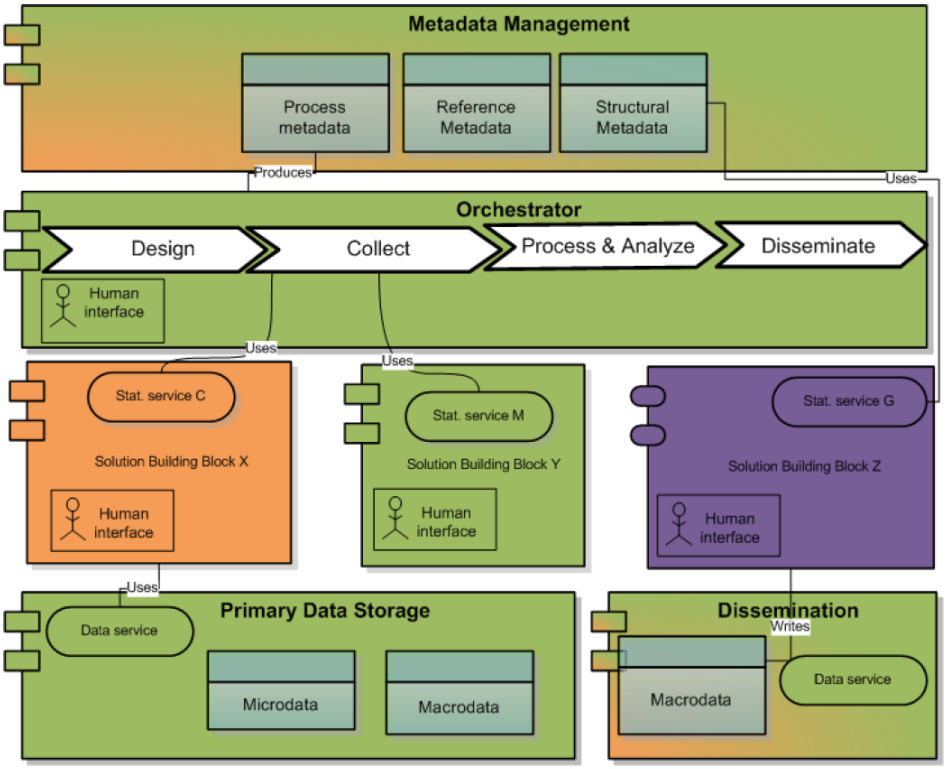

Figure 3 - GPBS (high level) Architecture

Adopting the conceptual model of the SPRA Architecture[2], Figure 3 illustrates the main components of the TO-BE integrated architecture, where a statistical production process corresponds to a number of GSBPM phases and steps. Each phase, for instance 'Design', will use one or more functionalities implemented in a Statistical Service. Services are invoked providing both input datasets and corresponding structural and referential metadata managed by the Process Orchestrator.

The output of the Service will then be 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.

The assessment of GPBS performances is under definition. Preliminary considerations are related to the identification of indicators to assess (i) the efficiency and (ii) the quality

[2] The international standard ESS Statistical Production Reference Architecture (SPRA) is part of a more general framework on the Enterprise Architecture at the level of the European Statistical System. SPRA models the statistical services to be implemented for GSBPM processes.

improvement before and after the adoption of GPBS tools. The main expected outcome is the decrease of manual revisions, overlapping steps and overediting, resulting from the standardization of data processing. Also, the reduction of the time lag to produce the final estimates should encourage the survey managers to use GPBS tools, instead of their own procedures. In general, the target improvements to measure will be related to the six dimensions of the ESS Quality Assurance Framework and particularly to accuracy and reliability, coherence and comparability. Nevertheless, in case this improvements should be difficult to assess by explicit indicators, we will consider collecting survey staff feedbacks and opinions concerning specific topics.

4. GPBS USE CASES

In order to describe the challenges to be faced while implementing the GPBS architecture, a list of use cases is reported in the following section.

4.1. **Service Invocation Modes:** Detection of units with influential suspicious values

Overview

This task relates to GSBPM sub-process [5.3] 'Review & validate'. In this phase, data editing activities allow to detect and correct data inconsistencies and errors. The aim is to improve the quality of disseminated statistical output.

In order to increase efficiency and timeliness, the editing strategy adopted is based on the selective editing[3]. This approach aims at identifying relevant errors, thus limiting a very accurate revision only to a subset of units.

Architecture components

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

[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.

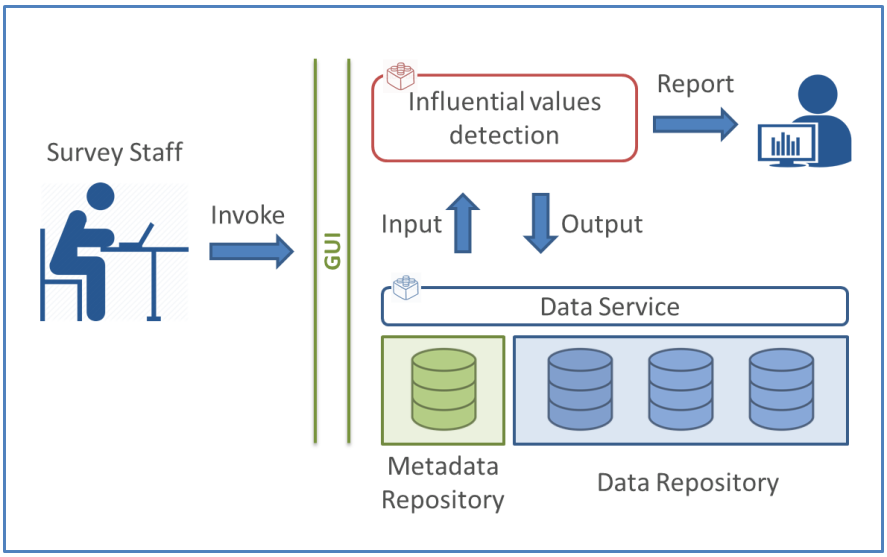

Figure 4 - Invocation of a Statistical Service for influential values detection

The invocation of a Statistical Service is shown in Figure 4. 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.

Technical remarks

Statistical Services can be invoked either in a synchronous or asynchronous mode. Each approach has pros & cons reported in the following scheme.

Invocation pattern	Pros	Cons
Synchronous	✓ Easy to implement (REST apis) ✓ Does not need a complex server architecture	− Not suitable for huge amount of data − Difficult to manage multiuser access to the statistical service
Asynchronous	✓ Suitable for huge amount of data ✓ Suitable to manage multi user access to data	− Need an environment to schedule & submit jobs − Need to implement a communication protocol with the scheduler

Table 2 – Pros & Cons of service invocation patterns

4.2. **Orchestration:** manual revision of influential suspicious values

Overview

Assuming the output of the influential values detection (as described in the previous section) is a subset of units to be manually reviewed, the survey staff may:

1. Use the available information (other sources or historical survey data) to check the coherence of survey data;

2. Contact the respondents to verify collected data.

Architecture components

1. Metadata: survey structural metadata retrieved from the metadata repository;

2. Process orchestrator: invoke in a sequential way Data Services and Statistical Services, according to a defined process flow;

3. Statistical Services: 1) influential values detection;
   2) interactive review of influential values;

4. Data Services: read/write data from/to the Data Repository.

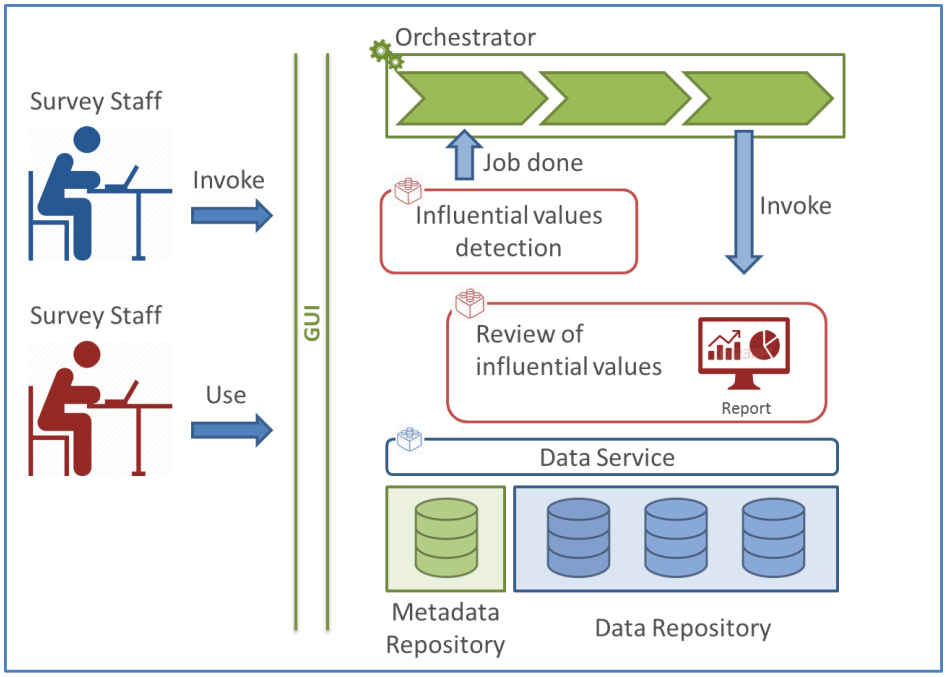

Figure 5 – Orchestration of Statistical and Data Services

Figure 5 shows two users involved in data editing activities. The upper user invokes the 'Influential values detection' Statistical Service. Once the execution of the task is completed, the Process Orchestrator (using the output of 'Influential values detection') runs the 'Review of influential values' Statistical Service. Such service allows the user to visualize and correct the influential values to be manually reviewed.

Technical remarks

The Process Orchestrator should check the execution of the invoked services, according to the sequence of the workflow and manage simultaneous users interactions. The Process Orchestrator should perform the following tasks:

- Execution of the control flow on a process instance;
- Scheduling of different process instances;
- Avoid overlapping of different users revisions (transactional support in coordination with the RDBMS);
- Provide a graphical user interface to (at least): (i) pass parameters to services; (ii) invoke services; (iii) monitor process execution; (iv) handle exceptions; (v) access (read/write) and visualize available data.

4.3. **Event management:** Business demographic events

Overview

Referring to GSBPM sub-process 'Create frame & select sample', a relevant issue is the management of the changes of both identification and stratification variables (eg, location of the business, legal person, etc).

Usually, in a sample survey, collected data and the Register information used as sampling primary data source, are not time aligned. Indeed, the list of units to interview at time *t* is extracted from the Register that contains data validated to a time *t-x* (where *x* is the reference time of different integrated data sources in the Register). During data collection, some business demographic events may be observed, and the updated information may be used to refresh the Register.

It would be useful to store and centrally manage all changes transmitted by the respondents of different surveys, in order to reduce the time gap between the Register and survey data.

While short-term surveys need immediate updates (data changes collected by one survey need to be shared with all Register users), structural statistics (released typically on annual basis and referred to a defined period) may need different update schedules, in order to limit the impact of changes on data coherence.

To face these issues, the adopted solution is to hold two versions of key variables, the 'current' version, updated as soon new information is received, and the 'frozen' version that remains stable for a certain period and is updated from the current version periodically, or in specific situations agreed with users. The challenge is to design the most appropriate data architecture to standardize the link between the 'current' version and the 'frozen' one.

Architecture components

1. Metadata: tracking of changes due to a demographic event;

2. Statistical Service: 1) update 'current' version;
   2) quality check before updating 'frozen' version;
   3) update 'frozen' version;

3. Data Service: read/write data from/to the frame 'current' and 'frozen' version in the Data Repository.

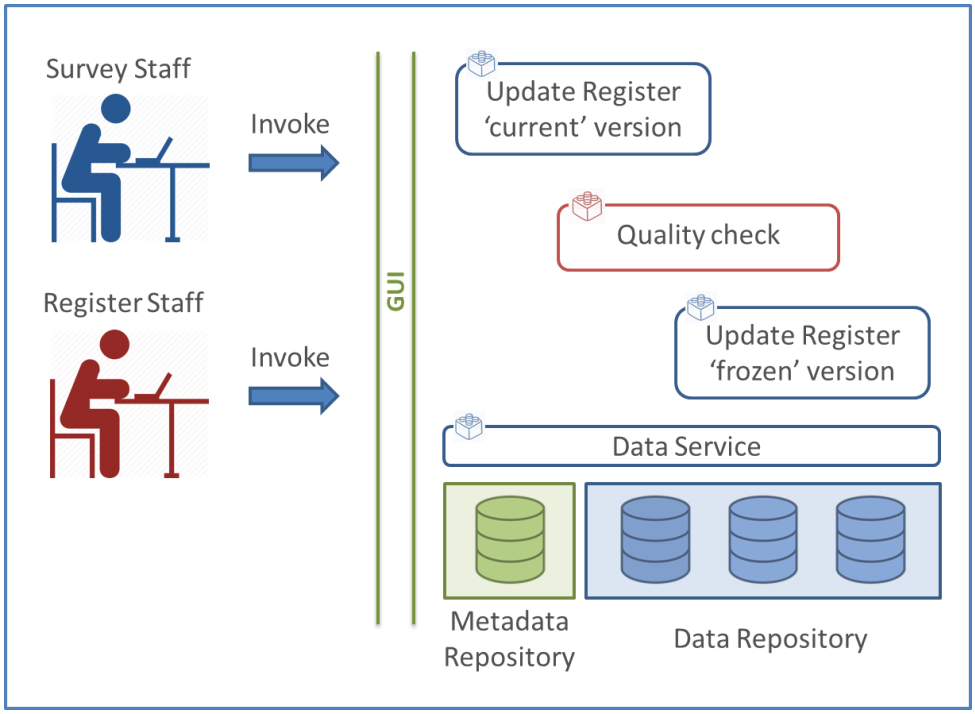

Figure 6: Demographic event management

A user involved in data collection, records a business demographic event and using the 'Update Register current version' Data Service stores this information in the Data Repository. The Register Staff, using the 'Quality check' Statistical Service, performs accuracy checks of data changes and periodically runs the 'Update Register frozen version' Data Service.

Technical remarks

Considering the scheduling of updates from 'current' to 'frozen' version, the first issue is to evaluate the best timing that reduces the trade-off between timeliness and accuracy of the stored information and statistical outputs to be built from the Register.

The second issue concerning the workflow, arises from the need to inform and prioritize the different stakeholders about data changes and updates. A possible solution could be the implementation of a publish-subscribe messaging pattern.

5. QUESTIONS

1. How can we manage the integration of ad-hoc legacy applications in a standard environment built on generalized functionalities? How to cope with data processing parameters edited by end user at runtime, through graphical interactive interfaces?
2. How can we implement a user-friendly workflow, easy to configure and manage by all type of users, despite their knowledge of software applications? Which is the most suitable environment?
3. As the process chain has a low degree of complexity, is it better to develop an in-house solutions or to evaluate a commercial product. (e.g. enterprise service bus, orchestrator, etc.) to manage the workflow?
4. The generalized statistical production process can be potentially abstracted as a "data-driven" workflow, in line with the scientific nature of Istat's processes. Can we rely on the characteristics of data-driven workflows for designing and implementing more efficient solutions for the GPBS?
5. Concerning the integration of various statistical services in the workflow, how to design structural metadata to describe the data structures passed as input/output to services and/or accessed as central data repository?
6. We do have process metadata describing for instance the process flow that could have an "active" way in being interpreted for the execution. We do however have also other kinds of process metadata related for instance to the naming of services and processes themselves. How do we manage such kinds of process metadata?

6. REFERENCES

[1] Generic Statistical Business Process Model (GSBPM)
https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0

[2] Generic Statistical Data Editing Models (GSDEMs)
https://statswiki.unece.org/display/kbase/GSDEMs

[3] Business registers - recommendations manual (2010 Edition)
https://unstats.un.org/unsd/EconStatKB/KnowledgebaseArticle10171.aspx

[4] Enterprise Architecture Reference Framework (EARF)
https://ec.europa.eu/eurostat/cros/content/ess-enterprise-architecture-reference-framework_en

[5] Statistical Production Reference Architecture (SPRA)
https://ec.europa.eu/eurostat/cros/content/spra_en

[6] Quality Assurance Framework of the European Statistical System
http://ec.europa.eu/eurostat/web/quality

ANNEX: INDUSTRIAL PRODUCTION SURVEY AS-IS PROCESS DESCRIPTION

The industrial production index is one of the main indicators to assess a country's economic performance, and measures changes over time in the physical volume of goods produced by industry, excluding the construction sector. This survey is performed monthly to collect information concerning the physical quantities of a representative basket of products, selected according to the Prodcom Classification. During the interview, the analysis has been focused on the main features of the statistical process, thus considering only a subset of GSBPM phases. The language used for modeling each step of the statistical process is ArchiMate.

The following figure summarize, the reference population, the unit type, the collected unit and the sample design strategy.

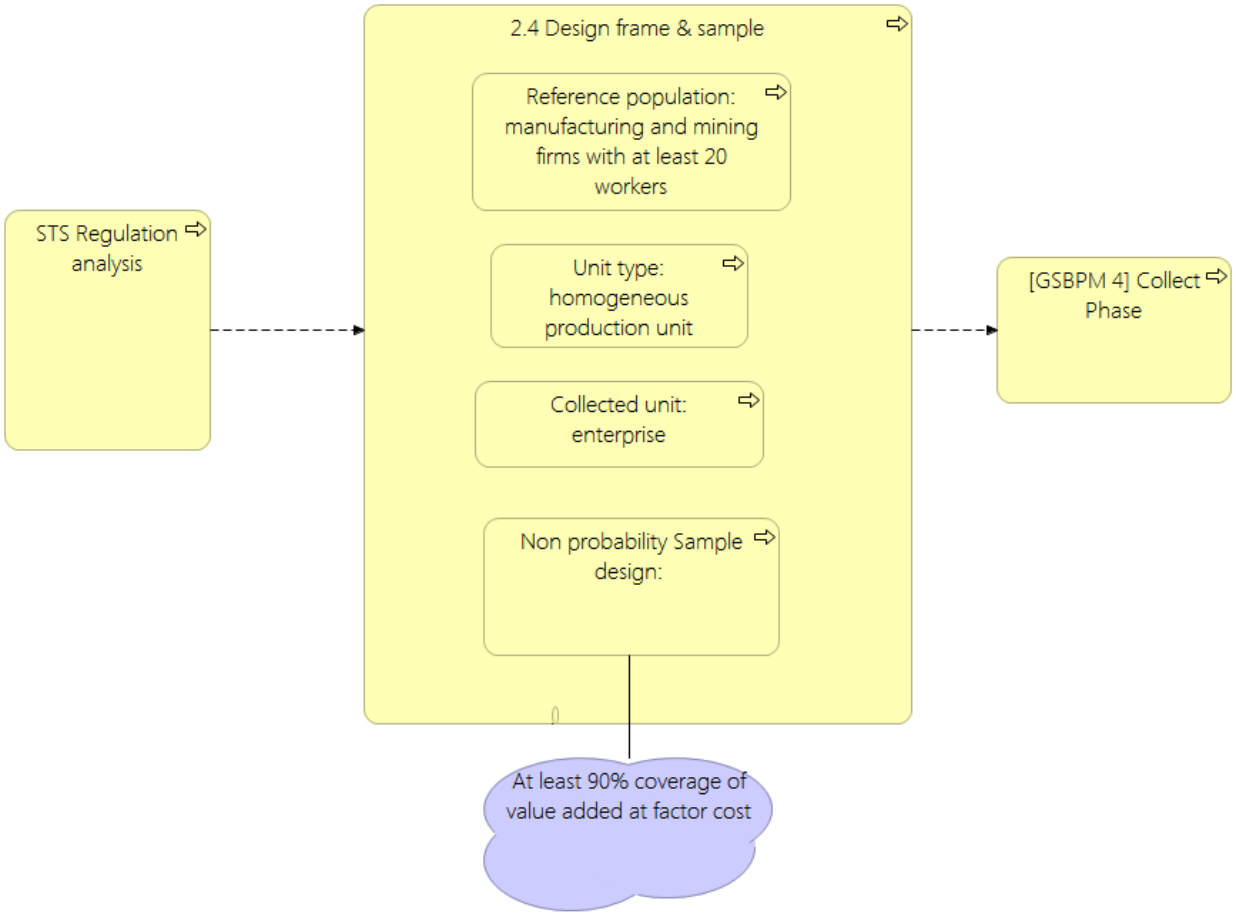

Figure 1: GSBPM phase 2.4

An overview of data collection phase, and the auxiliary information used, is shown in Figure 2. The application layer corresponds to the blue objects.

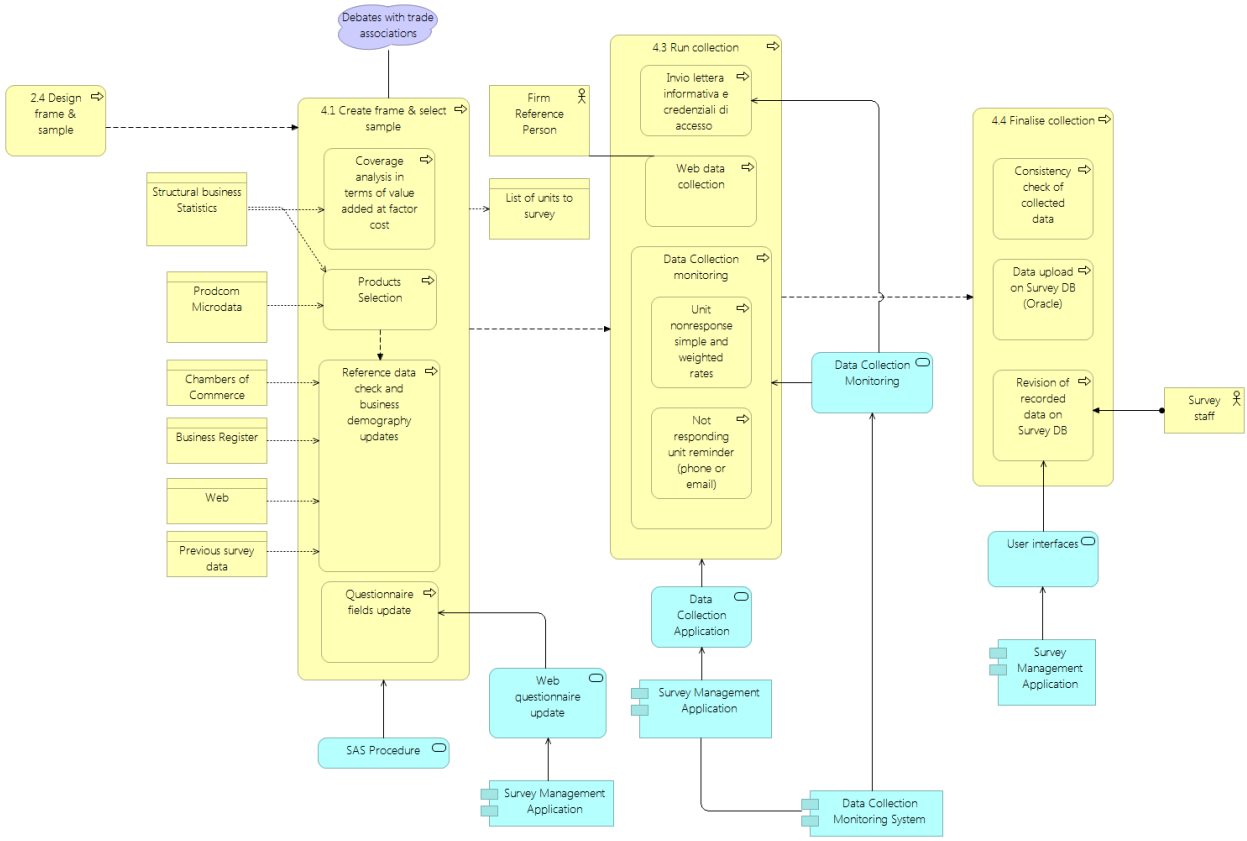

Figure 2: GSBPM Collect phase

The main steps of data processing following data gathering and the related application framework are represented in Figure 3. In this case, data editing and imputation are performed by an ad-hoc survey application.

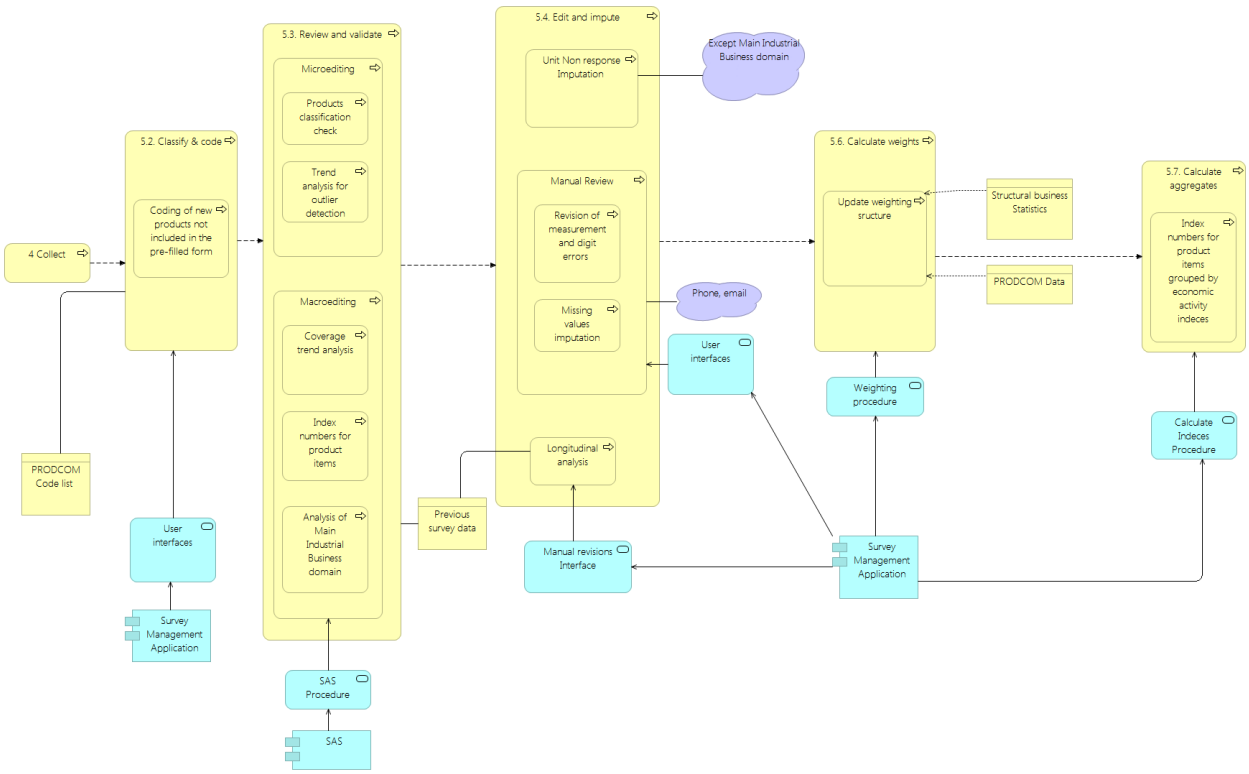

Figure 3: GSBPM Process phase

The last phase modeled, describes the steps to produce and validate draft outputs for data dissemination.

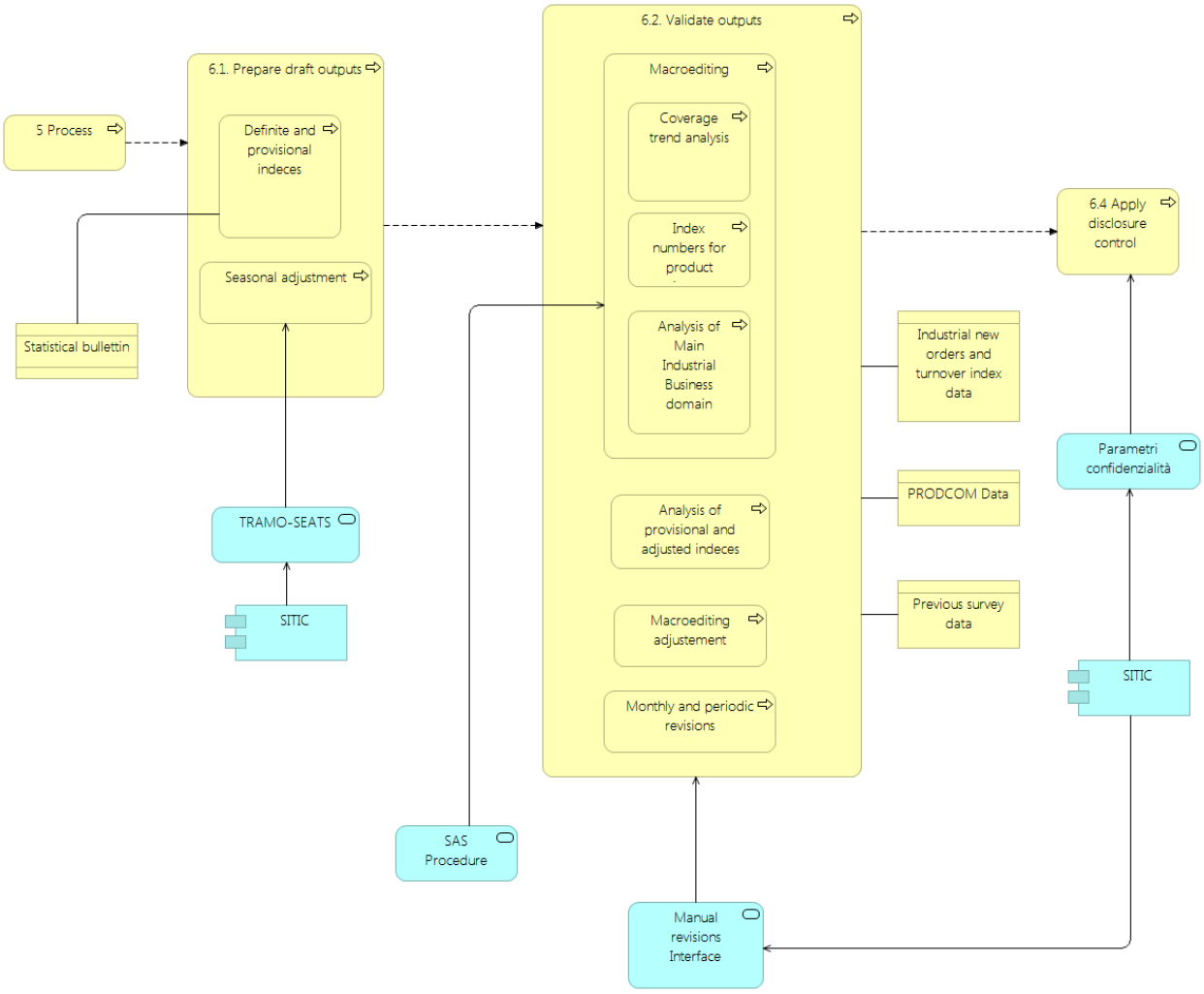

Figure 4: GSBPM Analyse phase