Distr. GENERAL

8 October 2012

WP. 18
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Seminar on New Frontiers for Statistical Data Collection
(Geneva, Switzerland, 31 October-2 November 2012)

Topic (iii): Legal and institutional aspects of using new data sources

THE STATISTICAL DATA WAREHOUSE: A CENTRAL DATA HUB, INTEGRATING NEW DATA SOURCES AND STATISTICAL OUTPUT

Contributed Paper

Prepared by Harry Goossens (Coordinator ESSnet on Data warehousing, Statistics Netherlands)

I. Introduction

1. Within the European Statistical System (ESS) there is a growing need for modernisation of statistical data collection. Key words are decreasing costs and administrative burden, increasing efficiency and flexibility. Statistical institutes are investigating ways to disclose all kinds of new data sources that become available through the global use of modern technologies like internet, mobile phones, automated scanning techniques etc. At the same time, these technologies offer broad new possibilities to modernise data collection processes and in that way also collecting new data sources.

2. One of the challenges in this process of change is the integration of sources and collection modes and following to that, the standardisation of collection methods and technologies. Besides this, a second and probably even bigger challenge is the integration of the modernised statistical data collection into the statistical production:

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

3. 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.

4. Next to the influence on the data collection and –processing aspects, this modernisation also has an important organisational impact. Of course, first there is the need for a complete new way of organising the statistical data collection. But in addition, it also has higher and stricter demands for the data and metadata management. Often these two activities are decentralised and implemented in various ways, depending on the needs of specific statistical system., whereas realising maximum re-use of available statistical data demands just the opposite: a centralised and standardised, flexible and transparent metadata catalogue that gives insight and easy access to all available statistical data.

5. Building a statistical data warehouse (S-DWH) is considered to be a crucial instrument in the process of reaching this goal. The S-DWH approach can be used as a tool that helps NSIs to identify the particular phases and elements in their process of statistical production that must be common/ 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 new data sources and statistical output.

II. Background

A. The ESSnet project on data warehousing

6. Within the European Statistical System (ESS) the programme on the Modernisation of European Enterprise and Trade Statistics (MEETS) aims at *"the implementation of a more efficient way of collecting data"*. One of the main actions of MEETS 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.'

7. In this context, in October 2010 the "ESSnet on micro data linking and data warehousing in statistical production" is 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.**

8. 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. Therefore the activities in the first year concentrated on an inventory ('stocktaking') of the current situation in member states, with the following deliverables:

   (a) a detailed overview of current best practices in the use of integrated data warehouse systems within the ESS;

   (b) a prioritised list of problems and desired solutions as indicated by the Member States;

   (c) a conceptual model of the statistical data warehouse.

9. Main conclusions of an ESS-wide questionnaire was that on the one hand 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.

10. In this paper the work of the 'ESSnet on micro data linking and data warehousing in statistical production' is presented All results of the work in the first year can be found on the projects website:

    http://www.essnet-portal.eu/project-information/data-warehouse/data-warehouse-sga1

    The scope and status of the ongoing activities are also on the projects website:

    http://www.essnet-portal.eu/project-information/data-warehouse/data-warehouse-sga2

B. Defining the S-DWH

11. The broad definition of a statistical data warehouse (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.'

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

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

13. Based upon this definition a first draft version of the conceptual model of the statistical warehouse was defined, setting the scope/boundaries of the statistical warehouse and representing all various stages and elements between input and output. This model consists of two idealised representations of perspectives on statistical processes, which the ESSnet called the "Data Model" and "Process Model" . This model was used as basis for an ESS-wide questionnaire.

14. Main conclusions of was that on the one hand 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. But on the other hand, it also learned that a straighter and more distinctive definition of a statistical data warehouse was needed as the conceptual model of the statistical data warehouse was not distinctive enough. Therefore the ESSnet made an explanation of the S-DWH[1]:

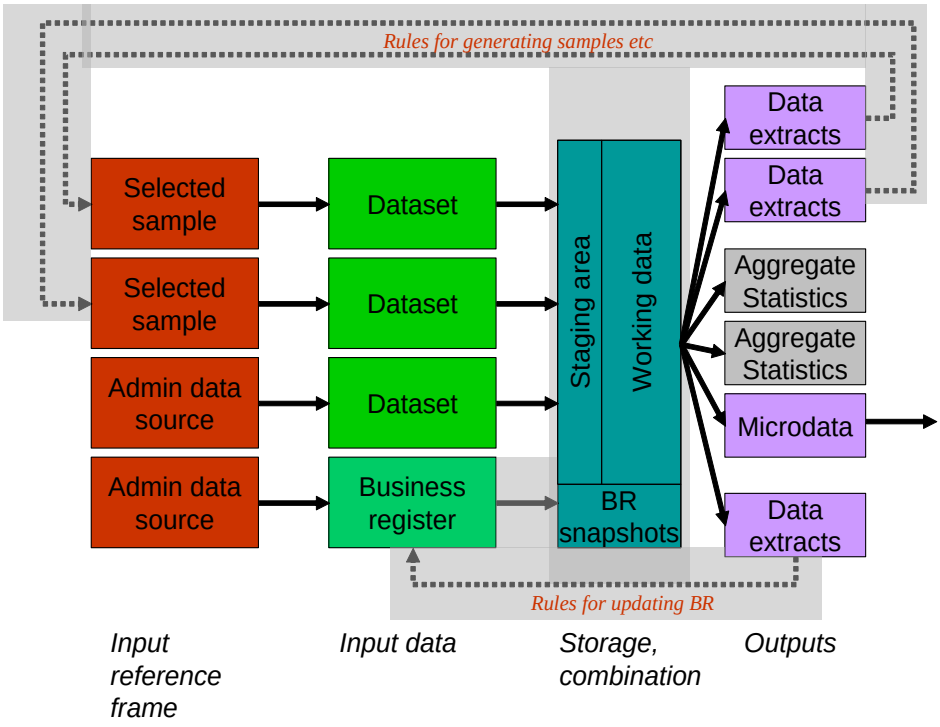

Figure 1: Explaining the S-DWH 1

15. The shaded areas on the above diagram can be considered the 'statistical data warehouse' comprising:

[1] http://www.essnet-portal.eu/data-warehouse/data-warehouse-sga1/final-report/annex-09-explaining-s-dwh

(a) technical facilities for storing and processing data, receiving data in and producing outputs in a flexible way

(b) rules for updating the sources for the DWH

(c) rules for generating samples

(d) definitions necessary to achieve those sample/source generation

(e) the data flow model

16. In daily statistical practice, the statistical data warehouse 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 not only support statistical production processes but also data collection processes by providing:

(a) a detailed and correct overview/insight of already available data sources;

(b) a framework for adequate data governance, including metadata management, confidentiality aspects and data authorisation;

(c) access to registers sampling frames (BR, etc.);

(d) flexible data storage and data exchange between processes.

III. The Business Architecture of the S-DWH

A. The layered architecture

17. In essence, the data warehouse is a concept that intends to provide an architectural model for the dataflow from operational systems to decision support systems; in our specific case, the S-DWH, from data collection systems to statistical output systems. In this context, the architecture is the conceptualisation of 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 to *enable* people, involved on 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.

18. Goal is reduction of redundancy en replication of activities. In the context of statistics, this comes as an answer to the need "avoid silo-ed solution systems"[2] in order to optimize and ensure organization consistency. It follows that the S-DWH must be able to interface with tools usually used by the NSI for their statistical production activities for the phases of estimation, confidentiality, sampling, diffusion and etc. The generic elements that must be considered in order to design a S-DWH are the data sources, the management instruments, the effective data warehouse in terms of micro and macro data, as well as the meta data and the different types and number of possible users.

19. In order to implement a S-DWH the first step is the conceptualisation of an architectural model of the data flow from the sources, surveys or administrative archives, through processing till statistical outputs environments.

[2] See the definition of stovepipe production - the European parliament and the council on the production method of EU statistics: in a vision for the next decade (Brussels, 10.8.2009 COM(2009) 404)

20. To provide such a model for the generic S-DWH as detailed as possible, in the context of statistics production, we identified four functional layers, each for specific statistic activities[3]:

(a) access layer

(b) interpretation and data analysis layer

(c) integration layer

(d) source layer

(e) and used them to define a layered Business Architecture for the S-DWH, representing the various statistical data used by each layer (figure 2). It is a bottom up representation, in which the ground level corresponds to the area where the process starts while the top of the architectural pile is where the DWH process ends.

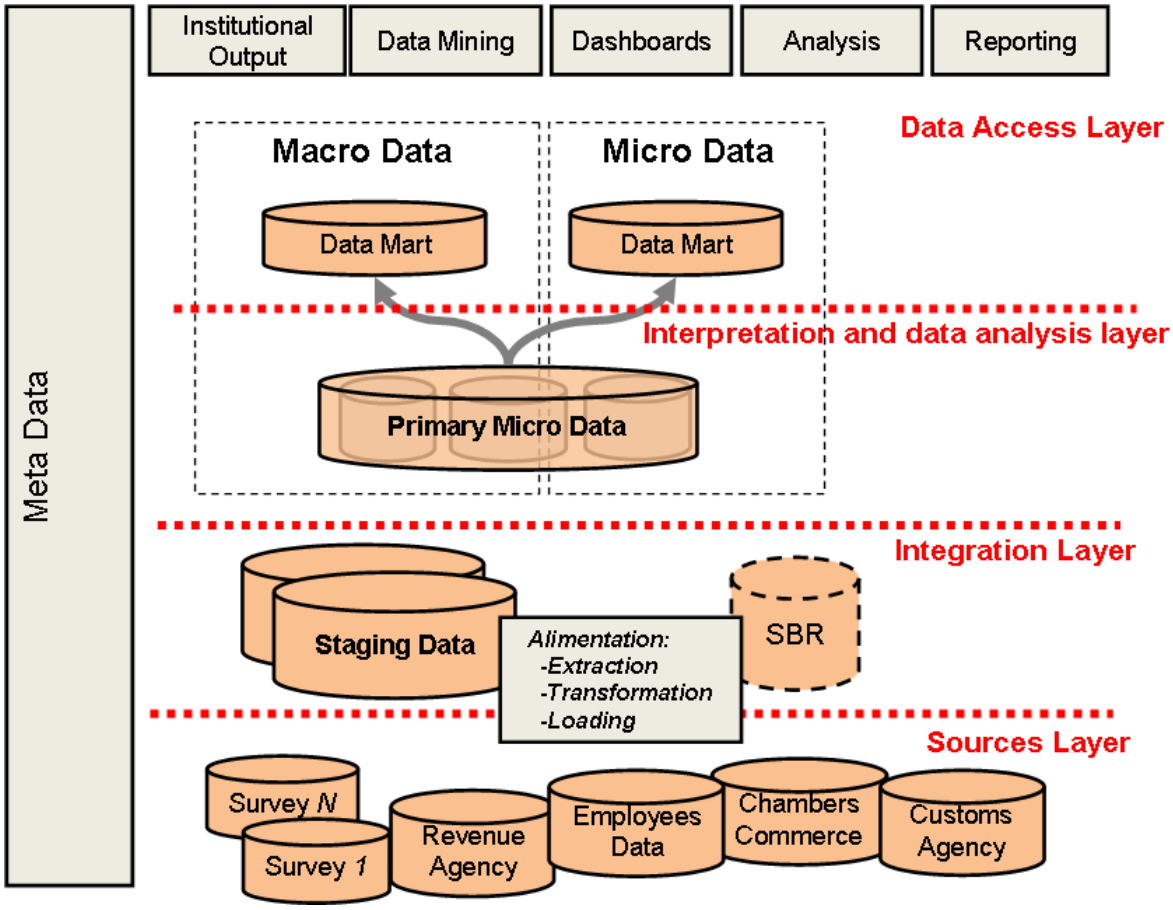

Figure 2: The layered architecture of the S-DWH with focus on "statistical data" used by each layers.

21. 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:

(a) 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 statistical purpose. The source layer is the interface towards all external actors participating in data collection. Generally, all data must be carefully checked before any promoting to the integration layer.

[3] For examples, see *"The Data Warehouse Toolkit"* - R. Kimball and M. Ross

(b) 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.

(c) 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. This is accomplished through structures defined as data marts, regarded as subsets of the DW, usually oriented to a specific business line or team.

(d) The Access layer is the layer for the final presentation, dissemination and delivery of the information sought. This layer is open for a wide range of users and using various exploration tools. In this layer the data organization must support automatic dissemination systems and free analysis, in both cases, statistical information is macro data.

B. Mapping the S-DWH on the GSBPM

22. 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].

23. For getting good understanding of the influence of a S-DWH approach for statistical production we analysed the architecture of five EU-regulated business statistics (SBS, STS, ProdCom, Trade statistics and Business Register) as managed by the partners of this ESSnet. 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 (figure 3). The GSBPM phases are on horizontal axe on different columns and the SDWH layers are on vertical axe on different rows. Their cross produce a cell-matrix which represent a potential position of a GSBPM sub-processes in the SDWH layer. The pink shaded areas are showing the classical 'generic' mapping of the sequential GSPBM sub-processes on the layered S-DWH.

24. A presentation of the S-DWH architecture is available on the ESSnet portal.[5]
The final version of the 'S-DWH Business Architecture' will be available by October 2012.

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

[5] http://www.essnet-portal.eu/data-warehouse/data-warehouse-sga2/workshop-3-essnet-data-warehousing-cardiff-uk-23-and-24-may-2012

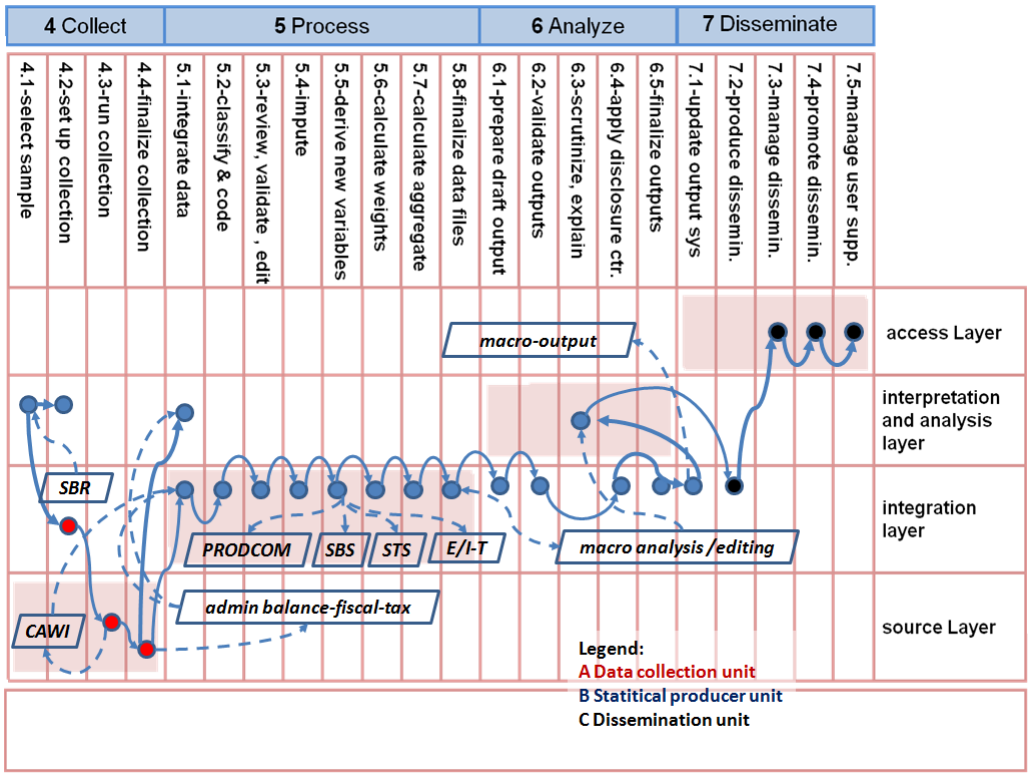

Figure 3: GSBPM mapping on the S-DWH

To effectively include a possible sub-process in this cell, we will fill the cell with a ball and connect subsequent ball with arrows in an common work flow. In which the arrows describe the versus of the work flow. To identify the actor responsible of each sub-process we fill each circle with different colour and associate each actors to a colours, their association will be described in separated legend. Rhombus, like in the BPMN, is representing data objects, i.e. shows the reader which data is required or produced in each activity. The position of each rhombus is relevant only in relation to SDWH information architecture. The rhombus must be positioned each time in one of four possible layers. To describe the process of population of a data object is used a dotted line with the arrow from the process toward the rhomb. Otherwise, everywhere the process use data as input the arrow is from the data object to the sub-process.

IV. Metadata in the S-DWH

A. The metadata Framework

25. 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.

26. Although most authors of data warehousing literature agree on the important role of metadata, you can find surprisingly little practical support on how to implement a metadata layer. An article by Panos Vassiliadis[7] of the University of Ioannina, Greece, summarizes well the requirements of data warehouse metadata. They should include information on:

(a) the contents of the data warehouse, their location and their structure

(b) the processes that take place in the data warehouse

[6] Kimball, The Data Warehouse Lifecycle Toolkit (Second Edition), Wiley, 2008, p. 117
[7] Data Warehouse Metadata, Encyclopedia of Database Systems, Springer, 2009

(c) the implicit semantics of data along with any other kind of data that aids the end-user exploit the information of the warehouse

(d) the infrastructure and physical characteristics of components and the sources of the data warehouse

(e) security, authentication, and usage statistics that aids the administrator tune the operation of the data warehouse as appropriate.

27. In order to better identify the role of the metadata in a S-DWH the ESSnet defined a (first version) of the metadata framework of the S-DWH.[8] 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:

(a) 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.

(b) 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.

(c) The SDWH requires *structural* metadata, especially *technical* metadata. Active metadata must be structural, at least to some part.

(d) *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.

28. The table below 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.

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

29. A crucial metadata requirement that is desirable in any environment, but is necessary in the S-DWH, is that all metadata must be consistent throughout the data warehouse. All metadata items (concepts as well as physical references) must be uniquely identified and there must be one-to-one relationships between identity and definition, and identity and name.

B. Metadata and the layered S-DWH

30. After defining metadata in the context of the S-DWH, the next step is to link it to the statistical production lifecycle: what metadata are produced during a process, what metadata are needed to perform a process, and what metadata are forwarded from one process to the next one.

[8] http://www.essnet-portal.eu/data-warehouse/data-warehouse-sga2/21-wp1-metadata/deliverables (see related documents)

31. In the layered architecture (Figure 2), the metadata layer at the left-hand side indicates the necessity of metadata support from start to finish. The table below gives a rough overview of where in the S-DWH layers three important metadata categories are created (indicated by *c*) and used (*u*).

Layer	Statistical metadata	Process metadata	Quality metadata
Data access	u	cu	u
Interpretation	cu	cu	cu
Integration	cu	cu	c
Source	c	c	c

32. The table shows that the lower layers mainly provide metadata, but can't make much use of them, while in the higher layers metadata are used, but relatively few are added. This very much agrees with the rule that metadata should be captured as close to the source, or as early in the process as possible.

33. The S-DWH architecture should make it possible to trace any changes made to data as well as metadata by using process metadata and versioning both data and metadata. Thus, a metadata item is normally never changed, updated or replaced. Instead, a new version is created when necessary, which means that there will always be a possibility to identify which metadata was considered correct at a certain point in time even if it has later been revised. A more detailed analysis on the metadata subsets and their use in the S-DWH layers, and also in relation to the GSPBM processes will be carried out in the upcoming work of this project.

V. Managing the S-DWH

A. (Meta) data governance

34. 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. 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. From the requirements on data follow similar demands on metadata: 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.

35. 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:

(a) **correct and reliable** (the metadata must give a correct picture of the statistical data),

(b) **consistent and coherent** (the metadata driving the statistical processes and the reporting metadata presented to the end users must be compatible with each other),

(c) **standardised and coordinated** (the data of different statistics are described and documented in the same standardised way).

36. 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. The metadata framework (see IV – a) defines a metadata model as follows:

[Def 3.6.1] *A metadata model is a special case of a data model: an abstract documentation of the structure of metadata used by business processes.*

37. 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) a conceptual model that usually gives a high-level overview on how the metadata is organised, managed, maintained etc.

(b) a physical model that describes the details of the metadata objects and attributes, including relations between the metadata objects.

Simpler, you could say that a conceptual metadata model is a description of the overall metadata process(es), where the physical model is a structured description of the metadata elements.

38. In the context of the term (metadata)**model** also the term **standard** needs to be reconsidered, as they are often used in relation or even mixed. The following general definition of a model is commonly accepted:

'A model is a simplified description of an analogue part of the reality.'

For the term standard, often also **norm** is used as a synonym. The following general definition of standard/norm is commonly accepted:

'A standard or norm is a document with recognized agreements, specifications or criteria about a product, service or method.'

Looking at the coherence of and/or the differences between both terms a standard/norm generally defines WHAT to be done, a model describes HOW to do it.

B. Organisational aspects

39. 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 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, also people must change. They have to learn and except new ways of working and stick to them consistently.

40. In this context, metadata management is a subject that is often neglected. The questionnaire showed some remarkable results. On the one hand, almost all NSIs mention that metadata are "important" or "extremely important" in statistical production. But on the other hand, 80% of the NSIs admit that metadata are currently implemented in only a few systems, not managed on a central level. Next to complexity on of the main reasons is that metadata management primarily is felt as a burden, as extra work with no direct gain. It needs the awareness that good and reliable

information about the available statistical data sources is the key to new sources and the enabler of more efficient statistical production and analysis.

41. In modern society 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, as well in production but certainly also in data collection. The challenge is the statistical disclosure of all new data sources that become available through the global use of modern technologies.

42. 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.