Building the statistical data warehouse to improve statistics

Harry Goossens

Coordinator ESSnet on Data Warehousing

Statistics Netherlands (CBS), e-mail: hct.goossens@cbs.nl

Abstract

Within the European Statistical System (ESS) there is a strong focus on the modernisation of statistical production, where decreasing costs and administrative burden versus increasing efficiency and flexibility are key words. In order to further improve and optimise statistical production, statistical institutes (NSIs) are searching for ways to make optimal use of all available data sources, existing and new. Next to possibilities to disclose all kinds of new data sources that become available through the global use of modern technologies (like internet, mobile phones, etc.), one of the major challenges in this process of change is the integration of the re-use of statistical data that is already available within an NSI:

How to make optimal use of all available data sources (existing and new) ?

Hereby the focus is not only on the use as input for producing the statistics they are designed and collected for. More and more NSIs are looking for possibilities to re-use already available data as source/input to match (new) data demands from statistics, in order to further improve and optimise statistical production.

This modernisation has also an important organisational impact. Next to the need for a complete new way of organising the statistical production process, it also comes with higher and stricter demands for the data **and** metadata management. These two activities are often decentralised and implemented in various ways, depending on the needs of specific statistical systems, whereas realising maximum re-use of available statistical data just demands the opposite: a centralised and standardised, flexible and transparent metadata catalogue that gives insight in and easy access to all available statistical data.

To reach this goal, building a statistical data warehouse (S-DWH) is considered to be a crucial instrument. The S-DWH approach enables NSIs to identify the particular phases and elements in their statistical production process that need to be common and reusable. Of course there are several ways of defining the S-DWH: a strong focus on data access and output or also process integration (process driver), static/data storage or dynamic/data flow ?

But in all various concepts the goal is the same:

To create a central data hub, integrating all available data sources and statistical output.

Keywords: S-DWH (statistical data warehouse), re-using statistical data, metadata management.

1. Background - The ESSnet project on data warehousing

One of the main actions of the MEETS programme (Modernisation of European Enterprise and Trade Statistics foresees to *"make better use of data that already exist in the statistical system, including the possibility of estimates"*, with as ultimate aim:

'To create fully integrated data sets for enterprise and trade statistics at micro level:

- a data warehouse approach to statistics.'

In this context, in October 2010 the "ESSnet on micro data linking and data warehousing in statistical production" was established to provide assistance in the development of more integrated databases and data production systems for (business) statistics in ESS Member States. The ESSnet's main goal in daily statistical practice is to increase the efficiency of data processing in statistical production systems and to maximize the reuse of already collected data in the statistical system.

As the field of data warehousing, and thus the scope of this ESSnet, is very broad, activities should focus on the specific, detailed and prioritized subjects to explore and study in depth, determined by the ESS members. Main conclusions of an ESS-wide questionnaire was that there is great interest in the topic, ('data warehousing is hot') and that there is a clear need for advice and active support in setting-up and building a statistical data warehouse. All information and documents about the scope, results and status of the on-going activities are available on the projects website: http://www.cros-portal.eu/content/data-warehouse

2. Defining the S-DWH

First it is essential to create a clear and common understanding of the statistical data warehouse (S-DWH), so the broad definition of the S-DWH in this ESSnet is defined as:

> *'A central statistical data store for managing all available data of interest, enabling the NSI to (re)use this data to create new data / new outputs, to produce the necessary information and perform reporting and analysis, regardless of the data's source.'*

But for the ESSnet it is important to find out what the members states define as statistical data warehouse. Therefore this definition is redefined as:

> *'A system or set of integrated systems, designed to handle the processing of statistical data in the production of (business) statistics'.*

These definitions are used to design two idealised representations of perspectives on statistical processes, which the ESSnet called the "Data Model" and "Process Model"[1]. In the data model' perspective, the core is a unit for storing and processing data, irrespective of where it has come from or where it is going to. The process/store is not designed around either the type of input or output, but around the data item. From a technical point of view this perspective can be considered as a "top-down approach", meaning that the data warehouse is built from a (new) overall concept. All processes can be fitted in this data warehouse in one or several steps. In the 'process model' perspective, the data warehouse for storing and processing data is the set of production processes needed to manage the inputs and generate the outputs. From a technical point of view this perspective can be considered as a "bottom-up" approach", meaning that the data warehouse is built up from (stepwise implementing) current processes.

Based upon these definitions a conceptual model of the statistical data warehouse is defined, setting the scope/boundaries of the statistical warehouse and representing all various stages and elements between input and output [figure 1][2].

The shaded areas be considered the 'statistical data warehouse' comprising:

- technical facilities for storing and processing data, receiving data in and producing outputs in a flexible way
- rules for updating the sources for the DWH
- rules for generating samples
- definitions necessary to achieve those sample/source generation
- the data flow model

In daily statistical practice, the S-DWH is the central data hub, which enables the connection and integration of all kinds of (new) data sources with statistical output. Therefore the S-DWH must support all statistical production processes (including data collection) by providing:

- a detailed and correct overview/insight of already available data sources;
- a framework for adequate data governance, including metadata management, confidentiality aspects and data authorisation;
- access to registers sampling frames (BR, etc.);
- flexible data storage and data exchange between processes.

---

[1] Ritchie, F. and Goossens, H. (2011) "Data Warehousing: The conceptual perspective" http://www.cros-portal.eu/content/data-warehouse (see *Data Warehouse (SGA1)/Documents)*

[2] Ritchie, F. (2011) "Explaining the Statistical Data Warehouse" http://www.cros-portal.eu/content/data-warehouse (see *Data Warehouse (SGA1)/Final Report/Annex 9)*

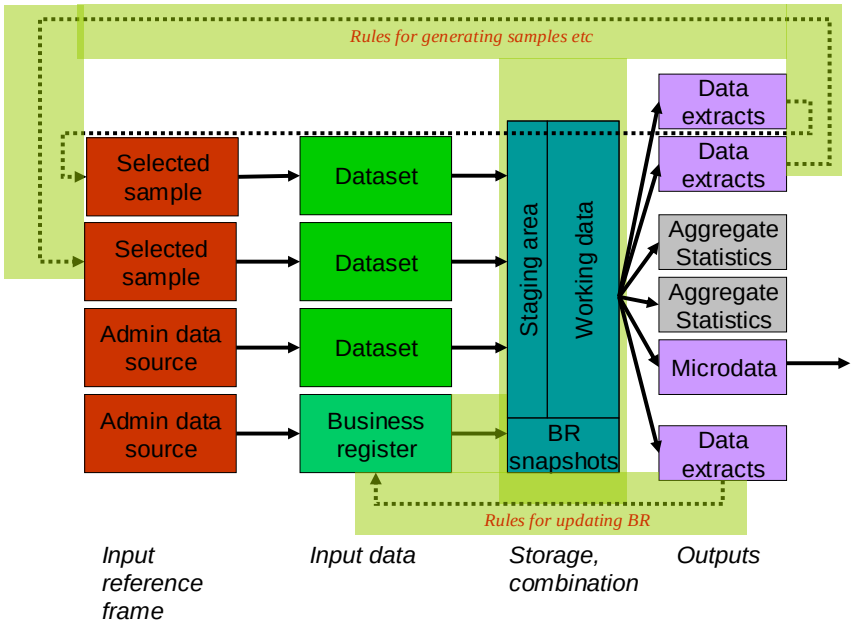

Figure 1: Explaining the S-DWH

3. The layered business architecture of the S-DWH

A data warehouse is a concept that intends to provide an architectural model for the dataflow from operational systems to decision support systems, in the case of the S-DWH, from data collection systems to statistical output systems. In this context, the architecture is the conceptualisation on how to build up a statistical data warehouse. This means defining a common model for the total statistical production as an integrated, comprehensive production system, covering all different statistic domains. The structured data in a S-DWH must be organized in a way that *enables* statisticians, involved in statistical production, to "reuse" data by creating new statistical information or data output and enabling users to re-use the produced information for any possible new needs. So in order to implement a S-DWH, the first step is the conceptualisation of such an architectural model of the data flow from the sources, surveys or administrative archives, through processing till statistical outputs environments. To provide such a model for the generic S-DWH, the ESSnet identified four functional layers, each for specific statistic activities:

- access layer
- interpretation and data analysis layer
- integration layer
- source layer

and used them to define a layered Business Architecture for the S-DWH, representing the various statistical data used by each layer [figure 2][3].

For a better understanding of the 4 layers, you have to go inside them from the perspective of a generic functional analysis of statistical production:

- The Source layer is the level for, physically or virtually, storing the data from internal (surveys, existing micro data) or external (administrative data, archives) sources for tatistical purpose. The source layer is the interface towards all external actors participating in data collection.

[3] Quaresma, S., Laureti Palma, A.(2011) "Hypothesis of a DW architecture for business statistical production" http://www.cros-portal.eu/content/data-warehouse (see *Data Warehouse (SGA1)/Final Report/Annex 10)*

- The Integration layer is used for all integration and reconciliation activities of data sources in order to become a first integrated staging area, independent from sources. This layer has set of applications/tools to perform all operational activities for regular statistical production, carried out, automatically or manually, by users.
- The Interpretation and Data Analysis layer is specifically for statisticians and enables any data manipulation or unstructured activities. In this layer expert users can carry out data mining or design new statistical processes. In general, the output of these activities is aggregate data for the next access layer or specific engineering of the next iterations.
- The Access layer is the layer for the final presentation, dissemination and delivery of the statistical information sought. It is open for a wide range of users and using various exploration tools. The data organization must support automatic dissemination systems and free analysis, in both cases, statistical information is macro data.

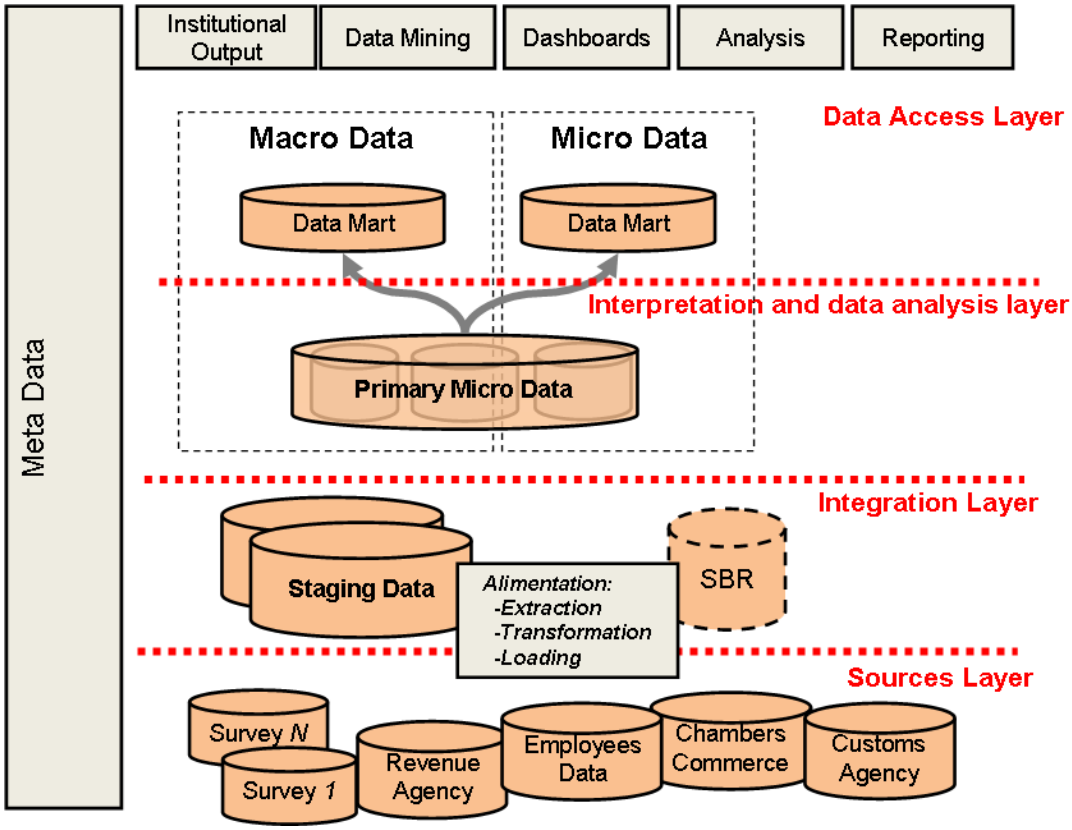

Figure 2: **The layered architecture of the S-DWH with focus on "statistical data", used by each layer.**

After the identification of the architecture of a S-DWH the next step is to find a common language to identify and locate the different phases of a generic statistical production process on the different functional levels of the S-DWH. This common language is best represented by the Generic Statistical Business Process Model (GSBPM), which intends to define and share a common statistical framework for statistical production[4]. For getting good understanding of the influence of a S-DWH approach for statistical production the architecture of five EU-regulated business statistics (SBS, STS, ProdCom, Trade statistics and Business Register) were analysed, as managed by the partners of this ESSnet.

[4] http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model

For a homogeneous approach of this analysis we used a simplified Business Process Model Notation, based upon the GSBPM glossary and thus created a generic graphical representation of the statistical (sub-) processes by mapping it on the layered S-DWH. After merging the results of the separate analysis we were able to represent a generic workflow of the production process in the S-DWH.[5]

4. Metadata in the S-DWH

One of the key factors and drivers in a S-DWH is the information about one or more aspects of the data itself, are usually referred to as "metadata" (or meta content).

> *'Metadata is the DNA of the data warehouse, defining its elements and how they work together. [...] Metadata plays such a critical role in the architecture that it makes sense to describe the architecture as being metadata driven'.*[6]

The metadata provides the access to the data and must enable a clear and unambiguous description of the data and its elements. In order to better identify the role of the metadata in a S-DWH the ESSnet defined the metadata framework of the S-DWH.[7] In this document we identified the various metadata categories and metadata subsets. Based upon these definitions and keeping in mind the specific metadata requirements of statistics production it is possible to assess metadata requirements of the S-DWH:

- The SDWH requires *active* metadata. The amount of objects (variables, value domains, etc.) stored makes it necessary to provide the users (persons and software) with active assistance finding and processing the data.
- The SDWH requires formalised metadata. The amount of metadata items will be large and the requirement for metadata to be active makes it necessary to structure the metadata very well.
- The SDWH requires *structural* metadata, especially *technical* metadata. Active metadata must be structural, at least to some part.
- *Process* metadata are vital to a SDWH. Since the data warehouse supports many concurrent users it is very important to keep track of usage, processing results, performance, etc.

Metadata Subset	Metadata category							
	Formalised				Free-form			
	Reference		Structural		Reference		Structural	
	Act	Pas	Act	Pas	Act	Pas	Act	Pas
Statistical			dw			gen		
Process	dw		dw	dw	gen			gen
Quality		dw				gen		
Technical			dw					
Authorisation			gen					
Data model				dw				dw

Table 1: Combining metadata categories and subsets

Table 1 shows the possible combinations of metadata categories and subsets. In the cells are indicated which combinations are of general interest for statistics production ("gen") and which ones are of particular interest for a S-DWH ("dw"). Most of the remaining combinations are possible, but less common or less likely to be found useful.

[5] Randlepp, A.(2013) "The S-DWH Modular Workflow"- version 2.0 http://www.cros-portal.eu/content/data-warehouse (see *Data Warehouse (SGA2)/2.3 WP3 Technical/Deliverables)*

[6] Kimball, R. The Data Warehouse Lifecycle Toolkit (Second Edition), Wiley, 2008, p. 117

[7] Lundell, L.G. (2012) "Framework of metadata requirements and roles in the S-DWH, version 1.0 http://www.cros-portal.eu/content/data-warehouse (see *Data Warehouse (SGA2)/2.1 WP1 Metadata/Deliverables)*

5. Managing the S-DWH

The ESSnet has defined the S-DWH as "a central statistical data store, regardless of the data's source". This definition should be understood as a logically coherent data store, not necessarily as one single physical unit. The logical coherence means that it must be possible to uniquely identify a data item throughout the data warehouse, to trace it on its way through the logical layers from input to dissemination, and to follow it longitudinally. From the requirements on data follows a crucial metadata requirement for the S-DWH: all metadata items (concepts as well as physical references) must be consistent throughout the data warehouse and must be uniquely identified, with a one-to-one relationships between identity and definition, and identity and name. A user must be able to search the entire metadata layer and, if permitted, to access data in the logical S-DWH without actual knowledge of their physical locations. All data in the S-DWH must have corresponding metadata ('no data without metadata'), all metadata items must be uniquely identifiable, metadata should be versioned to enable longitudinal use, etc. Finally, metadata must provide "live" links to the physical statistical data.

Thus, metadata plays a vital role in the S-DWH, satisfying 2 essential needs:

a. to *guide* statisticians in processing and controlling the statistical production
b. to *inform* end users by giving them insight in the exact meaning of statistical data

In order to meet these 2 essential functions, the statistical metadata must be:

correct and reliable (the metadata must give a correct picture of the statistical data),
consistent and coherent (the metadata driving the statistical processes and the reporting metadata presented to the end users must be compatible with each other),
standardised and coordinated (the data of different statistics are described and documented in the same standardised way).

Since the different users of the (meta)data have diverse needs, it is essential to ensure an effective management of the statistical metadata in the S-DWH. To realise this, the use of a metadata model is a key element in structuring and standardising the statistical metadata within a NSI in a generic way. In the context of the S-DWH, a metadata model is a standardized representation used to define all necessary metadata elements of statistical information systems, based upon and using 1 or more standards/norms. In these implementations, standards act as checklists for controlling the completeness and correctness of all metadata elements as described by the model. At least 2 types of metadata models can be distinguished:

- a conceptual model that usually gives a high-level overview on how the metadata is organised, managed, maintained etc.: a description of the overall metadata process(es);
- a physical model that describes the details of the metadata objects and attributes, including relations between the metadata objects: a structured description of the metadata elements.

6. Conclusions

Of course, the design and implementation of a S-DWH as a huge impact on a NSI. It means developing new IT-systems, using new tools etc. asking for a high financial investment. It needs a complete redesign of the statistical production processes, moving from single operations to integrated generic statistical production. But in addition, it is also a major organisational operation, which is often underestimated. Not only systems need to change, specifically people must change. They have to learn and except new ways of working and stick to them consistently. Nowadays NSIs are confronted with a rapidly changing demand for information.

Next to a growing need for more information on more topics also the political lifecycle of policymakers is decreasing which means quicker delivery. To be able to meet this requirement ask for modernisation of statistics, in production, in data collection and the re-use of data. The concept of the Statistical Data Warehouse as data hub is one of the ways to meet this challenge. The S-DWH supports the total statistical chain from data collection to data dissemination by offering unambiguous insight in all available data sources.