SDMX self-learning package No. 2

Student book

The SDMX Information Model

Produced by	Eurostat, Directorate B: Statistical Methodologies and Tools Unit B-5: Statistical Information Technologies
Last update of content	November 2010
Version	1.0

TABLE OF CONTENTS

1 Scope of the student book

The student book aims at providing a general introduction to users interested in:

- The SDMX Information Model for data;
- The SDMX Information Model for metadata;
- The main objects of the SDMX Information Model;

At the end of the student book, the reader should be able to understand the basics of the SDMX Information Model.

This student book is the second one of a set of student books (see Table 1) which altogether provides the complete set of information to master SDMX, with a particular focus on the data model.

Ref.	Title
[01]	Introduction to SDMX
[02]	SDMX Information Model
[03]	SDMX-ML Messages
[04]	Data Structure Definitions
[05]	Metadata Structure Definitions
[06]	XML based technologies used in SDMX
[07]	SDMX architecture using the pull method for data sharing – Part 1
[08]	SDMX architecture using the pull method for data sharing – Part 2

Table 1 – Students books on SDMX

Prerequisites

This student book may require as prerequisite the reading of the first student book.

2 Statistical Data and Structural and Reference Metadata

2.1 Statistical data

Statistical 'data' are sets of often numeric observations which typically have time associated with them. They are associated with a set of metadata values, representing specific Concepts, which act as identifiers and descriptors of the data. These metadata values and Concepts can be understood as the named Dimensions of a multi-dimensional co-ordinate system, describing what is often called a 'cube' of data (Figure 1).

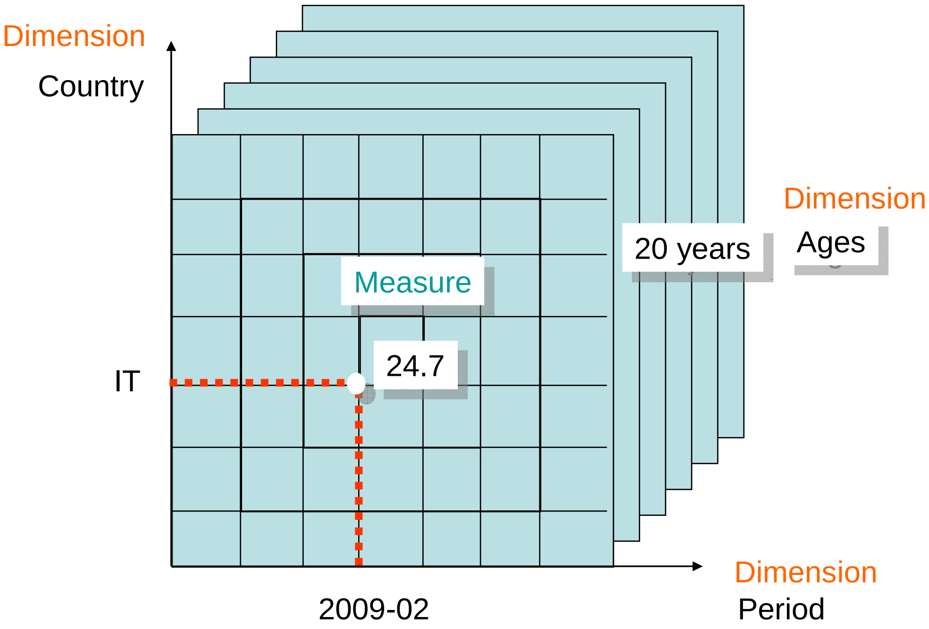


Figure 1 - Multidimensional 'Cube' of data

Different data organisations exist to present statistical data. Tabled data can be formatted as a Multidimensional table, in a Time series format or in a Cross-sectional data format.

Time Series data organisation example (Table 2) presents the statistical data according to yearly observation periods. For this table it is looked at only one geographical (GEO) entity (FR = France) for the development in the number of tourism establishment types over a specified period.

Cross-sectional data organisation is used for the exchange of data with more than one observation type in a Data set (Table 3). This means it is intended for situations where the statistical data consist of multiple observations (A100 - Hotels and similar, B010 - Tourist Campsites and B020 - Holiday dwellings) at a single point in time.

An example of a multidimensional statistical table is presented in Figure 3 (on page 8). In this example two Dimension layers are presented in the columns of the table (Activity and Time).

2.1.1 Data in Time series Format

Number of touristic establishments – Time series

FREQ: A – Annual
GEO: FR – France
TOUR_INDICATOR: A001 – Establishments
UNIT: NBR - Number

Activity	A100	B010	B020
Time	Hotels and similar	Tourist Campsites	Holiday dwellings
2002A00	18768	8354	1934
2003A00	18617	8331	1968
2004A00	18598	8289	2251
2005A00	18689	8174	2329

Table 2 - Data in Time series format

2.1.2 Data in Cross-Sectional Format

Number of touristic establishments – Cross-sectional

TIME: 2007A00
TOUR_INDICATOR: A001 - Establishments
UNIT: NBR - Number

Activity	A100	B010	B020
Country	Hotels and similar	Tourist Campsites	Holiday dwellings
AT	14204	540	3388
ES	17827	1220	4843
FR	18135	8052	2406
IT	34058	2587	61810

Table 3 - Data in Cross-sectional format

2.2 Statistical Metadata

The term «metadata» is very broad indeed. A distinction can be made between ‘structural’ metadata – those Concepts used in the description and identification of statistical data and metadata – and ‘reference’ metadata (additional explanatory metadata, for example on the methodology used or quality aspects). The following paragraphs will provide a deeper explanation of these two types - structural & reference metadata - used to express the Data and Metadata Structures and to understand Data and Metadata sets in relation with the SDMX Information Model.

2.2.1 Structural Metadata

Statistical Data sets are described by a set of metadata values, taken from specific Concepts. The Concepts act as identifiers and descriptors of the data. This system of Concepts identifying and describing the data can be recognized as the named Dimensions of a 'multi-dimensional cube' of data.

Structural metadata is arranged in Structure Definitions. A 'Data Structure Definition (DSD)' / 'Metadata Structure Definition (MSD)' describes how Data sets / Metadata sets are organized and defines the mechanism for referencing those Data / Metadata sets, which are described by the structural metadata.

A Concept used in a Data Structure Definition is given a 'usage role' of Dimension, Attribute, and Measure in that definition. The Data Structure Definition, being the modelled structure of the data 'cube', can also include special Concepts, for example the Measure Dimension to represent the multiple Measures of a cross-sectional data organisation.

When Concepts take their value from a set of known values (codes), objects called 'code lists' can be linked as the representation of the Concepts or, more usual, assigned in the DSD ('Key Family') to the related Concept.

2.2.2 Reference Metadata

The Metadata sets mentioned above are related to the SDMX model for additional explanatory metadata, often referred to in SDMX as reference metadata. Reference metadata are generally in a textual format, using Concepts describing the content, methodology and quality of the data, thus it can be broken down into:

- Conceptual metadata, describing the Concepts used and their practical implementation;
- Methodological metadata, describing methods used for the generation of the data;
- Quality metadata, describing the different quality aspects of the statistical data.

These are mostly metadata which are reported not as an integral part of the statistical Data set. It concerns for example metadata related to entire collections of data. Reference metadata is content metadata that gives more information about statistical data, so as to make its interpretation more meaningful.

The reference metadata is structured according to a 'Metadata Structure Definition' (MSD). A Metadata Structure Definition describes how Metadata sets, containing reference metadata are organized and defines the mechanism for referencing the statistical data or structural metadata to which this reference metadata relates.

Eurostat has defined a Metadata Structure called the Euro-SDMX Metadata Structure (ESMS). It contains the description and representation of statistical metadata Concepts to be used for documenting statistical data and for providing summary information useful for assessing data quality and the production process in general.

3 Brief introduction to the SDMX Information Model

The SDMX Information Model provides a broad set of formal objects and their relationships to represent statistical data and metadata, actors, processes, and resources within statistical exchanges. Besides the objects described in detail in the student book, the SDMX Information Model includes detailed technical model definitions and further components, which are briefly summarized below:

Structure Set and **Mappings:** Structure sets allow grouping structural metadata together to form a complete description of the relationships between specific, related sets of data and metadata. The Structure sets can be used to map dimensions and attributes to one another, to map concepts, to map codelists and to map category schemes.

Reporting Taxonomy: A reporting taxonomy allows an organisation to link (possibly in a hierarchical way) a number of data cubes or data flow definitions, which together form a complete ‘report’ of data or metadata.

Process and **Transitions:** In any system that processes data and metadata the system itself is a series of processes and in each of these processes the data or metadata may undergo a series of transitions. The process model is a generic model that can capture key information about these stages in both a textual way and also in a more formalised way by use of expressions.

Transformation and **Expressions:** In the model, this package allows to track.how data is acquired or derived. The model can be used equally to define data transformations /derivations and to define integrity checks with the help of expressions (e.g. the Sum of A+B must equal C).

The Figure 2 below depicts the essential characteristics supported in the SDMX Information Model for data and metadata reporting.

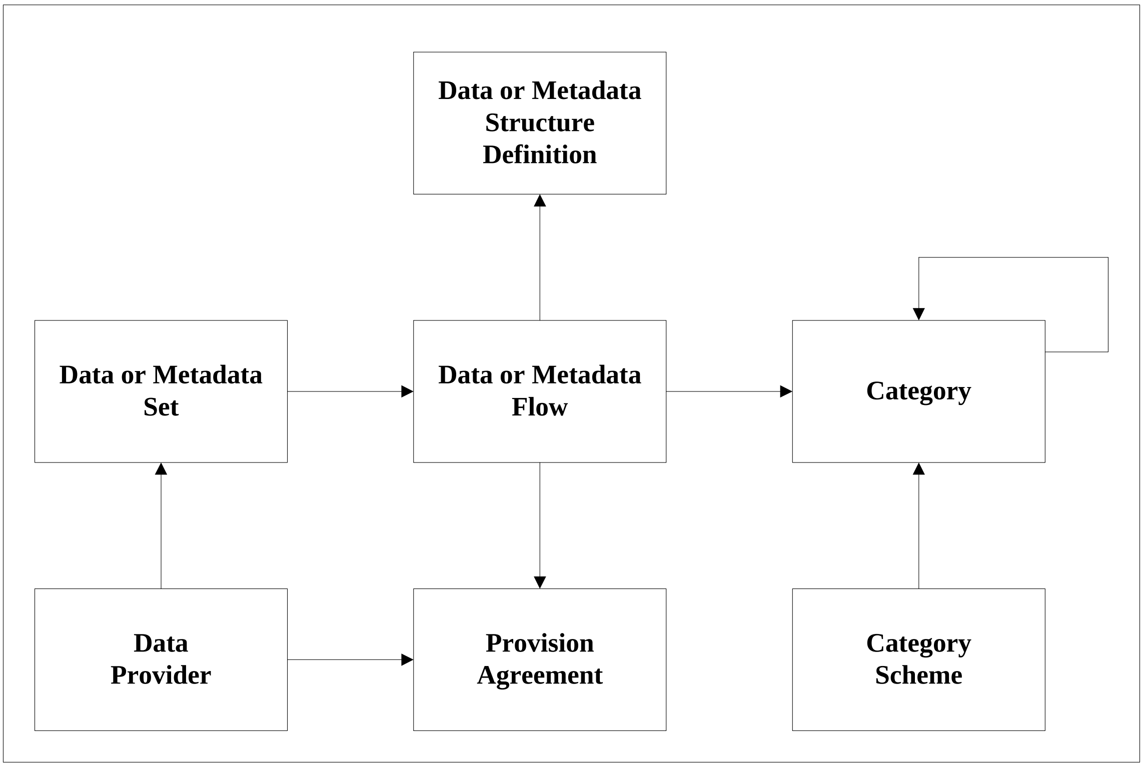


Figure 2 - Main Elements of the SDMX Information Model

The pivot of this diagram is the Data or Metadata Flow, maintained by the organisation that collects data or metadata. A Data Flow is linked to a 'Data Structure Definition' (DSD) while a Metadata flow is linked to a 'Metadata Structure Definition' (MSD). DSD or MSD define the structure of data or metadata and identifies the elements to which the related metadata can be attached.

Data or metadata may be made available by many providers. Any provider may report or publish data or metadata for several Data or Metadata flows, according to a Provision Agreement.

The Data or Metadata Flow may also be linked to one or more statistical topics (Category) in a subject-matter scheme (Category Scheme). A Category Scheme provides a way of classifying data for collection, reporting or publication.

Example:

- For Short-Term Business Statistics (STS) within "SDMX Open Data Interchange (SODI) project" four Categories are defined:
  - SSTSCONS – SODI – Short-term Statistics on Construction;
  - SSTSIND – SODI – Short-term Statistics on Industry;
  - SSTSRTD – SODI – Short-term Statistics on Retail Trade;
  - SSTSSERV – SODI – Short-term Statistics on Services.
- Each of these is linked to one or more Dataflows. In the case of SSTSCONS, for example the following Dataflows are defined for which data is collected:
  - SSTSCONS_PROD_M (Production in construction – total, building construction, civil engineering – monthly);
  - SSTSCONS_PROD_Q (Production in construction – total, building contraction, civil engineering – quarterly);
- Data flows like SSTSIND_PROD_M are linked to a unique Data Structure Definition (e.g. ESTAT+STS+2.1.xml)
- A Provision Agreement represents the frame or the arrangement between a data provider and recipient for a data set or Metadata set to be exchanged (for example SSTSCONS_PROD_M_DE1.xml – Data Provider: Statistisches Bundesamt, Germany; Recipient: Eurostat)

The SDMX standards offer a common model and multiple data formats which support the exchange of any type of data 'cube'. To allow automated processing of the data from a variety of sources, the structure of this 'cube' needs to be defined in a way that conforms to the SDMX Information Model.

In the SMDX Information Model, structural metadata are represented by a list of Concepts organised into lists called 'Concept Schemes'. Concepts are the basic building blocks in SDMX: the Concepts exist and are maintained separate from any structure that uses them.

The SDMX Information Model provides objects for the structuring not only of data, but also of ‘reference’ metadata.

The SDMX Information Model allows the exchange and storage of reference metadata independently from the data that it is describing. The reference metadata can be kept in a so called metadata repository. Consequently, a dissemination system can disseminate data with the related reference metadata, which is extracted on request from the reference metadata repository.

Reference Metadata can also be indexed to support search facilities. For this, a metadata report could for example be processed by a registry service to extract its structural information. This structural information allows to catalogue the metadata and a user may query for it.

The process of providing and managing the flow of Data or Metadata sets is also covered by the SDMX Information Model with associated metadata concerned with ‘data provisioning’. This metadata is useful to those who need to understand the content and form of a data provider’s output. Each data provider can describe in standard fashion the content of and dependencies within the data and Metadata sets which they produce, and supply information about the scheduling and mechanism by which their data and metadata is provided. This allows for automation of some validation and control functions, as well as supporting management of data reporting.

To organize and manage the exchange and dissemination of data and metadata, SDMX Information Model includes also information about classification schemes and domain Categories, along with their relationships to data and Metadata sets (please compare section 4.12).

4 The main SDMX objects

4.1 Introduction

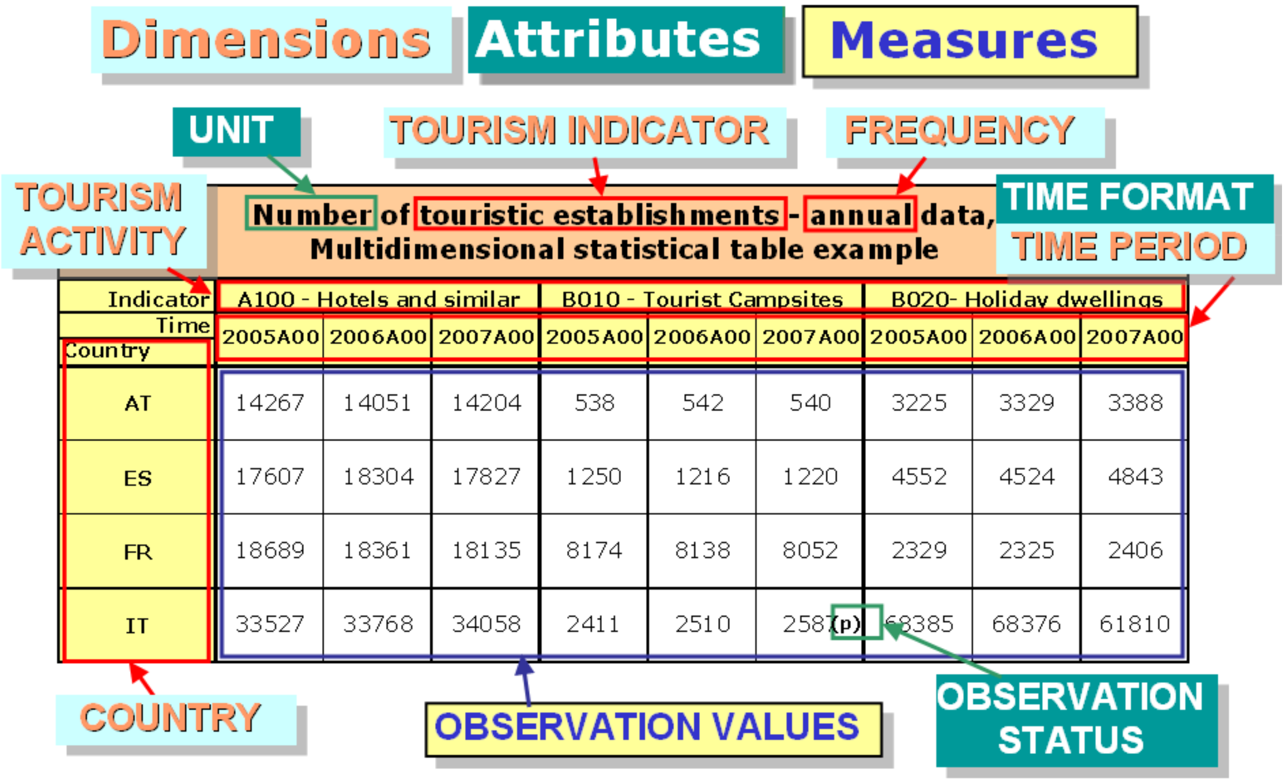
The SDMX Information Model (SDMX-IM) describes a set of formal objects so as to present a standard view of the statistical exchange process.

It is based on a set of objects that conceptualizes the real world in the framework of statistical data/metadata sharing and exchanges. The following paragraphs intend to clarify the main objects required to understand the basic of the SDMX Information Model.

4.2 Concept and Concept Scheme

Concepts play an important role in the SDMX Information Model as they are used to describe the structure of a multidimensional statistical table (Figure 3 - Multidimensional statistical table with its Concepts) or the structure of a metadata report. In the SDMX Information Model, Concepts can have a specific value representation (coded value, numeric format, date format, string, etc.) that can be defined in the Concept Scheme.

In the example below, Concepts identify the different elements of a TOURISM multidimensional statistical table. The Table 4 shows the list of Concepts used as well as their coded representation.



Number of touristic establishments - annual data, Multidimensional statistical table example

Indicator	A100 - Hotels and similar			B010 - Tourist Campsites			B020- Holiday dwellings		
Time	2005A00	2006A00	2007A00	2005A00	2006A00	2007A00	2005A00	2006A00	2007A00
Country									
AT	14267	14051	14204	538	542	540	3225	3329	3388
ES	17607	18304	17827	1250	1216	1220	4552	4524	4843
FR	18689	18361	18135	8174	8138	8052	2329	2325	2406
IT	33527	33768	34058	2411	2510	2587(p)	68385	68376	61810

Figure 3 - Multidimensional statistical table with its Concepts

Concept structure for the Multidimensional table example						
Key	**Concept ID**	**Concept Name**	**Attachment level**	**Usage status**	**Code List ID**	**Code List Name**
DIMENSIONS						
1	FREQ	Frequency		Mandatory	CL_FREQ	Frequency code list
2	COUNTRY	Tourism Country		Mandatory	CL_COUNTRY	Country code list
3	INDIC_TO	Tourism Indicator		Mandatory	CL_TOUR_INDICAT	Tourism Indicator code list
4	ACTIVITY_TO	Tourism Activity		Mandatory	CL_TOUR_ACTIVITY	Tourism Activity code list
	TIME_PERIOD	Time period		Mandatory		
MEASURES						
	OBS_VALUE	Observation value		Conditional		
ATTRIBUTES						
	OBS_STATUS	Status of the observation	Observation	Conditional	CL_OBS_STATUS	Observation status code list
	UNIT	Unit	Series	Mandatory	CL_UNIT	Unit code list
	TIME_FORMAT	Time format	Series	Mandatory	CL_TIME_FORMAT	Time format code list

Table 4 – Concept structure example of the above table

Concepts are identified in Concept Schemes by an ID and a name in at least one language, and optional multi-lingual descriptions of the Concepts can be added.

Example: A Concept describing the country of reference could be defined as follows:

Concept: Reference area		
ID	REF_AREA	
Name	(English)	Reference country code
	(French)	Code du pays de reference
Description	(English)	Country from which the population migrate
	(French)	Pays à partir duquel la population migre

Table 5 - Concept Reference area

A Concept Scheme is a SDMX object maintained by an agency and containing a list of Concepts from which Data and Metadata Structure Definitions are built. Many Concept Schemes can be created. Concept Schemes generally group Concepts relevant to a single structure although a Data/Metadata Structure Definition can use Concepts from different Concept Schemes.

4.3 Code lists

In the SDMX Information Model, a code list is an object containing a list of codes and maintained by an agency. A Code list is simply a set of values to be used in the representation of a Concept (Dimension or Attribute) in Data/Metadata Structure Definitions.

Example: CL_UNIT_MULT	
Code	**Description**
0	Units
1	Tens
2	Hundreds
3	Thousands
4	Tens of thousands
6	Millions
9	Billions

Table 6 - Example of the code list Unit multiplier

Each code is defined uniquely by a value and a description that can be provided in several languages.

The model allows a code list to have simple hierarchy of codes. In that case, the hierarchy is made by defining at maximum one parent code for child codes.

Example : CL_NUTS		
Code	**Description**	**Parent code**
BE2	VLAAMS GEWEST	
BE3	REGIONE WALLONE	
BE31	Prov. Brabant Wallon	BE3
BE32	Prov. Hainaut	BE3
BE321	Ath	BE32
BE322	Charleroi	BE32
BE323	Mons	BE32
BE324	Mouscron	BE32
BE33	Prov. Liege	BE3
BE34	Prov. Luxembourg (B)	BE3
BE35	Prov. Namur	BE3

Table 7 - Hierarchical Code list of Regions (NUTS)

4.4 Data Structure Definition (DSD)

The Data Structure Definition is the name for a set of explanations on how a Data set is built and how it shall be interpreted. This data description is built on statistical Concepts. Many organisations name this Data Structure Definition a ‘Key Family’ and so the two names are synonymous.

The Data Structure Definition, maintained by a maintenance agency (e.g. Eurostat), is a description of all the structural metadata needed to understand the structure of the Data set. In fact the Data Structure Definition links the statistical data to its structural metadata by assigning descriptor Concepts to the elements of the statistical data.

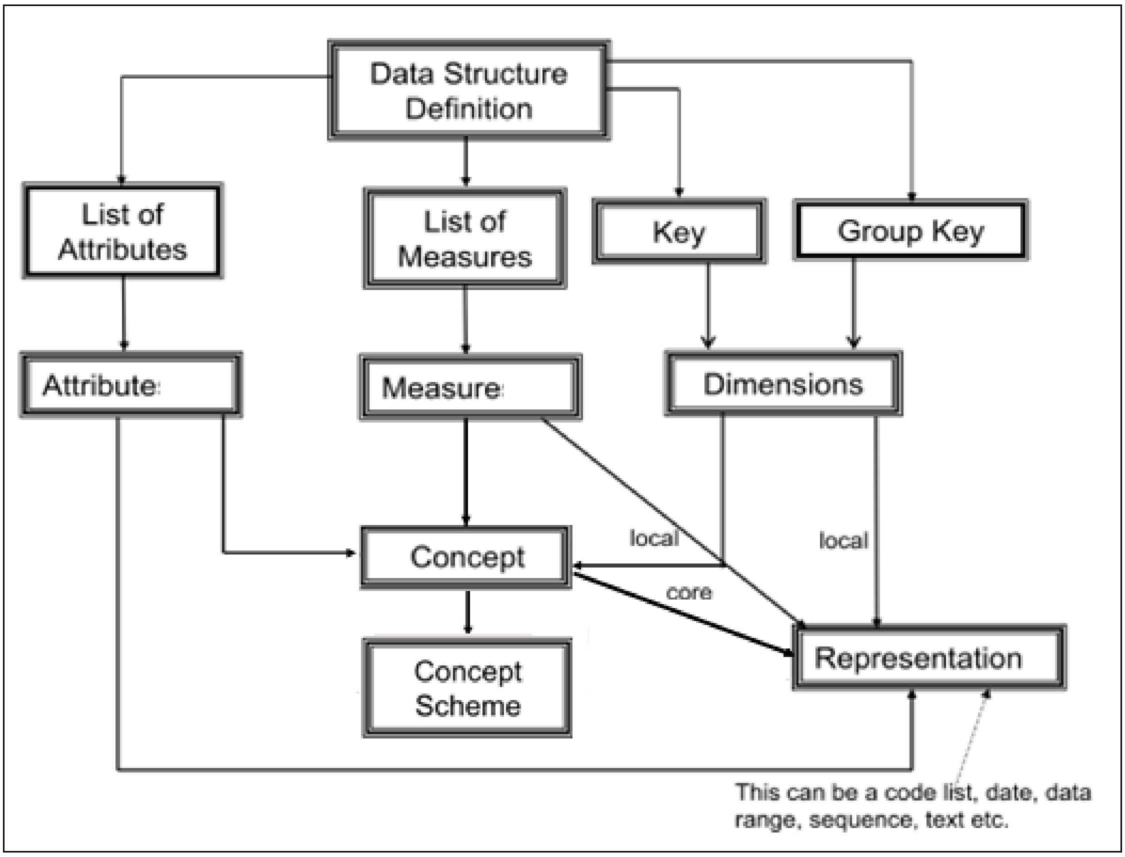


Figure 4 - Data Structure Definition

As shown in the figure, the Data Structure Definition is formed by three sets of Concepts:

- **Dimensions** comprising Non-Key Dimensions and the Key lists:
    - Key descriptor;
    - Group Key descriptor(s) - there may be more than one Group Key descriptor;
- **List of Measures;**
- **List of Attributes.**

The Concepts used for **Dimensions, Measures, and Attributes** can be taken from any maintained Concept Scheme, and need not all be from the same agency or scheme.

Dimensions are those Concepts, which describe the statistical data and form the identifier (Key) to the related data. For this, Dimensions are grouped into Keys, which allow the identification of a particular set of data (for example, a series).

The List of Measures comprises for time series data one Measure - the Primary Measure. This is conventionally associated with the OBS_VALUE Concept. It measures a statistical phenomenon over a time period. There can be only one Primary Measure declared in the DSD. This means that in SDMX time series there can be only one Measure Concept.

On the other hand, for the cross-sectional data organisation, a specific Measure Dimension can be declared containing the multiple cross-sectional Measures (see section 4.5.2).

The List of Attributes comprises one or more Attributes. Attributes are Concepts used to provide more information about some part of the Data set. Each Attribute in the Data Structure Definition must be assigned to an identified part of the Data set (in the model this is called the ‘attachment level’ or ‘grouping level’).

The Keys consist of Dimensions, whose combined values in a dataset uniquely identify observed data values (series or section). A particular observation value in a time series is identified with the Key and the Time for this particular value. The SDMX Information Model allows creating subsets of Keys, named Group Key. This subset of Key Dimensions form a Partial Key whose combined values identify a subset of the ‘cube’ to which Attributes are linked giving metadata about the identified object. Thus, the purpose of a Group Key descriptor is to define a subset of the full Key descriptor to which data Attributes can be attached.

Within a group, some descriptor Concepts have values that are the same for all Series within the group, while other descriptor Concepts are changeable. The rule is that descriptor Concepts are ‘attached’ to the grouping level where they become variable. Thus, if, within a single set of data, all the contents of a Series share a single value for a descriptor Concept, then that descriptor Concept should be attached at the Series level. This rule also assumes that the chosen level is the highest structural level where all sub-groups will share the same value.

Attachment levels of descriptor Concepts are always at least at the level where the Concept is meaningful: thus, you cannot attach the descriptor Concept frequency at the Observation level, because as a Concept it only operates at the level of Series (that is, with multiple Observations made over time).

The following example in time series data organisation on Short-Term Business Statistics (STS) industrial production illustrates the elements of its related Data Structure Definition. In this context of this example a special focus is given to the grouping issues. The below Figure 5 - Data Structure Scheme for STS – is extended by grey boxes containing the related STS Concepts (e.g. STS_Indicator) and the coded value representations for the coded Concepts (e.g. CL_STS_INDICATOR, the code list related to the STS_INDICATOR Concept).

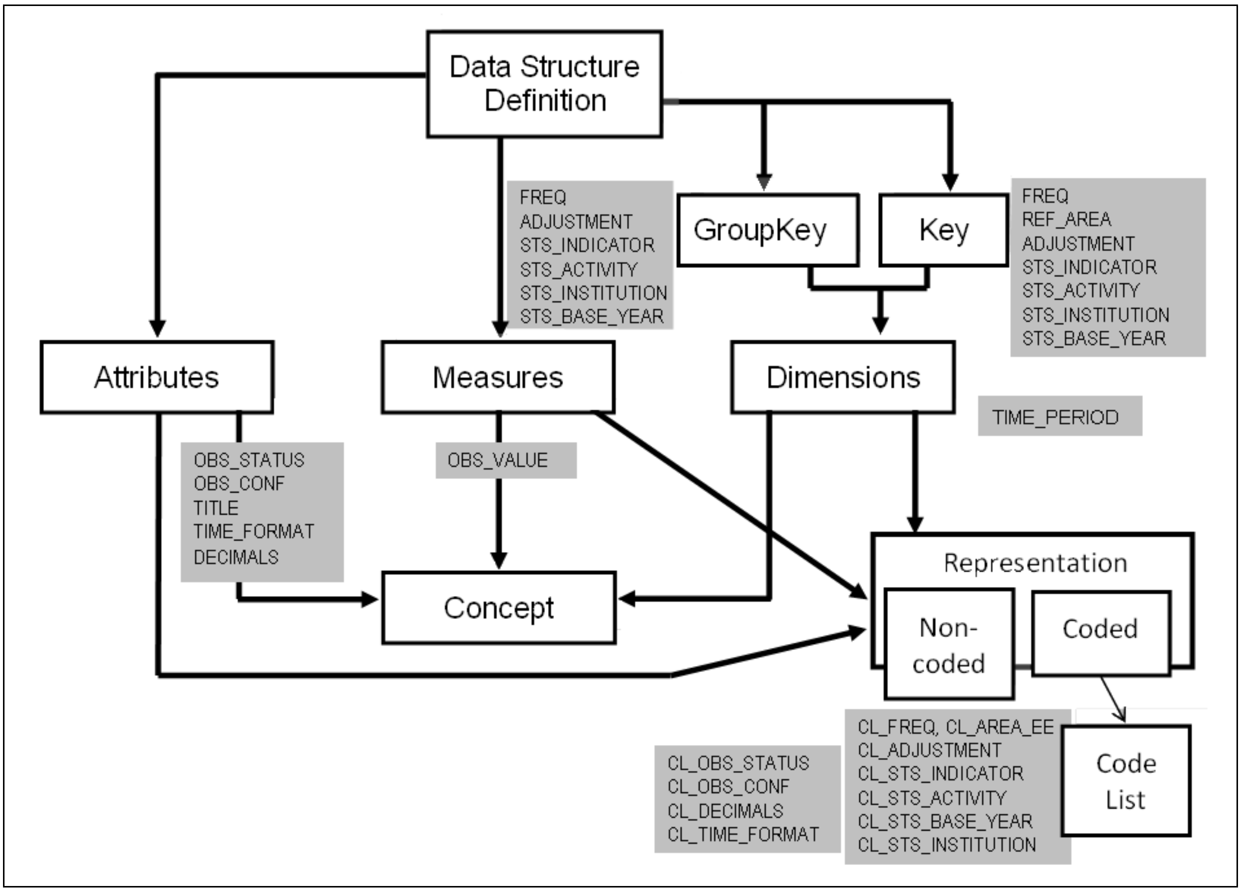


Figure 5 - Data Structure Scheme for STS

Time-series Data Structures, like the above STS example support the following 'attachment levels' or 'grouping levels':

LEVEL	IDENTIFICATION
Data set	Top level - the whole Data set;
Group Key	Identified by a value for each of the Dimensions forming the Group Key, for example a Group Key formed by FREQUENCY, ADJUSTMENT, STS_ACTIVITY, STS_INDICATOR, STS_INSTITUTION & STS_BASE_YEAR;
Series Key	Identified by a value for each of the Dimensions of the Key, for example FREQUENCY, ADJUSTMENT, REF_AREA, STS_ACTIVITY, STS_INDICATOR, STS_INSTITUTION & STS_BASE_YEAR;
Observation	Identified by a value for each of the Dimensions of the Key plus the time value), for example FREQUENCY, ADJUSTMENT, REF_AREA, STS_ACTIVITY, STS_INDICATOR, STS_INSTITUTION & STS_BASE_YEAR, and - in addition - the TIME_PERIOD.

Table 8 - Attachment levels

The most common group used to be the ‘SiblingGroup’ - in which all Dimensions are included, but the reporting frequency. This group was typically defined in the past, because most Attributes, which are not directly linked to the particular observation, will not vary in its values with the frequency. It would become very redundant to repeat the Attributes’ values for the series where only frequency is changing. Consequently those Attributes could be attached at the ‘SiblingGroup’ level.

Considering the above STS example a sibling group could be defined with the following combination of Key values:

- REF_AREA=”BE"
- ADJUSTMENT="N"
- STS_INDICATOR="TOTV"
- STS_ACTIVITY="NS0080"
- STS_INSTITUTION="1"
- STS_BASE_YEAR="2005"

The corresponding series are shown in Table 9 for two different frequencies (Monthly and Quarterly):

Total Industrial Turnover Index
Activity = NS0080
(Only for illustration - no real data)

Monthly	Index	Quarterly	Index
2008M01	95.19		
2008M02	87.13		
2008M03	97.09	2008Q01	94.24
2008M04	111.45		
2008M05	121.23		
2008M06	137.76	2008Q02	122.79
2008M07	108.19		
2008M08	112.15		
2008M09	103.09	2008Q03	107.24
2008M10	99.65		
2008M11	101.23		
2008M12	97.76	2008Q04	98.29

Table 9 - Sibling group tabled data

When grouping, there is no requirement to have only a single Dimension omitted from a Partial Key - it can be any subset of the set of ordered Dimensions (that is, all Dimensions except the time Dimension, which may never be declared as belonging to a Group / Partial Key). All groups declared in the Data Structure Definition must be unique - that is, you may not have duplicate Partial Keys. In SDMX data can be grouped to serve as useful package of information. One can imagine for STS data a grouping for the STS_ACTIVITY or Reference Area (REF_AREA) besides other Key Dimensions.
Below the example of a ‘REF_AREA Group’ is illustrated. This group is declared, because no variation exists for the Attributes for different countries. All Dimensions except the

REF_AREA Dimension forming the Partial Key (or Group Key). Thus, the Dimension REF_AREA, which is not part of a group, has a value which varies at the series level (for the STS time series data organisation).

In the example two sets of Partial Key values (Group Keys) are considered to organise the data (Group 1 and Group 2). The groups differ only in the Keys by the **Activity** value. In each group, we envisage two set of Keys identifying the time-series (Series 1 and Series 2) whose Key values differ only for the reference area. Once a grouping is defined, the relevant Attributes can be attached to the groups (e.g. the group-level Attributes in the example: Title and Decimals).

Example 'REF_AREA Group':

STS Data set represented as a time-series in CSV format:

```
M;BE;N;TOTV;NS0080;1;2005;2008-04;95.19;E;F
M;BG;N;TOTV;NS0080;1;2005;2008-04;138.05;A;F
M;BE;N;TOTV;NS0080;1;2005;2008-05;87.13;E;F
M;BG;N;TOTV;NS0080;1;2005;2008-05;137.76;A;F
M;BE;N;TOTV;NS0060;1;2005;2008-04;101.24;E;F
M;BG;N;TOTV;NS0060;1;2005;2008-04;165.59;A;F
M;BE;N;TOTV;NS0060;1;2005;2008-05;86.59;E;F
M;BG;N;TOTV;NS0060;1;2005;2008-05;168.55;A;F
```

STS Data Structure: **Dimensions** are *frequency; reference area; adjustment; indicator; activity; institution; base_year; time_period.* **Attributes** are *obs_status; confidentiality; decimals & title*

Please note that the Group Key needs to be shown in the dependent Series in order to establish the link, since the linked Series cannot be embedded in the Groups for SDMX-ML Compact data.

Group 1: *Frequency = M; adjustment ="N"; indicator = TOTV; activity = NS0080; institution = 1; base_year = 2005; Title = "Total Industrial Turnover Index for non-durable consumer goods"; Decimals = 2.*

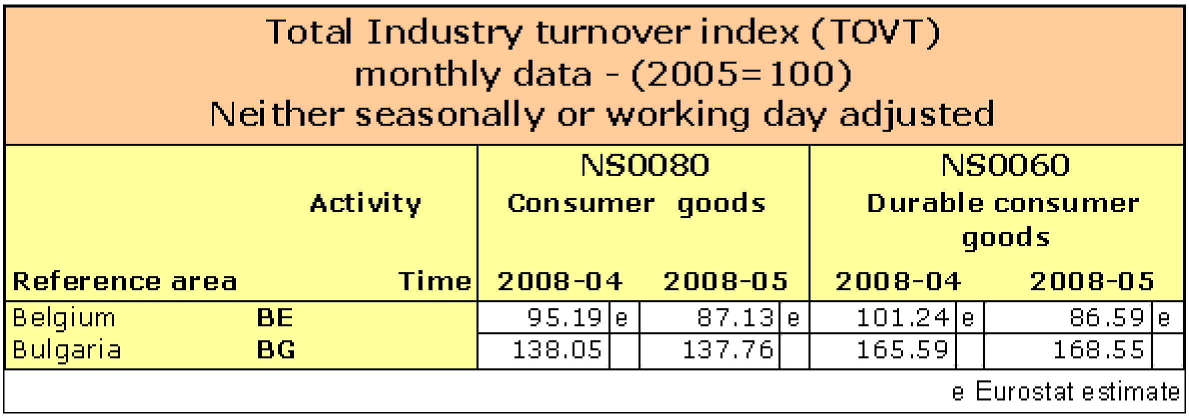
Group 2: *Frequency = M; adjustment ="N"; indicator = TOTV; activity = NS0060; institution = 1; base_year = 2000; Title = "Total Industrial Turnover Index for durable consumer goods"; Decimals = 2.*

Series 1: *Frequency = M; adjustment ="N"; indicator = TOTV; activity =NS0080; institution = 1; base_year = 2005,* ***reference area =BE***
Observations
time_period = 2008-04*; observation value =*95.19*; status =E; confidentiality = F*
time_period = 2008-05*; observation value =87.13; status =E; confidentiality = F*

Series 2: *Frequency = M; adjustment ="N"; indicator = TOTV; activity =NS0080; institution = 1; base_year = 2005,* ***reference area =BG***
Observations
time_period = 2008-04*; observation value =138.05; status = A; confidentiality = F*
time_period = 2008-05*; observation value =137.76; status = A; confidentiality = F*

Series 1: *Frequency = M; adjustment ="N"; indicator = TOTV; activity =NS0060; institution = 1; base_year = 2000,* ***reference area =BE***
Observations
time_period = 2008-04*; observation value =101.24; status =E; confidentiality = F*
time_period = 2008-05*; observation value =86.59; status =E; confidentiality = F*

Series 2: *Frequency = M; adjustment ="N"; indicator = TOTV; activity =NS0060; institution = 1; base_year = 2005,* ***reference area =BG***
Observations
time_period = 2008-04*; observation value =165.59; status = A; confidentiality = F*
time_period = 2008-05*; observation value =168.55; status = A; confidentiality = F*



Total Industry turnover index (TOVT)
monthly data - (2005=100)
Neither seasonally or working day adjusted

	Activity	NS0080 Consumer goods				NS0060 Durable consumer goods			
Reference area	Time	2008-04		2008-05		2008-04		2008-05	
Belgium	BE	95.19	e	87.13	e	101.24	e	86.59	e
Bulgaria	BG	138.05		137.76		165.59		168.55	

e Eurostat estimate

Table 10 - STS – Total Industrial Turnover Index (Multidimensional table)

Pos. in Key	Dimension or Attribute name	Identifier	Code List	Attachment level
	DIMENSIONS			
1	Frequency	FREQ	CL_FREQ	
2	Reference area	REF_AREA	CL_AREA_EE	
3	Adjustment	ADJUSTMENT	CL_ADJUSTMENT	
4	Type of index	STS_INDICATOR	CL_STS_INDICATOR	
5	Activity	STS_ACTIVITY	CL_STS_ACTIVITY	
6	Type of institution	STS_INSTITUTION	CL_STS_INSTITUTION	
7	Base year	STS_BASE_YEAR	CL_STS_BASE_YEAR	
	Reference period	TIME_PERIOD		
	MEASURES			
	Turnover index	OBS_VALUE		
	ATTRIBUTES			
	Observation status	OBS_STATUS	CL_OBS_STATUS	Observation
	Confidentiality	OBS_CONF	CL_OBS_CONF	Observation
	Time format	TIME_FORMAT	CL_TIME_FORMAT	Series
	Title	TITLE		Group
	Decimals	DECIMALS	CL_DECIMALS	Group

Table 11 - Data Structure Concepts of the Time series example

4.5 Data set

The Data Set contains data and related metadata whose content conforms to the specification of a Data Structure Definition.

As illustrated in Figure 6, the Data Structure Definition may link to a Dataflow Definition. The Dataflow (see section 4.8) defines metadata relating to a flow of data that is collected or disseminated. Such metadata contains for example periodicity of reporting and which organizations report a Data Set.

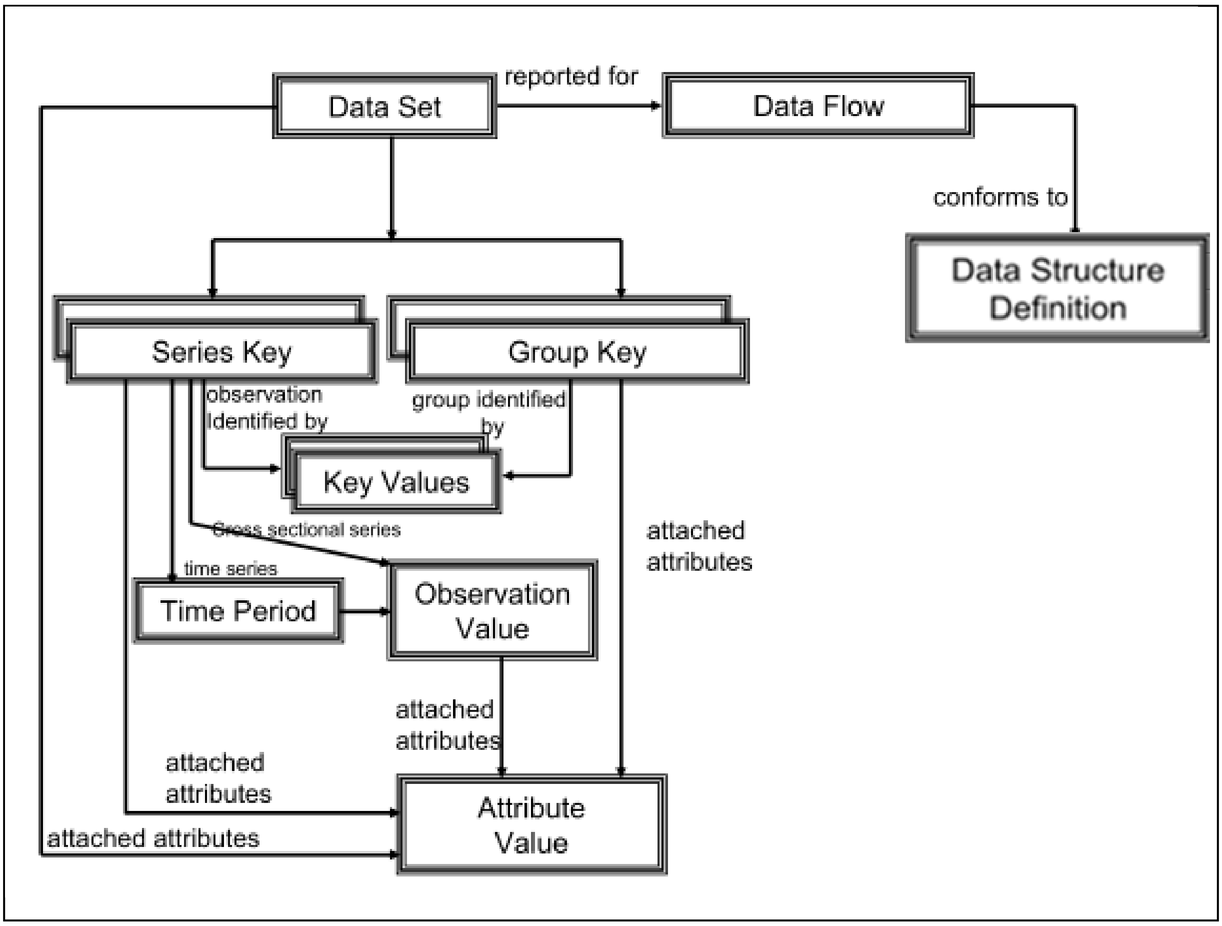


Figure 6: Dataset overview

4.5.1 Time series Data set

The time series Data Set comprises:

- Time series Keys, each of which defines the Key of a time series which, when combined with a Time Period uniquely identifies an Observation
- Optionally Group Keys, which (conceptually) comprise a set of Time series Keys for which Attribute values can be reported
- Attribute values, which are reported for a specific object as one of Data Set, Group Key, Time series Key or Observation

The main structure of the Data Set is a set of Keys and Group Keys. Each Key comprises a set of Key values, one value for each of the Dimensions defined in the Data Structure Definition. In case of time-series, for each Key there may be one or more Observation values at different times of a Time Period. Attribute values can be reported, and each of these values can be attached to the approriate level: Data Set, Series Key, Group Key or directly to an Observation value.

The Data Set may include a reference to the Data Flow, which in turn, is linked (obligatory) to the Data Structure Definition. Application can then use the Dataflow to retrieve the DSD in order to process or validate the Data Set.

4.5.2 Cross-sectional Data set

'Cross-sectional' data are types of statistical data which are not typically organized like time series. Data are organized around some other, non-time Dimension of the statistical data cube.

Cross-sectional representations of the data may be derived from the same Data Structure Definition from which time-series representations are structured, so long as the needed additional structural metadata is provided.

This functionality allows multiple Measures (so called cross-sectional Measures) to be declared in the Data Structure Definition, associated with the representational values of one Dimension. When data is structured to represent a set of multiple observations at a single point in time, the 'section' – one or more observations for each declared Measure – replaces the series in the Data Structure.

Each Measure carries at least one Dimension of the Key (the 'Measure Dimension') at the observation level, while the time period is attached at a higher level in the Data Structure (the Group level – see below). The remainder of the Key is found at the Section level (or above), similar to the way in which it is attached at the Series level for time series Data Structures.

For example, if the ACTIVITY Dimension is declared as Measure Dimension in the STS Data Structure Definition (Table 12), we define several sections that correspond to several possible values of the Dimension, let's say NS0080, NS0060, and NS0050.

Concepts describing the three cross-sectional Measures need to be defined to be then declared in the Data Structure Definition. For example:

Concept name	Corresponding value of ACTIVITY Measure Dimension
CONGIND	NS0080
MIGDCG	NS0060
MIGCDI	NS0050

Table 12 - Concepts of the ACTIVITY Measure Dimension

In such a situation the Concepts' definition for the STS data is enriched with three new Measures:

Pos. in Key	Dimension or Attribute name	Identifier	Code List	Attachment level
DIMENSIONS				
1	Frequency	FREQ	CL_FREQ	Section
2	Reference area	REF_AREA	CL_AREA_EE	Section
3	Adjustment	ADJUSTMENT	CL_ADJUSTMENT	Section
4	Type of index	STS_INDICATOR	CL_STS_INDICATOR	Section
5	Activity / Measure Dimension	STS_ACTIVITY	CL_STS_ACTIVITY	Observation
6	Type of institution	STS_INSTITUTION	CL_STS_INSTITUTION	Section
7	Base year	STS_BASE_YEAR	CL_STS_BASE_YEAR	Section
	Reference period	TIME_PERIOD		Group
MEASURES (of the Activity Measure Dimension)				
	Consumer goods industry	CONGIND		
	Durable Consumer Goods Industry	MIGDCG		
	Capital Goods Industry	MIGCDI		
ATTRIBUTES				
	Observation status	OBS_STATUS	CL_OBS_STATUS	Observation
	Confidentiality	OBS_CONF	CL_OBS_CONF	Observation
	Time format	TIME_FORMAT	CL_TIME_FORMAT	Section
	Title	TITLE		Group
	Decimals	DECIMALS	CL_DECIMALS	Group

Table 13 – Data Structure Concepts of the example with Cross-sectional Measures

The following tables illustrate the two data representations: time series and cross-sectional:

Total Industry turnover index (TOVT) - Time series			
REF_AREA: Italy INDICATOR: TOVT BASE YEAR: 2005			
Indicator **Time**	**NS0080** **Consumer goods**	**NS0060** **Durable consumer goods**	**NS0050** **Capital Goods Industry**
2007M01	100.8	82.4	93.7
2007M02	106.5	97.1	106.8
2007M03	121.2	121.3	134.9
2007M04	97.7	104.4	106.4
2007M05	111.9	121.8	134.5
2007M06	116	120.3	136.6
Total Industry turnover index (TOVT) – Cross-sectional			
TIME: 2007M01 INDICATOR: TOVT BASE YEAR: 2005			
Indicator **Country**	**NS0080** **Consumer goods**	**NS0060** **Durable consumer goods**	**NS0050** **Capital Goods Industry**
ES	99.6	104.6	108.4
FR	102.3	92.8	93.4
IT	100.8	82.4	93.7

Table 14 - Time series and Cross-sectional table

4.6 Metadata Structure Definition (MSD)[1]

A Metadata Structure Definition defines the valid content of a Metadata Set (see 4.7) in terms of the Concepts comprising the structure of the Metadata Set, how the Concepts are related in terms of their role in the Metadata Set, and the valid content of each of the Concepts when used in a Metadata Set.

Metadata Structure Definitions are structured in a fairly simple fashion – because reference metadata is less structured than aggregated statistical data. The simpler structure for it consists either in a flat list or a hierarchy.

A MSD carries a set of Concepts (which might originate from different Concept Schemes) and organizes them for example into a hierarchy. Also it states for each Concept, whether its reporting is required or optional and assigns a Concept's representation (code lists, text format, etc.). In addition, it identifies the subject(s) - so called Target identifiers - of the reported metadata, i.e. in relation with which data the reference metadata is reporting.

[1] MSD and Metadata set are described more in detail in the Student book 5 – Metadata Structure Definition

4.7 Metadata set

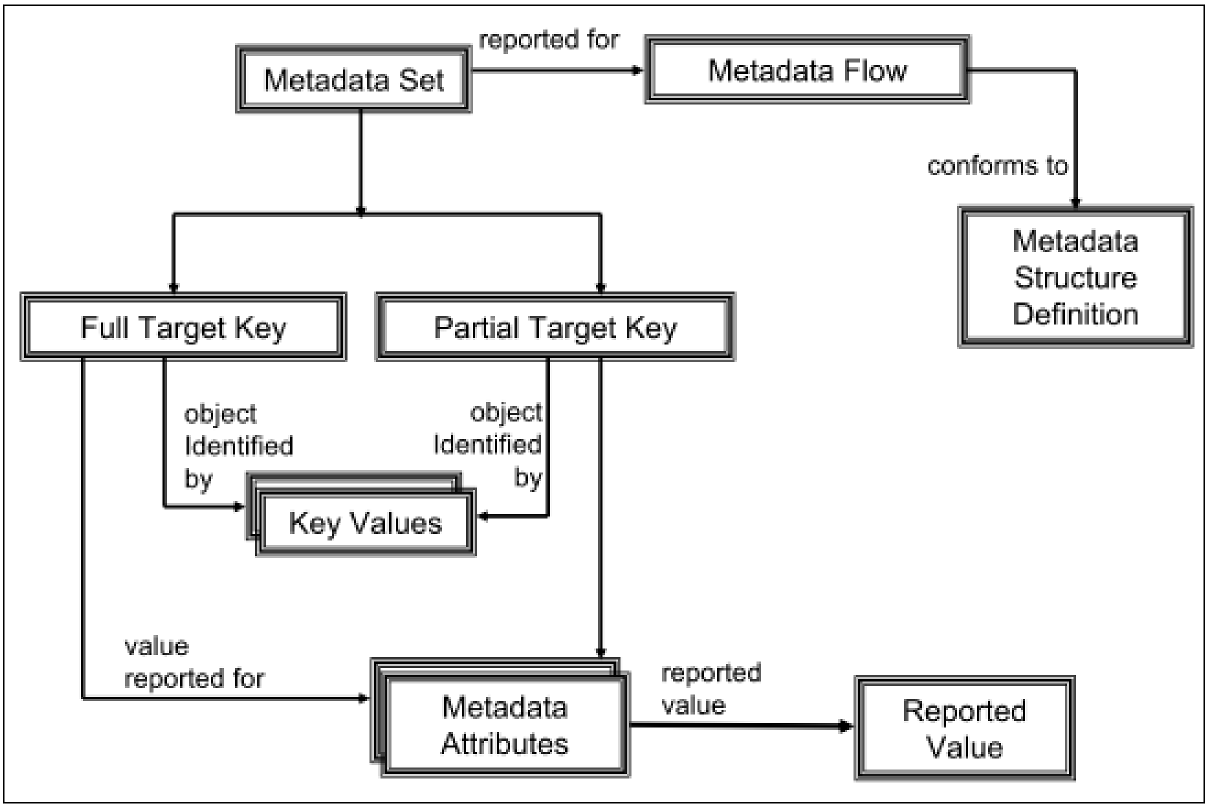


Figure 7 - Metadata set

A Reference Metadata set is reported for a Metadata flow, which has to conform to a Metadata Structure Definition. Target Keys identify the target data or metadata object for which the Metadata Attributes (with its particular values) are reported. The fact that the Metadata Set includes a reference to the Metadata flow, which is itself strictly linked to the Metadata Structure Definition, allows an application to retrieve the MSD so that the Metadata report can be processed, and validated if required.

A Reference Metadata set is a set of information regarding almost any object within the formal SDMX view of statistical exchange, they may describe:

- The maintainers of data or structural definitions;
- The schedule on which data is released;
- The flow of a single type of data over time;
- The quality of data (in accordance with a data quality framework).

The Metadata Set generally comprises multiple metadata reports, while each report defines:

- A single target object, to which the reference metadata is attached;
- The reported values for the metadata Attributes (as specified in the MSD) those make up the specific report structure.

4.8 Dataflow & Metadata flow definition

In SDMX, Data sets are reported or disseminated according to a Dataflow definition. The Dataflow definition is linked to the Data Structure Definition and may be associated with one or more subject matter domains , this facilitates the search for data according to organised Subject Matter Scheme (called Category Scheme in the model) as they provides a way of classifying data for collection, reporting, or publication.

Dataflows defined for STS		
DSD	**SDMX**[2]	**GESMES**
	ESTAT+STS+2.0.xml	EUROSTAT_STS_TS.gsm
Dataflows	**SDMX**[2]	**GESMES**
Production in industry	SSTSIND_PROD_M	STSIND_PROD_M, _Q
Turnover in industry, total, domestic, non-domestic, split of non-domestic for Euro area, non-Euro area		STSIND_TURN_M, _Q
New orders received in industry, total, domestic, non-domestic, split of non-domestic for Euro area, non-Euro area	SSTSIND_ORD_M	STSIND_ORD_M, _Q
Number of persons employed, Number of employees, in industry		STSIND_EMPL_M, _Q
Hours worked in industry		STSIND_HOUR_M, _Q
Gross wages and salaries in industry		STSIND_EARN_M, _Q
Output prices in industry, total, domestic market, non-domestic market, split non-domestic market Euro area, non Euro area, Import prices (total, Euro area, non-Euro area)	SSTSIND_PRIC_M	STSIND_PRIC_M
Production in construction, total, building construction, civil engineering	SSTSCONS_PROD_M, _Q	STSCONS_PROD_M, _Q
New orders received in construction, total, building construction and civil engineering		STSCONS_ORD_M, _Q
Number of persons employed, Number of employees, in construction		STSCONS_EMPL_M, _Q
Hours worked in construction		STSCONS_HOUR_M, _Q
Gross wages and salaries in construction		STSCONS_EARN_M, _Q
Output prices in construction, construction costs, material costs, labour costs		STSCONS_PRIC_M, _Q
Building permits, number of dwellings or square metres of useful floor area		STSCONS_PERM_M, _Q
Turnover in retail trade, value or deflated	SSTSRTD_TURN_M	STSRTD_TURN_M
Number of persons employed, Number of employees, in retail trade		STSRTD_EMPL_M, _Q
Turnover in repair and other services, value or deflated	SSTSSERV_TURN_M, _Q	STSSERV_TURN_M, _Q
Output prices in other services	STSSERV_PRIC_Q	STSSERV_PRIC_Q
Number of persons employed, Number of employees, in repair and other services	STSSERV_EMPL_Q	STSSERV_EMPL_M, _Q
Any other indicator not mentioned in the list above		STSOTHER_OTH_M, _Q

Table 15 – STS - DSD and Dataflows (SDMX & GESMES)

[2] DSD and Dataflows from Eurostat's “SDMX OPEN DATA INTERCHANGE” (SODI) project

A ‘Dataflow’ can be seen as the on-going publication of a Data set, as new observations are added to the existing ones, or as subsequent Data sets with the same subject and structure are published. ‘Dataflows’ can be understood as Data sets which are not bounded by time.

Data structures and Dataflows are owned and maintained by agencies. In a data collection environment both, Data Structures and Dataflows will normally be maintained by the data collector agency.

The diagram below depicts the essential characteristics supported in the SDMX Information Model for reporting or publishing data and metadata. The pivot of this diagram is the Data or Metadata Flow.

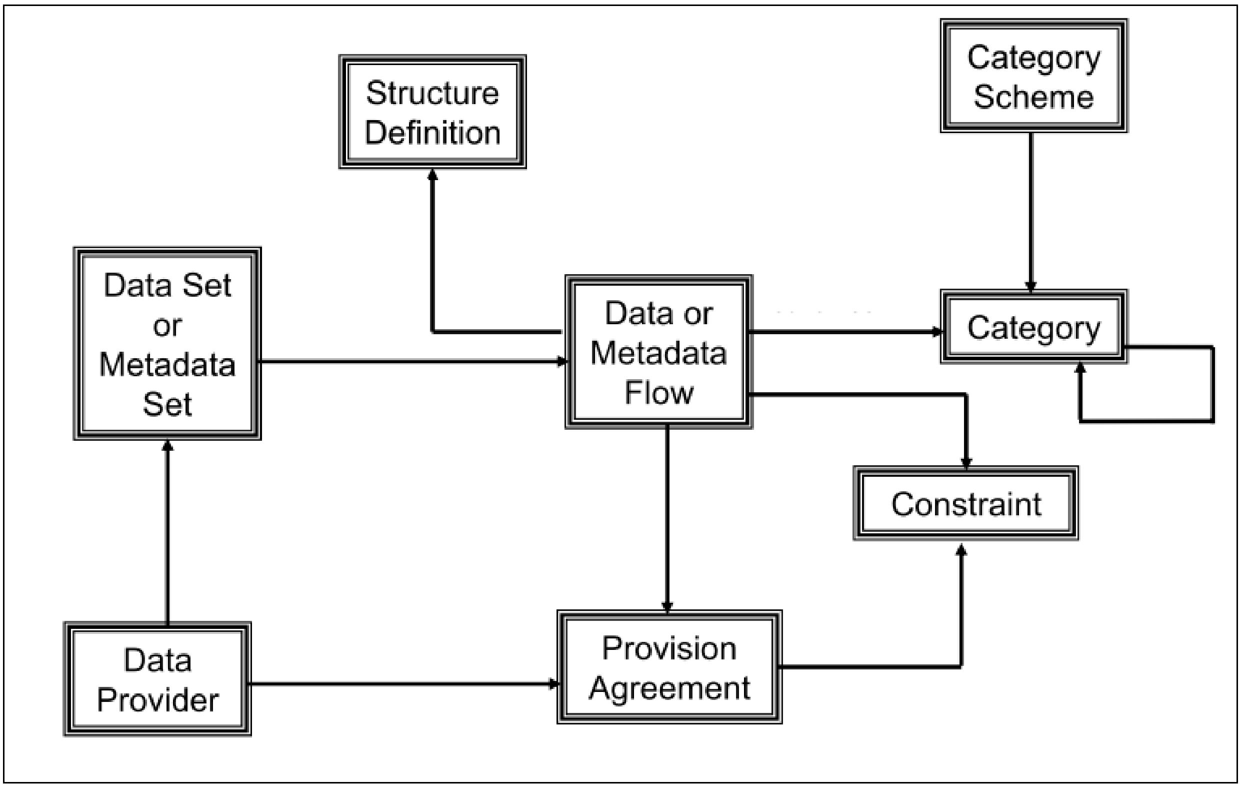


Figure 8 - Data and Metadata flow definition

The Data or Metadata Flow is maintained by the organization that collects or ‘harvests’ the data or metadata in order to use it or to publish it. The Data Flow is linked to a single Data Structure Definition (Structure Definition on the diagram), but in contrary a DSD can be the basis for more than one Dataflow. Similarly, the Metadata flow is linked to a Metadata Structure Definition. The Data or Metadata sets for the data or reference metadata Flow may be provided by many Data Providers and any one Data Provider may report or publish Data or Metadata sets for many Data or Metadata Flows – typically a Data Provider may supply Data or Metadata sets for many topics or Categories of statistical data.

The SDMX Information Model allows Constraints to be attached to the Dataflow definition. Constraints may concern reporting periodicity or sub set of possible Key Dimensions that are allowed in a Data set. Constraints are tackled in a dedicated paragraph below.

A Metadata flow definition is very similar to a Dataflow definition, but describes, categorises, and constrains Metadata sets. In a Metadata Set there need only be a reference to the Metadata flow Definition as this enables an application to identify the relevant Metadata Structure Definition which will enable it to validate and process the metadata.

4.9 Data provider

In SDMX a data provider is an organization which produces data or reference metadata and makes them available for other organisations. Data providers may provide Data sets to many different Dataflows. Data flows may incorporate data coming from more than one data provider. In order to manage this process the Data Provider is linked with the Dataflow Definition or Metadata flow Definition by means of a Provision Agreement, objects which tell you which data providers are supplying what data to which Dataflows. The same is true for Metadata flows.

4.10 Provision Agreement

Inherent in any statistical exchange – and in many dissemination activities - is a concept of ‘service level agreement’, even if this is not formalized or made explicit. SDMX incorporates this idea in objects termed ‘provision agreements’.

It is the set of information which describes the way in which Data sets and Metadata sets are provided by a data provider. A provision agreement can be constrained in terms of the sub set of Keys or code values in much the same way as a Data or Metadata flow definition.

Thus, a data provider can express the fact that it provides a particular Dataflow covering a specific set of countries and topics, with a particular publication schedule (Release calendar).

4.11 Constraints

The term ‘Constraints’ in the SDMX Information Model embeds a set of information: the specific topics about which data or metadata is reported within the theoretically possible set of data (as described by a Data Structure Definition or reference Metadata Structure Definition), and the time period covered by the statistical data and metadata.

Constraints are associated:

- With Dataflows - typically describing the topics covered (= **content related Constraints**);
- With the provision agreement - where a full description of **time-related Constraints** and **content related Constraints** is given.

Time-related Constraints

Time-related Constraints are generally defined in the provision agreement. The data provision may be constrained by a data provider with regard to the periods, which are reported. For example that data is only provided since the year 2000, because incompleteness or reliability problems for older data.

Another time-related constrained could be that a data provider constraints the data provision to only yearly Data sets for one statistical indicator, while the majority of other data providers report quarterly and yearly Data sets for the same indicator.

Content related Constraints:

Data or Metadata Provider can apply Constraints on the scope of the data or metadata that can be supplied, in terms of Key ranges or complete Key sets. These Constraints on the content of a data provision can be specified for a Dataflow Definition, Metadata flow Definition, and a Provision Agreement. They constrain the use of a corresponding Structure definition for a Data / Metadata set in relation to its Dataflow Definition or Provision Agreement (provisioning Constraints). Thus, the Data set or Metadata set reported based on a DSD or MSD may be constrained by:

- Specifying one or more sets of Keys from the theoretical set that could be created from the Dimensions (often called the 'Cartesian product') which may be included in or excluded from the complete set of Keys used in the Data Structure Definition (resp. the Metadata Structure Definition). The complete set of Keys represents all possible combinations of values of the Key Dimensions. A Data Provider might supply data for a sub set of the coded values in any one of the Dimensions comprising the Key.

Example:

A Data Provider might report only a limited set of indicators (e.g. in STS – only the activities of the previous example: NS0080, NS0060 & NS0050) or limit the reporting of Trade data to imports and exports with specific partner countries (e.g. to provide only data regarding trade with the EU Member states).
Thus, in both cases the data provision is constrained, since a part of the values from Key Dimensions are excluded from the data exchange.

- Specifying one or more 'Cube Regions', each of which comprises a set of 'specific selections' which, together, may be included in or excluded from the set of values that are valid for a Dimension, Attribute, or Measure. Each selection specifies the constrained list of values that are valid. This list must be either
  - a complete set or a sub set of the full representation specified for the relevant Dimension, Attribute, or Measure in the Data Structure Definition;

Example:

A Data Provider might suppress the reporting of several sensitive products in External Trade data, where the value per unit exceeds a specific amount (e.g. 10 million Euros).

  - or a complete set or a sub set of the list of valid values (Code List) specified for an identifier component.

Example:

A Data Provider might suppress the reporting of several sensitive products in External Trade data (like special machinery or weapons). Consequently the respective coded values in the product code list are excluded from reporting.

4.12 Category Scheme

A Category List is maintained by a Maintenance Agency. A Category Scheme provides a way of classifying data for collection, reporting, or publication. They are made up of a simple hierarchy of Categories (a Category may have one or more child Categories), which in SDMX may include any type of useful classification for the organization of data and metadata.

In order to support data and metadata discovery, the Dataflow Definition and the Metadata flow Definition may also be linked to one or more Categories in one or more Category Schemes such as a scheme of subject matter domains or reporting categories. Typical Category Schemes may comprise many high level Categories such as financial, economic, health, tourism, transport, demography for one organisation. Each of these would be further segmented into lower level Categories forming a hierarchy of Categories.

Stepping down the hierarchy will lead to the lowest level Category, where the Data and Metadata Flows that are linked to. These flows identify the Data and Metadata Sets and their publishers (Data Providers).
In the fictitious example in Table 16 this breakdown pattern is shown for the first Category STS for STS_IND and STS_CONS. An assumed SDMX Dataflow (STS_IND_PROD_M) would consequently be linked to the Category 'STS_IND_PROD'.

Category Scheme (no real SDMX Categories – only for illustration)	
CATEGORYSCHEME_ID	**Description**
INDU_TRADE_SRV_SCHEME	**A Category Scheme for Industry, trade & services**
CATEGORY_ID	**Description**
ICTS	**Industry, Trade and Services**
STS	**Short-term Business Statistics (STS)**
STS_IND	STS Industry
STS_IND_PROD	STS Industry Production index
STS_IND_TOVT	STS Industry Turnover index
…	…
STS_CONS	STS Construction
STS_CONS_PRO	STS Construction Production index
STS_CONS_LAB	STS Construction Labour input index
…	…
STS_TS	STS Trade and Services
SBS	**Structural Business Statistics (SBS)**
SBS_NA	SBS Main indicators
SBS_IND_CO	SBS Industry and Construction
SBS_DT	SBS Trade
SBS_SERV	SBS Services
…	…
TOUR	**Tourism Statistics**
PROM	**Production of Manufactured Goods Statistics**
ISCO	**Information Society Statistics**

Table 16 – Example: Category Scheme - Industry, trade and services

5 Glossary

Table 17 presents the list of concepts and acronyms with their definition.

Concept	Definition
DSD	Data Structure Definition
ESMS	Euro SDMX Metadata Structure
GESMES	Generic Statistical Message
ID	Identifier
IT	Information Technology
MSD	Metadata Structure Definition
NUTS	Nomenclature of Territorial Units for Statistics (for the French: nomenclature d'unités territoriales statistiques)
SBS	Structural Business Statistics
SDMX	Statistical Data and Metadata eXchange.
SDMX-IM	SDMX Information Model
SDMX-ML	SDMX Markup Language - XML format for the exchange of SDMX-structured data and metadata
SODI	SDMX Open Data Interchange
STS	Short Term Business Statistics
XML	EXtensible Markup Language

Table 17 – Glossary