Functional Architecture of the Statistical Data Warehouse

Sónia Patrícia F. C. B. Quaresma Gonçalves

Information Systems and Computing Department

National Statistical Institute

Portugal

sonia.quaresma@ine.pt

Functional Architecture

We propose a generic business process on the S-DWH divided in four, focused on functionalities groups each specialized in a data layer. The metadata used and produced in the different layers of the warehouse are defined in the linking[1] and framework[2] Metadata.

To describe the main high level functionalities of the S-DWH from users' viewpoints we will introduce a Functional Architecture diagram (FA), this will be described by the Generic Statistical Information Model (GSIM), using the Generic Statistical Business Process Model (GSBPM) convention when needed. The 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 (information objects).

A Functional Diagram (FD) reflects a software product's architecture from a usage perspective. In the S-DWH context this work is performed by the NSI users, or official statistics producers. In order to enable FA to communicate with stakeholders, even with no specialization in software architecture, we borrow the functional diagram notation from the Enterprise Architecture (EA) modelling technique, which is used to model the primary process of an enterprise and its physical and administrative functions. Consequently, a FD will contain modules that represent the basic functions of a software product.

We start the description focusing our attention on the management functionalities that interact with the S-DWH system, afterwards we will analyse internal functionalities to describe hierarchical functional representation.

In the follow discussion we will use these four conceptual groups to connect the nine statistical phases with the over-arching management process of the GSBPM.

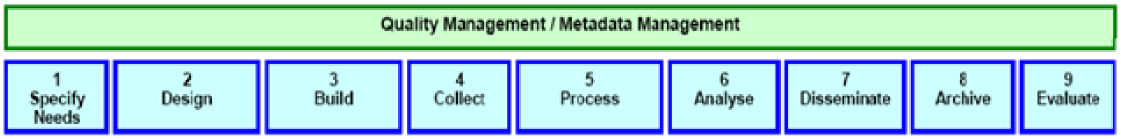

Figure 1 – Management and the Nine statistical phases of the GSBPM

[1] Ennok M et al. (2013) On Micro data linking and data warehousing in production of business statistics, ver. 1.1.

[2] Lundell L.G. (2012) Metadata Framework for Statistical Data Warehousing, ver. 1.0.

FD Strategic functionalities

The strategic management processes among the over-arching processes stated in GSBPM and in the extension for the S-DWH management functionalities falls outside S-DWH system but are still vital to it. Those strategic functions are:

1. Statistical Program Management.
2. Business Register Management.
3. Web Communication Management.

The functional diagram below illustrates the relationship between the strategic over-arching processes and the operational management.

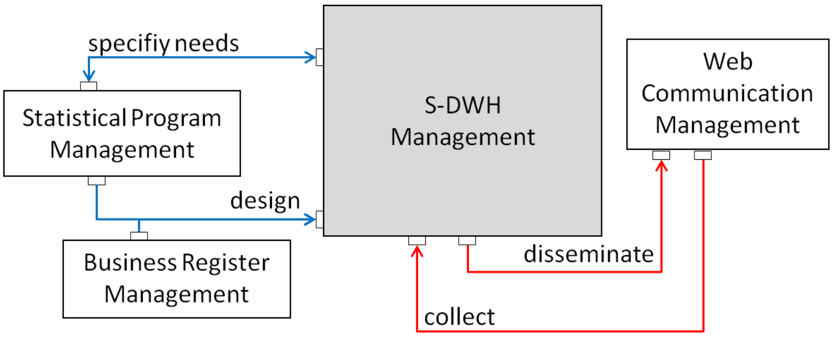
specifiy needs
Statistical Program Management
S-DWH Management
Web Communication Management
design
Business Register Management
disseminate
collect

Figure 2 – High level functional diagram (FD) representation

In the functional diagram the utilities are represented by modules whose interactions are represented by flows. The diagram is a collection of coherent processes, which are continuously performed. Each module is described with a box and contains everything necessary to execute the represented functionality.

As far as possible the GSBPM and GSIM are used to describe the functional architecture of an S-DWH, thus the colours of the arrows in the functional diagrams refers to the four conceptual categories already used inside the GSIM conceptual reference model. The Structures Group (yellow) contains sets of information objects that describe and define the terms used in relation to data and their structure (e.g. Data Sets).

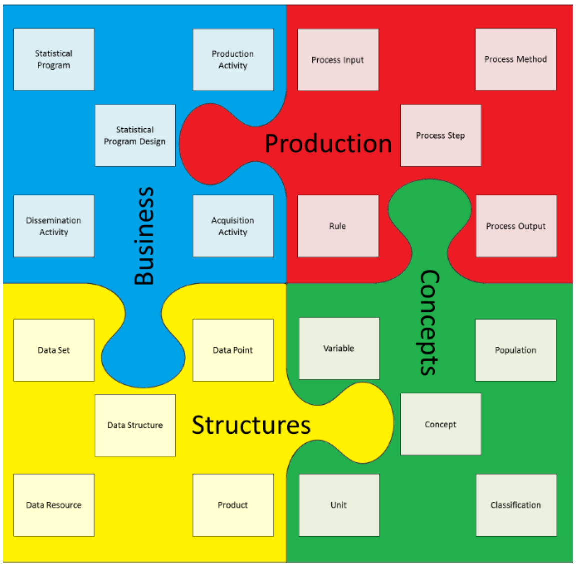

Figure 3 – General Statistical Information Model (GSIM)

The functional diagram in Figure 2 shows that the identification of new statistical needs (Specify Needs phase) will trigger the initiation of a *Statistical Program*. This, in turn, will then trigger a design phase (in GSIM, the *Statistical Program Design*, will lead to the development of a set of *Process Step Designs* - i.e. all the sub-processes, business functions, inputs, outputs etc. that are to be used to undertake the statistical activity).

The basic input process for new statistical information derives from the natural evolution of the civil society or the economic system. During this phase, needs are investigated and high level objectives are established for output. The S-DWH is able to support this process by allowing the use of all available information to analysts to check if the new concepts and new variables already are managed in the S-DWH. The design phase can be triggered by the demand for a new statistical product, as a consequence of a change associated with process improvement, or perhaps as a result of new data sources becoming available. In each case a new *Statistical Program* will be created, and a new associated *Statistical Design*.

The web communication management is an external component with a strong interdependency with the S-DWH since it is the interface for external users, respondents and scientific or social society. From an operational point of view the assurance of a contact point accessible over internet, e.g. a web-portal is a key factor for good respondent relationships, services related to direct or indirect data capturing and information products deliverance.

FD Operational Functionalities

In order to analyze the functionalities which support a generic statistic business process we describe the functional diagram of Figure 2 in more detail. Expanding the module representing the S-DWH Management, we can identify four more management functionalities within it:

1. Statistical Framework Management.
2. Provider Management.
3. Process Metadata Management.
4. Data Management.

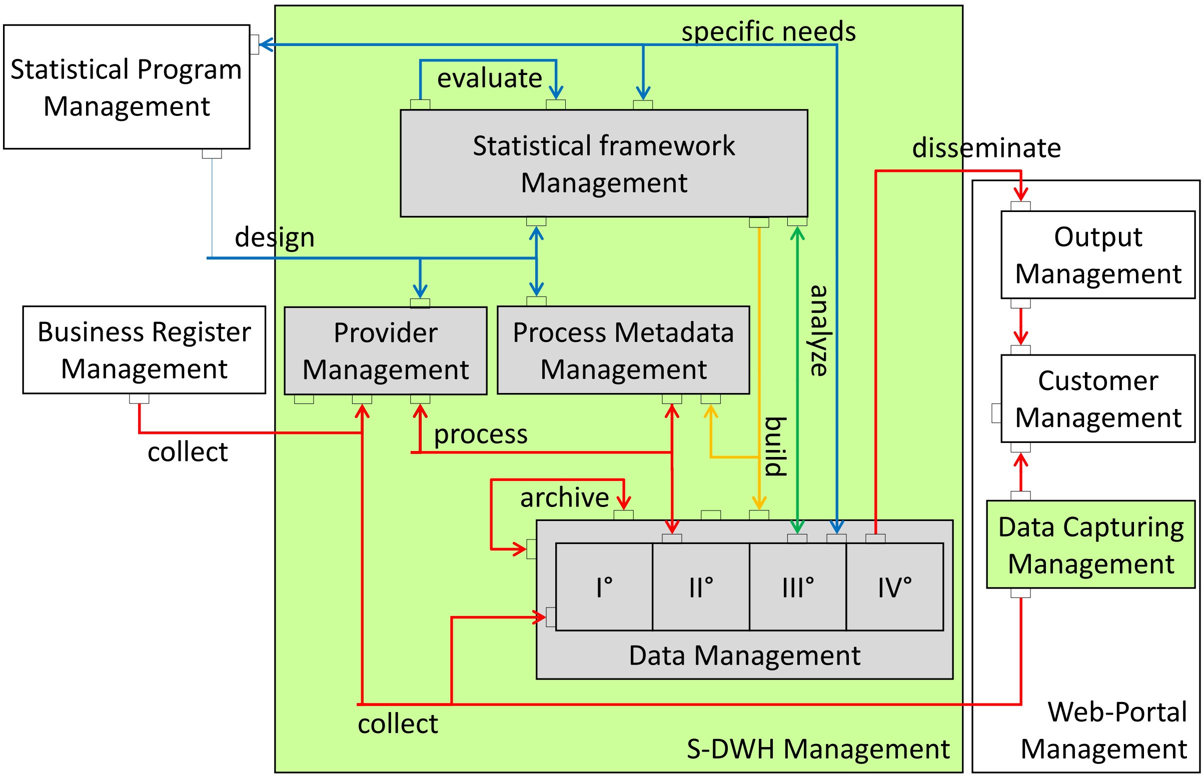

Figure 4 – Functional Diagram, expanded representation

Furthermore, by expanding the Web-Portal Management module we can identify three more functionalities: Data Capturing Management, Customer Management and Output Management.

The details in Figure 4 enable us to contextualize the nine stages of the GSBPM in a S-DWH functional diagram. We represent those nine parts using connecting arrows between modules. For the arrows we used the same four colours used in the GSIM to contextualize the objects.

The flows depicted which map to nine phases of the GSBPM will be discussed in the next sections.

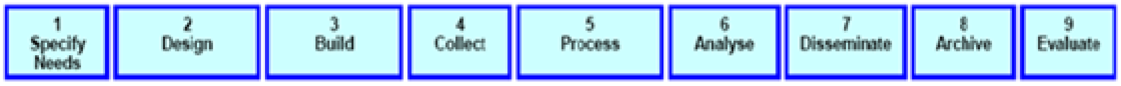

Figure 5 – Nine statistical phases of the GSBPM

"Specify Needs" path

This segment represents the request for new statistics or an update on current statistics. The flow is blue since this phase represents the building of Business Objects from the GSIM, i.e. activities for planning statistical programs. This phase is a strategic activity in a S-DWH approach because a first overall analysis of all available data and meta data is made.

In the diagram we identify a sequence of functions starting from the Statistical Program passing through the Statistical framework and ending with the Interpretation layer of Data Management. This module supports executives in order to "consult needs", "identify concepts", "estimate output objectives" and "determine needs for information".

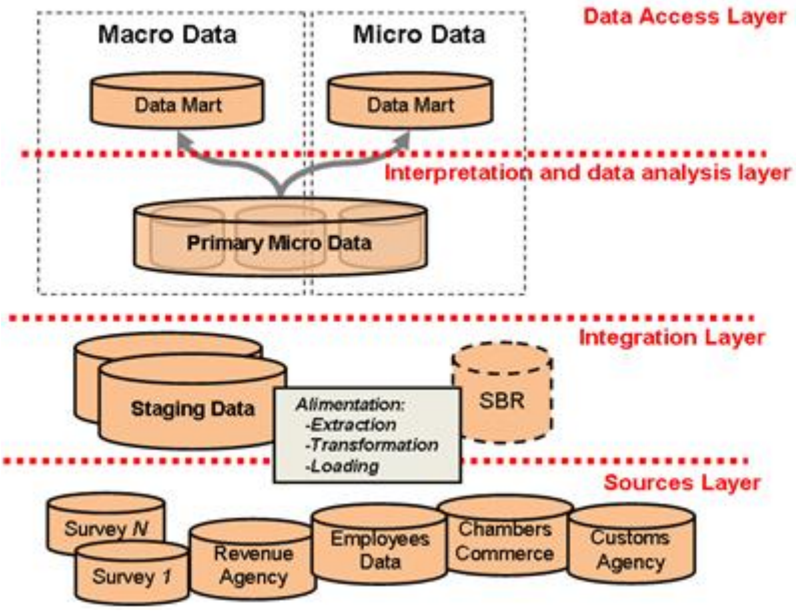

Figure 6 – S-DWH Layers simplification

The connection between the Statistical framework and the Interpretation layer data indicates the flow of activities to "check data availability", i.e. if the available data could meet the information needs or the conditions under which data would be available. This action is then supported by the "interpretation and analysis layer" functionalities in which data is available and easy to use for any expert in order to determine whether it would be suitable for the new statistical purposes.

At the end of this action, statisticians should prepare a business case to get approval from executives or from the Statistical Program manager.

"Design Phase" path

This pointer stands for the development and design activities, and any associated practical research work needed to define the statistical outputs, concepts, methodologies, collection instruments and operational processes.

All these sub-processes can create active and/or passive metadata, functional to the implementation process. Using the GSIM reference colours we colour this flow in blue to describe activities for planning the statistical program, realized by the interaction between the statistical framework, process metadata and provider management modules. Meanwhile the phase of conceptual definition is represented by the interaction between the statistical framework and the interpretation layer.

The information related to the "design data collection methodology" impacts on the provider management in order to "design the frame" and "sample methodology". These designs specify the population of interest, defining a sample frame based on the business register, and determine the most appropriate sampling criteria and methodology in order to cover all output needs. It also uses information from the provider management in order to coordinate samples between instances of the same statistical business process (for example to manage overlap or rotation), and between different processes using a common frame or register (for example to manage overlap or to spread response burden).

The operational activity definitions are based on a specific design of a statistical process methodology which includes specification of routines for coding, editing, imputing, estimating, integrating, validating and finalizing data sets. All methodological decisions are made using concepts and instruments defined in the Statistical Framework, and the workflow definition, able to support the production system, is managed inside the Process Metadata. If a new process needs new concepts, variables or instruments these are defined then in the Statistical Framework.

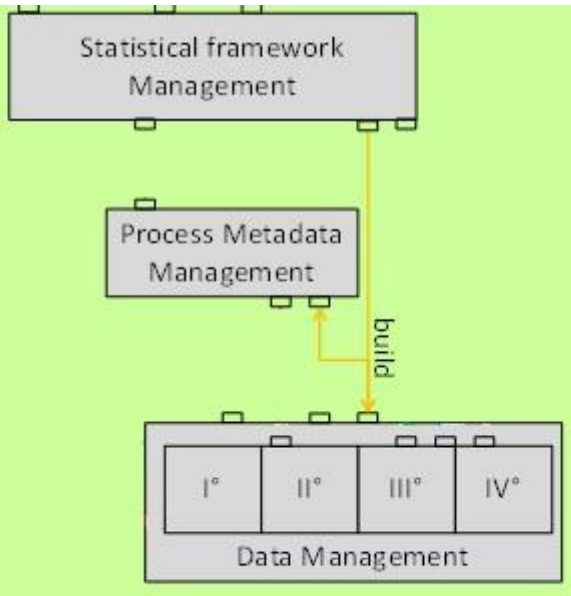

Figure 7 – Build path

"Build Phase" path

In this part all sub processes are built and tested for the systems component production. For statistical outputs produced on a regular basis, this phase usually occurs for the first iteration, and following a review or a change in methodology, rather than every time.

In a S-DWH which represents a generalized production infrastructure this action is based on code reuse and each new output production line should only consist in a work flow configuration. This has a direct impact on active metadata managed by process metadata in order to execute the operational production flows properly. Maintaining the consistency with the GSIM, we colour this flow in yellow. Therefore, in a S-DWH the build phase can be seen as a metadata configuration able to interconnect the Statistical Framework with the DWH data structures.

"Collect Phase" path

This stage includes all collection activities for all necessary data, and loads data into the source layer of the S-DWH. This represents the first step of the operational production process and for that reason, in analogy with the GSIM, we colour this flow in red.

The two main modules involved with the collection phase in the functional diagram are Provider Management and Data Capturing Management. Provider Management includes:

- Cross-Process Burden
- Profiling
- Contact Information Managements.

This is done by optimizing register information using three information inputs':

1. From the external official Business Register;
2. From respondents' feedback;
3. From the identification of the sample for each survey;

Data capturing management collects external data into the source layer. Typically this phase does not include any data transformations.

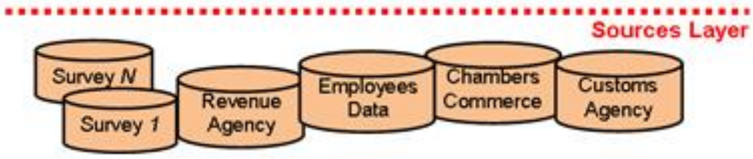

Figure 8 – Sources layer in the S-DWH

We distinguish between two kinds of typologies of data capturing: controlled and not controlled systems. The first is data collection directly from respondents using instruments which should include the sharing of variable definitions and first checks. A typical example is a web questionnaire. The second typology is represented by data collected from an external archive. In this case, before any data uploading a conceptual mapping between internal and external statistical concept is necessary. Data mapping involves combining data residing in different sources and providing users with a unified view of these data. These systems are formally defined as a triple <T,S,M> where T is the target schema, S is the heterogeneous set of source schemas, and M is the mapping that maps queries between the source and the target schemas.

"Process Phase" path

Processing encompasses the effective operational activities made by reviewers. It is based on specific explanation steps and corresponds to the typical ETL phase of a DWH. In a S-DWH it describes data records cleansing and their preparation for output or analysis. The operational sequence of activities follows the design of the survey configured in the metadata management. This phase corresponds to the operational use of modules and for this reason, in accordance with the managing of production objects of the GSIM, we colour this flow in red.

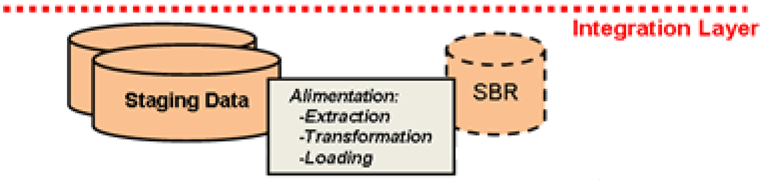

Figure 9 – Integration Layer in the S-DWH

All the sub process "classify & code", "review", "validate & edit", "impute", "derive new variables and statistical units", "calculate weights", "calculate aggregate", "finalize data files" are made up in the "integration layer" following ad hoc sequences in function of the typology of the survey. The "integrate data" connects different sources and uses the provider management in order to update asynchronous business register status.

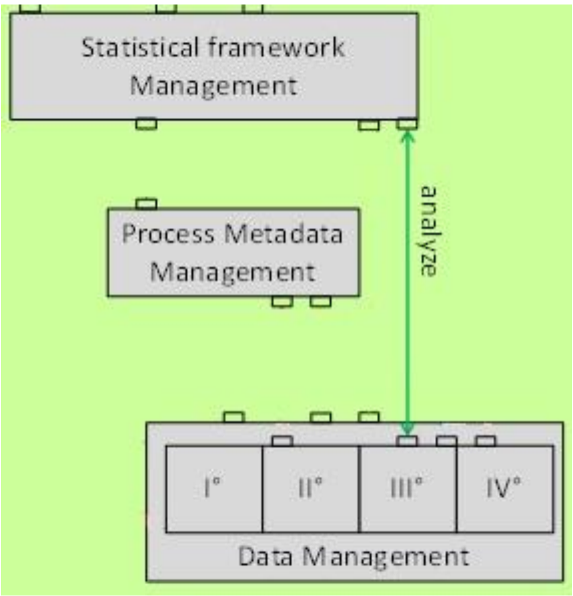

Figure 10 – Analyze path

"Analyze Phase" path

This phase is central for any DWH, since during this phase statistical concepts are produced, validated, examined in detail and made ready for dissemination. We therefore colour the activities flow of this phase in green in agreement with the GSIM.

In the diagram the flow is bidirectional connecting the statistical framework and the interpretation layer of the data management. This is to indicate that all non consolidated concepts must be first created and tested directly in the interpretation and analysis layer. It includes the use or the definition of measurements such as indexes, trends or seasonally adjusted series. All the consolidated draft output can be then automated for the next iteration and included directly in the ETL steps to produce an output directly.

The Analysis phase includes primary data scrutinizing and interpretation to support the data output. The inspection provides statisticians with a profound knowledge of the statistic data. They use that understanding to explain the statistics produced in each cycle by evaluating and measuring the effective fitting with their initial expectations.

“Disseminate Phase” path

Dissemination phase manages the release of the statistical products. It occurs always for all regularly produced statistical products. From the GSBPM we have five sub processes: “updating output systems”, “produce dissemination products”, “manage release of dissemination products”, “promote dissemination products” and “manage user support”. All of these sub process can be directly related to the operational data warehousing.

The “updating output systems” sub process is the effective arrow connecting the Data Management with the Output Management. We colour this flow in red, to indicate the operational data uploading. The Output Management produces dissemination products, manages the release and promotes dissemination products using the information stored in the “access layer”.

At last the finalized output sub process ensures that the statistics and associated information are fit for purpose, reach the required quality level, and are thus ready for use. This sub process is manly executed in the “interpretation and analysis” and their evaluations are available at the access layer.

“Archive Phase” path

This part manages the archiving and disposal of statistical data and metadata. Thinking about the S-DWH as an integrated data system, this phase must be considered to be an over-arching activity; i.e. it is a central structured generalized activity for all S-DWH levels. We include in this phase all operational structured steps needed to the Data Management, therefore we colour in red this flow to indicate the family of objects managed in this phase to maintain all kind of data.

In the GSBPM four sub processes are considered: “definition archive rules”, “management of archive repository”, “preserve data and associated metadata” and “dispose of data and associated metadata”. Among those the “definition archive rules” is a typical activity on metadata while the others are operational functions.

The archive rules sub process defines structural metadata, for the definition of the structure of data (data mart and primary), metadata, variables, data dimensions, constraints, etc., and it defines process metadata, for specific statistical business process as a general archiving policy of the NSI or standards applied across the government sector.

The other sub processes concern the management of one or more data bases, the preservation of data and metadata and their disposal, these functions are operational on a S-DWH and depend on its design.

“Evaluate Phase” path

This phase provides the basic information for the overall quality evaluation management. The evaluation is applied to all the S-DWH layers through the statistical framework management. It takes place at the end of each sub process and the gathered quality information is then stored into the relative metadata structures of each layer. Evaluation material may take many forms, from monitoring systems to log files, feedback from users or staff suggestions.

For statistical outputs produced regularly, evaluation should, at least in theory, occur once for each iteration, determining whether future iterations should take place, and if so, whether any improvements should be implemented. In a S-DWH context the evaluation phase always involves the evaluation of groups of business process for an integrated production.