eurostat

sdmx
Statistical Data and Metadata eXchange

SDMX self-learning package No. 1
Student book

Introduction to SDMX

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

TABLE OF CONTENTS

1 Scope of the student book

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

- Its history.
- Its components.
- The information model for data and metadata.

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

This student book is the first one of a set of 8 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]	The SDMX Information Model
[03]	SDMX-ML Messages
[04]	Data Structure Definition
[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 – Student books on SDMX

Prerequisites

The student book is a general introduction. No specific prerequisite is needed to understand the content.

2 Introduction and History

2.1 Scope of this chapter

This chapter explains the origin of SDMX, its purpose and history.

2.2 Origin and purpose of SDMX

The Statistical Data and Metadata eXchange (SDMX) initiative was launched in 2001 by seven organisations working on statistics at the international level: the Bank for International Settlements (BIS), the European Central Bank (ECB), Eurostat, the International Monetary Fund (IMF), the Organisation for Economic Co-operation and Development (OECD), the United Nations Statistical Division (UNSD) and the World Bank. These seven organisations act as the sponsors of SDMX.

With the Internet and the world-wide web, the electronic exchange and sharing of data has become easier and more common, but the exchange has often taken place in an ad hoc manner using all kinds of formats and non-standard concepts. This creates the need for common standards and guidelines to enable more efficient processes for exchange and sharing of statistical data and metadata. As statistical data exchange takes place continuously, the gains to be realised from adopting common approaches are considerable both for data providers and data users.

SDMX aims to ensure that metadata always come along with the data, making the information immediately understandable and useful. For this reason, the SDMX standards and guidelines deal with both data and metadata.

The stated aim of SDMX was to develop and use more efficient processes for exchange and sharing of statistical data and metadata among international organisations and their member countries. To achieve this goal, SDMX provides standard formats for data and metadata, together with content guidelines and an IT architecture for exchange of data and metadata. Organisations are free to make use of whichever elements of SDMX are most appropriate in a given case.

Common standards and guidelines followed by all players not only help to give easy access to statistical data, wherever these data may be and without demanding prior agreement between two partners, but they also facilitate access to metadata that make the data more comparable, more meaningful and generally more usable.

2.3 History

The Version 1.0 SDMX standards include the information model as well as the XML-based data format SDMX-ML and the GESMES/TS data format, renamed SDMX-EDI.

The Version 1.0 SDMX standards were approved by the sponsors in September 2004 and accepted as an ISO technical specification (ISO/TS 17369:2005) in April 2005.

In November 2005, the sponsors approved Version 2.0 of the SDMX standards, which are fully compatible with Version 1.0 but in addition provide for the exchange of reference (explanatory) metadata, and include the registry interface specification. SDMX 2.0 standards are being submitted to ISO in 2008, with some adjustments and corrections to take account of comments received since Version 2.0 was released in 2005.

The first draft of the Content-Oriented Guidelines was released for public review in March 2006, and a consolidated version was released for public review in February 2008. The full release of the Content-Oriented Guidelines, which has been extensively revised to take account of comments received from many organisations, took place in January 2009.

In March 2007, the sponsoring institutions signed a Memorandum of Understanding (MoU), which is intended to set out the arrangements for a durable collaboration by the sponsors on all aspects of SDMX. The MoU explicitly excludes the formation of any legal entity or common budget for SDMX; each sponsoring institution and its member countries will continue to use its existing procedures to agree on arrangements for transmission and publication of statistics.

In the conclusions of the 39th Session of the UN Statistical Commission (New York, February 2008), SDMX was recognised and supported as ‘the preferred standard for exchange and sharing of data and metadata in the global statistical community’ . This acceptance of SDMX at UN level is a major step forward towards the broader use of SDMX at world-wide level.

3 SDMX components

3.1 Scope of this chapter

SDMX is more than a data transmission format since it answers to several questions:

- What is the information model underlying the data and metadata exchanged between the partners?
- How to increase the interoperability and statistical harmonisation?
- How to exchange the data?

This chapter explains the 3 components of the SDMX standard which provide the answers to the above questions:

- The SDMX Information model
- Content-Oriented Guidelines IT Architecture for the Data Exchange

A global overview of these components is provided. Further detailed information is available in the other student books.

An additional paragraph explains available SDMX tools and how they can support the data transmission and production process.

3.2 The SDMX Information Model

3.2.1 What is an Information Model?

An information model is a representation of concepts, relationships, constraints, rules, and operations to specify data semantics for a chosen domain of discourse. It can provide sharable, stable, and organised structure of information requirements for the domain context.

Within the field of software engineering and data modelling an information model is an abstract, formal representation of entity types that includes their properties, relationships and the operations that can be performed on them. The entity types in the model may be kinds of real-world objects, such as devices in a network, or they may themselves be abstract, such as for the entities used in a billing system. Typically, they are used to model a constrained domain that can be described by a closed set of entity types, properties, relationships and operations.

An information model provides formalism to the description of a domain without constraining how that description is mapped to an actual implementation in software. There may be many mappings of the information model. Such mappings are called data models, irrespective of whether they are object models (e.g. using UML), entity relationship models or XML schemas.

3.2.2 Statistical Data and Metadata

3.2.2.1 Statistical Data Statistical data are sets of often numeric observation which typically have time associated with them. In order to understand their meaning we need a set of statistical concepts which act as identifiers and descriptor of the data. These metadata values can be understood as the named dimension of a multi-dimensional coordinate system, describing what is often called a 'cube' of data.

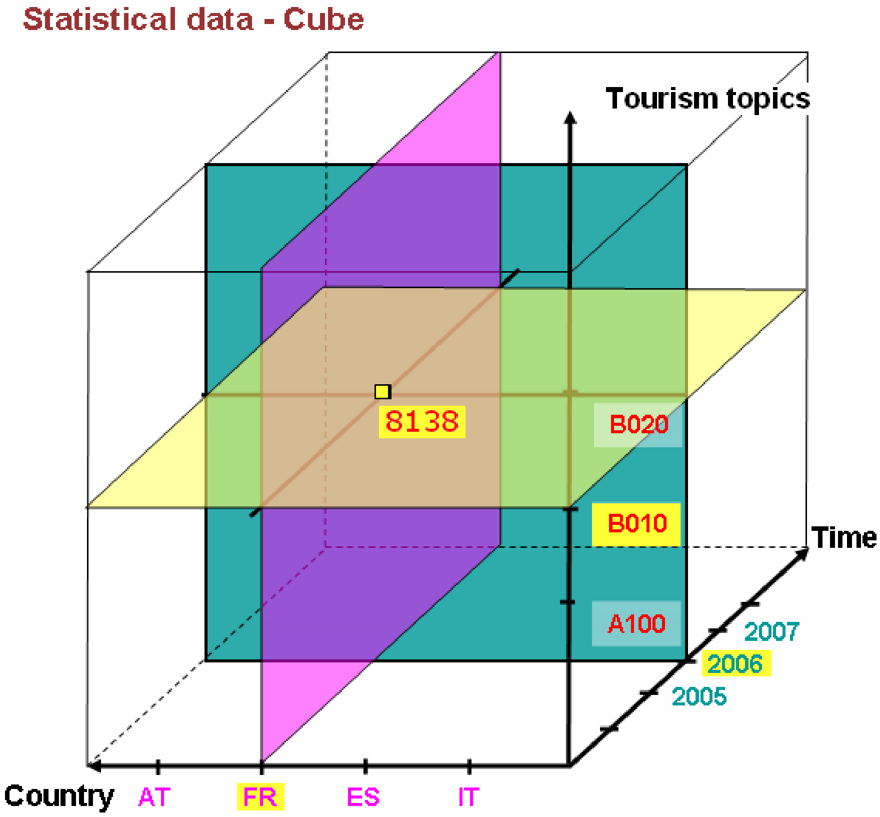


Figure 1 – A Cube in Tourism statistics

In the example in Figure 1, the number 8138 represents a statistical data observation, which can be described by the following statistical concepts:
The '**Unit:** *Number*' of '**Topic:** *B010 – Tourist Campsides*' in the '**Reference Area:** *France*', on '**Time:** *31 December 2006*'.

Descriptor Concepts

Statistical figures need to be described with concepts to make them meaningful. A simplified example could contain the following concepts:

- Concept: FREQUENCY (dimension – values provided by Frequency code list: Annual, Quarterly, Monthly, etc.);
- Concept: REFERENCE_AREA (dimension - values provided by a Country code-list);
- Concept: OBSERVATION_VALUE (measure – with a numeric data type);
- Concept: TOPIC (dimension - which names the topics of the statistical dataset);
- Concept: TIME (dimension - values provided by a Date representation);

- Concept: UNIT (attribute – unit in which the measure is expressed).

Dimensions, attributes and measures:

Descriptor concepts can be distinguished between dimensions, which describe the data and form the identifier (ID) of the related data, and attributes, which provide additional descriptive information to qualify the data. Attributes are for example the unit of measure or information on the ‘status’ of an observation (provisional, forecasted, estimated, revised, etc) as a qualifier with regards to the confidentiality of the data.

Measures include the phenomenon with is observed. While for the time-series data representation only one primary measure is declared, the cross-sectional data representation has usually multiple measures declared (array of measures) in a specific measure dimension.

3.2.2.2 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

Reference (or explanatory) metadata – the large set of concepts that describe and qualify statistical data sets and processes more generally, and which are often associated not with a specific observations or series of data, but with an entire collection of data or even the institution which provide that data. 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 dimensions of the statistical data.

Metadata is also associated, not only with data, but also with the process of providing and managing the flow of data. This kind of metadata concern with ‘data provisioning’ – metadata which is useful to those who need to understand the content and form of a data provider’s output. For example supplying information about the scheduling of data provision and mechanism by which their data and metadata are provided.

Finally, in order to organize and describe the exchange and dissemination of data and metadata, it is possible to express information about classification schemes and domain categories, along with their relationships to data and metadata sets.

3.2.3 SDMX Information Model (IM)

SDMX provides a way of modelling statistical data, structural metadata and the data exchange processes. SDMX also defines a model for additional explanatory metadata, so called reference metadata, which is generally in a textual format. In order to produce technical standards that could support different models of statistical exchange, the SDMX IM provides a broader set of formal objects which describe the actors, processes, and the resources within statistical exchanges.

The SDMX-IM through a structural pattern allows the specification of complex tabular structure which is often found in statistical data.

In the following a brief summary of the main objects is provided. A complete description of all objects detailed in the SDMX-IM will be provided in the student book 2 named ‘SDMX Information Model’.

Data Set

Data Set: a collection of similar data, sharing a structure, which covers a fixed period of time. Data sets are made up either of a number of time series, or a view of several related time series’data at a single point of time (a ‘time slice’).

Number of touristic establishments,
Multidimensional example

Indicator	A100 - Hotels and similar			B010 - Tourist Campsites			B020- Holiday dwellings		
Time Country	2005A00	2006A00	2007A00	2005A00	2006A00	2007A00	2005A00	2006A00	2007A00
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	68385	68376	61810

Table 2 - Multidimensional data representation

Time series data is a set of observations that represent the measurement of a statistical phenomenon over time (This allows us to see trends in the phenomenon being measured).

Number of touristic establishments,
in France - annual data

Indicator Time	A100 Hotels and similar	B010 Tourist Campsites	B020 Holiday dwellings
2005A00	18689	8174	2329
2006A00	18361	8138	2325
2007A00	18135	8052	2406

Table 3 - Time series data set

Cross-sectional data is intended to present more than one measurement, for example where the statistical data set consists of multiple diffent measurements / observations at a single point in time.

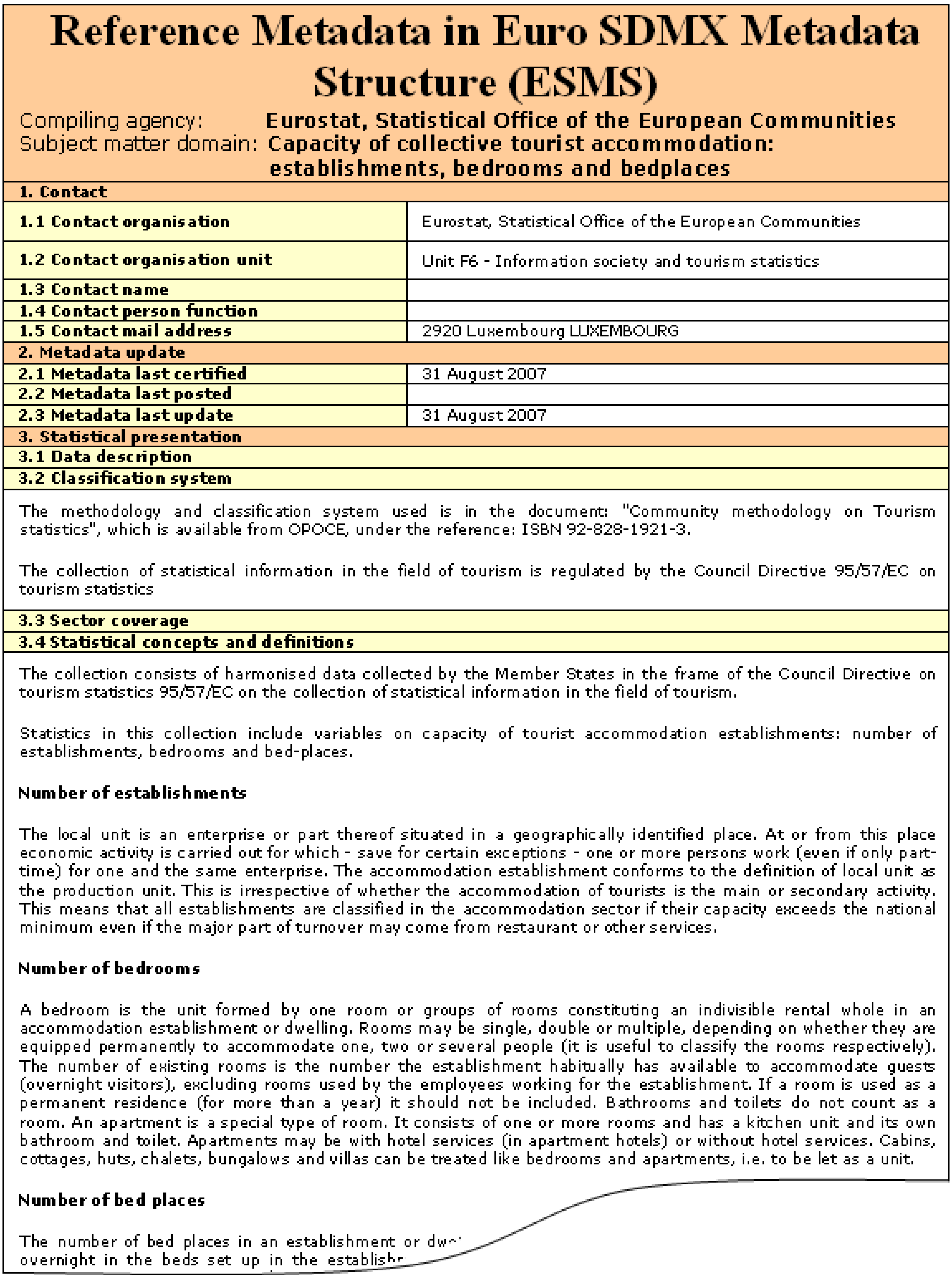
Number of touristic establishments,
annual data for 2006

Indicator Country	A100 Hotels and similar	B010 Tourist Campsites	B020 Holiday dwellings
AT	14051	542	3329
ES	18304	1216	4524
FR	18361	8138	2325
IT	33768	2510	68376

Table 4 - Cross-sectional data set

Metadata Set

A reference metadata set is information regarding the formal SDMX view of statistical exchange: it may describe the statistical approach; the maintainers of data or data structures; data release calendar; the quality of data, etc in relation with a respective structure definition (MSD). The below Table 5 presents an example of reference Metadata of the statistical domain 'Tourism' formatted according to Euro SDMX Metadata Structure (ESMS) reporting format.



Reference Metadata in Euro SDMX Metadata Structure (ESMS)

Compiling agency: **Eurostat, Statistical Office of the European Communities**
Subject matter domain: **Capacity of collective tourist accommodation: establishments, bedrooms and bedplaces**

1. Contact	
1.1 Contact organisation	Eurostat, Statistical Office of the European Communities
1.2 Contact organisation unit	Unit F6 - Information society and tourism statistics
1.3 Contact name	
1.4 Contact person function	
1.5 Contact mail address	2920 Luxembourg LUXEMBOURG
2. Metadata update	
2.1 Metadata last certified	31 August 2007
2.2 Metadata last posted	
2.3 Metadata last update	31 August 2007
3. Statistical presentation	
3.1 Data description	
3.2 Classification system	

The methodology and classification system used is in the document: "Community methodology on Tourism statistics", which is available from OPOCE, under the reference: ISBN 92-828-1921-3.

The collection of statistical information in the field of tourism is regulated by the Council Directive 95/57/EC on tourism statistics

3.3 Sector coverage

3.4 Statistical concepts and definitions

The collection consists of harmonised data collected by the Member States in the frame of the Council Directive on tourism statistics 95/57/EC on the collection of statistical information in the field of tourism.

Statistics in this collection include variables on capacity of tourist accommodation establishments: number of establishments, bedrooms and bed-places.

Number of establishments

The local unit is an enterprise or part thereof situated in a geographically identified place. At or from this place economic activity is carried out for which - save for certain exceptions - one or more persons work (even if only part-time) for one and the same enterprise. The accommodation establishment conforms to the definition of local unit as the production unit. This is irrespective of whether the accommodation of tourists is the main or secondary activity. This means that all establishments are classified in the accommodation sector if their capacity exceeds the national minimum even if the major part of turnover may come from restaurant or other services.

Number of bedrooms

A bedroom is the unit formed by one room or groups of rooms constituting an indivisible rental whole in an accommodation establishment or dwelling. Rooms may be single, double or multiple, depending on whether they are equipped permanently to accommodate one, two or several people (it is useful to classify the rooms respectively). The number of existing rooms is the number the establishment habitually has available to accommodate guests (overnight visitors), excluding rooms used by the employees working for the establishment. If a room is used as a permanent residence (for more than a year) it should not be included. Bathrooms and toilets do not count as a room. An apartment is a special type of room. It consists of one or more rooms and has a kitchen unit and its own bathroom and toilet. Apartments may be with hotel services (in apartment hotels) or without hotel services. Cabins, cottages, huts, chalets, bungalows and villas can be treated like bedrooms and apartments, i.e. to be let as a unit.

Number of bed places

The number of bed places in an establishment or dw
overnight in the beds set up in the establishr

Table 5 - Metadata set (truncated)

Concept Schema

A concept scheme is a maintained list of concepts that are used in data structure definitions (key family) and metadata structure definitions. There are many such concept schemes maintained according to specific statistical domain data structures. A ‘core’ representation of the concept can be specified (e.g. a core code list, or other representation such as for example ‘date’).

Data Structure Definition (DSD)

The DSD describes the structure of a particular set of data through a list of descriptor concepts. It defines which concepts are *dimensions* (identification and description - for example: Frequency, country, variable/topic, time period), and which are *attributes* (just description / qualification - for example the unit of measure, confidentiality flag or the data status flag).

In addition it determines the *attachment level* for each of these concepts, based on the packaging structure (*Data Set*, *Group*, *Series/Section* or *Observation*) as well as if they are mandatory or conditional.

Associated code lists provide possible values for the dimensions, as well as the possible values for the attributes (if attributes don’t use free text).

Thus a DSD is a set of descriptor concepts, associated with a set of data, which allows any user to understand what that data means.

Annual average population by sex
(in millions)

	MALE	FEMALE	MALE	FEMALE
	A:DE:2	**A:DE:3**	**A:US:2**	**A:US:3**
2008a0	39.77	41.07	114.96	118.11
2007a0	39.63	41.03	114.67	117.93
2006a0	39.57	40.97	114.43	117.75
2006a0	39.49	40.92	114.27	117.57

Table 6 - Sample simplified Demography data (example – non real data)

From reviewing the tables, we can derive the following *statistical (= descriptor) concepts* and their possible values, providing also sample code values:

1. Frequency (M = monthly, Q = quarterly, A = annual [for the ‘full year’ numbers]).
2. Reference Country (DE = Germany, US = United States).
3. Demographical indicator (Average population: 1= Total; 2=Male; 3=Female).
4. Unit of measure (Number of people).
5. Unit multiplier (6 = millions).

The time series are expressed in millions of People. ‘Unit of measure’ and ‘unit multiplier’ do not contribute to the identification of our series, but provide additional information / qualify them as *attributes* attached at the data set level.

The concepts 1 to 3 above are required to *identify* the time series, so they act as *dimensions* thus forming the key in the order (separator ':'): **A:DE:3.** This key would identify the time series of the annual (=Q) German male population.

We also indicate the 'compiling agency' for the data, which we attach to the data set level, too.

It should be noted that in this particular example, all concepts, apart from the observed topic - the demographic indicator - are SDMX Cross Domain Concepts, taken from the SDMX Content Oriented Guidelines. A simplified formal definition of the data structure is provided in Table 7 - DSD: Simplified example Demographical data.

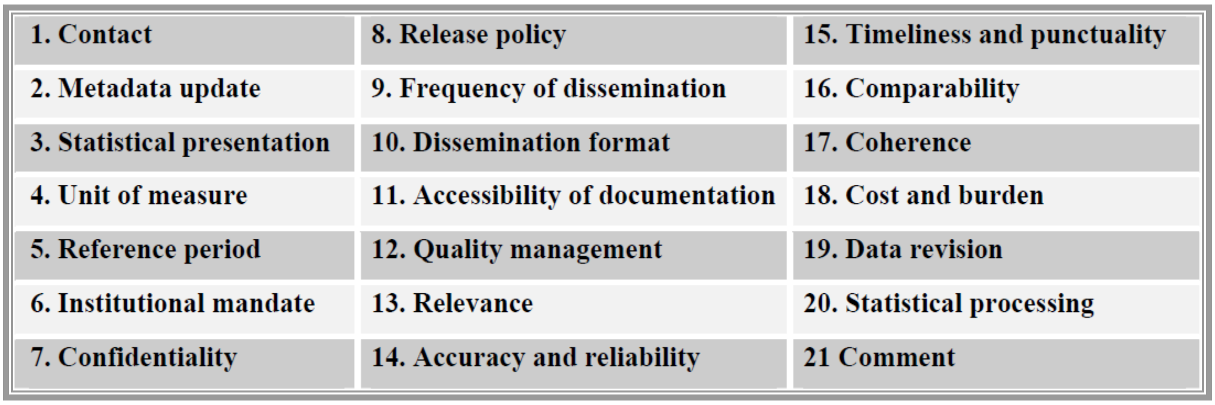
Pos. in key	Dimension or Attribute name	Identifier	Presentation[10]	Attachment level	Code list
1	Frequency	FREQ	A1		CL_FREQ
2	Reporting/reference area	REF_AREA	A2		CL_AREA
3	Demography indicator	DEMO_ITEM	A3		DEMO_ITEM
	Unit of measure	UNIT_MEASURE	AN3	Data set	CL_UNIT_MEASURE
	Unit multiplier	UNIT_MULT	N1	Data set	CL_UNIT_MULT
	Compiling agency	CL_COMPILING_ORG	AN3	Data set	CL_ORGANISATION

Table 7 - DSD: Simplified example Demographical data

Metadata Structure Definition (MSD)

A Metadata Structure Definition describes how metadata sets, containing reference metadata are organized. In particular, it defines which metadata are being exchanged, how these concepts relate to each other, how they are represented (either as free text or coded values) and with which object types (agencies, data flows, data providers, subsets of data flows, or others) they are associated. Any organization providing information about statistical data uses a set of metadata concepts (e.g. frequency of dissemination, reference area, timeliness, type of source data) in order to present the characteristics and quality of the data. Reference metadata may be attached to different data objects (for instance to a data set, a time series, or an observation). However, this kind of metadata is usually attached at a high level (data set, data flow or even at agency level). Some of these metadata concepts may be proprietary to the data provider, but interoperability will clearly be enhanced when the same concepts can be used by many exchange partners and across statistical domains.

SDMX therefore recommends the use of a set of common concepts to replace the existing SDDS format for representation of metadata. Recently the new Euro-SDMX Metadata Structure (ESMS) aims at documenting methodologies, quality and the statistical production processes in general. The ESMS, uses SDMX concepts and is supported by a 'generic' XML metadata format fully compliant with SDMX version 2.0 and the MSD structure. It uses 21 main concepts, with limited sub-items, derived from the SDMX list of cross-domain concepts (please compare Table 8). Most of the reference metadata in the ESMS are currently inserted as free text, although it is intended that some of them may follow a code list (e.g. frequency, or reference area).



1. Contact	8. Release policy	15. Timeliness and punctuality
2. Metadata update	9. Frequency of dissemination	16. Comparability
3. Statistical presentation	10. Dissemination format	17. Coherence
4. Unit of measure	11. Accessibility of documentation	18. Cost and burden
5. Reference period	12. Quality management	19. Data revision
6. Institutional mandate	13. Relevance	20. Statistical processing
7. Confidentiality	14. Accuracy and reliability	21 Comment

Table 8 - Metadata concepts

Codelists

In most cases the values for a dimension are defined in a code list. Each value on that list is given a language-independent abbreviation - the code (= DE) and a language-specific description (= Deutschland - in German or Allemagne - in French).

Attribute values can be represented with codes or only by free-text values, especially when they are associated to a higher level - e.g. the whole data set. This is also in the view that the purpose of an attribute is solely to describe & qualify and not to identify the data in a unique key.

An example of the Frequency (FREQ) code list is presented in Figure 5 – Example of cross-domain code list.

3.2.4 SDMX formats

SDMX-IM provides a way of modelling statistical data formats in two syntaxes guaranteeing that transformation between the different formats is lossless: SDMX-EDI and SDMX-ML (concering the features of SDMX standard version 1.0).

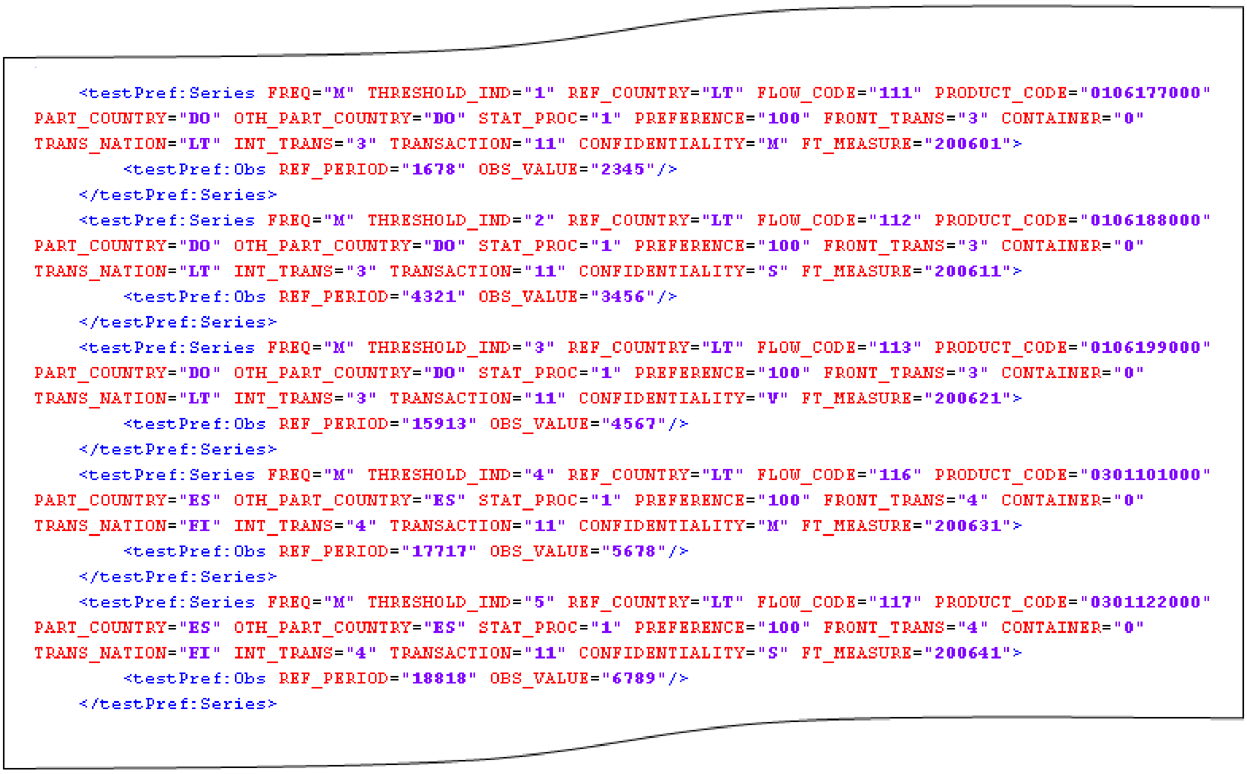
SDMX-ML, according to SDMX standard version 2.0, includes extended features, like for example the support for Metadata Structure Definitions and Metadata Messages and the interaction with an SDMX Registry / Repository.

Unlike SDMX-EDI, SDMX-ML offers different XML messages types for the exchange of data: Generic data message, Compact data message, Utility data message and Cross-sectional data message. The main characteristics for both formats are the following:

SDMX-EDI	SDMX-ML
SDMX-EDI format is a proprietary format based on the EDIFACT grammar and is fully compatible with GESMES/TS version 3.0 standard. It allows an efficient batch exchange of large amounts of data between systems or organisations.	SDMX-ML messages/files are based on XML. It allows standard XML web-based applications dealing with limited amounts of data, but taking advantage of XML open format flexibility and its multitude of functionalities.

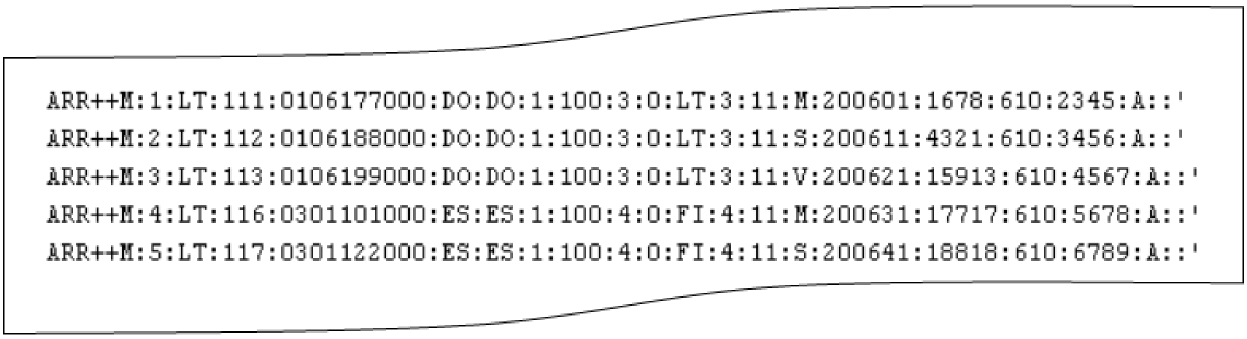
Table 9 - SDMX formats - main characteristics

The following two figures present examples of SDMX-ML and SDMX-EDI data files. More details on the SDMX formats and particularly on the SDMX-ML messages are provided in the student book 3 – Data Structure Definition.



```
    <testPref:Series FREQ="M" THRESHOLD_IND="1" REF_COUNTRY="LT" FLOW_CODE="111" PRODUCT_CODE="0106177000"
PART_COUNTRY="DO" OTH_PART_COUNTRY="DO" STAT_PROC="1" PREFERENCE="100" FRONT_TRANS="3" CONTAINER="0"
TRANS_NATION="LT" INT_TRANS="3" TRANSACTION="11" CONFIDENTIALITY="M" FT_MEASURE="200601">
        <testPref:Obs REF_PERIOD="1678" OBS_VALUE="2345"/>
    </testPref:Series>
    <testPref:Series FREQ="M" THRESHOLD_IND="2" REF_COUNTRY="LT" FLOW_CODE="112" PRODUCT_CODE="0106188000"
PART_COUNTRY="DO" OTH_PART_COUNTRY="DO" STAT_PROC="1" PREFERENCE="100" FRONT_TRANS="3" CONTAINER="0"
TRANS_NATION="LT" INT_TRANS="3" TRANSACTION="11" CONFIDENTIALITY="S" FT_MEASURE="200611">
        <testPref:Obs REF_PERIOD="4321" OBS_VALUE="3456"/>
    </testPref:Series>
    <testPref:Series FREQ="M" THRESHOLD_IND="3" REF_COUNTRY="LT" FLOW_CODE="113" PRODUCT_CODE="0106199000"
PART_COUNTRY="DO" OTH_PART_COUNTRY="DO" STAT_PROC="1" PREFERENCE="100" FRONT_TRANS="3" CONTAINER="0"
TRANS_NATION="LT" INT_TRANS="3" TRANSACTION="11" CONFIDENTIALITY="V" FT_MEASURE="200621">
        <testPref:Obs REF_PERIOD="15913" OBS_VALUE="4567"/>
    </testPref:Series>
    <testPref:Series FREQ="M" THRESHOLD_IND="4" REF_COUNTRY="LT" FLOW_CODE="116" PRODUCT_CODE="0301101000"
PART_COUNTRY="ES" OTH_PART_COUNTRY="ES" STAT_PROC="1" PREFERENCE="100" FRONT_TRANS="4" CONTAINER="0"
TRANS_NATION="FI" INT_TRANS="4" TRANSACTION="11" CONFIDENTIALITY="M" FT_MEASURE="200631">
        <testPref:Obs REF_PERIOD="17717" OBS_VALUE="5678"/>
    </testPref:Series>
    <testPref:Series FREQ="M" THRESHOLD_IND="5" REF_COUNTRY="LT" FLOW_CODE="117" PRODUCT_CODE="0301122000"
PART_COUNTRY="ES" OTH_PART_COUNTRY="ES" STAT_PROC="1" PREFERENCE="100" FRONT_TRANS="4" CONTAINER="0"
TRANS_NATION="FI" INT_TRANS="4" TRANSACTION="11" CONFIDENTIALITY="S" FT_MEASURE="200641">
        <testPref:Obs REF_PERIOD="18818" OBS_VALUE="6789"/>
    </testPref:Series>
```

Figure 2 - SDMX-ML Compact data message example (data part)



```
ARR++M:1:LT:111:0106177000:DO:DO:1:100:3:0:LT:3:11:M:200601:1678:610:2345:A::'
ARR++M:2:LT:112:0106188000:DO:DO:1:100:3:0:LT:3:11:S:200611:4321:610:3456:A::'
ARR++M:3:LT:113:0106199000:DO:DO:1:100:3:0:LT:3:11:V:200621:15913:610:4567:A::'
ARR++M:4:LT:116:0301101000:ES:ES:1:100:4:0:FI:4:11:M:200631:17717:610:5678:A::'
ARR++M:5:LT:117:0301122000:ES:ES:1:100:4:0:FI:4:11:S:200641:18818:610:6789:A::'
```

Figure 3 - SDMX-EDI / GESMES data message example (data part)

3.3 Content oriented guidelines

Content oriented guidelines are a set of recommendations within the scope of the SDMX standard. The aim is to produce maximum interoperability in the exchange of data and metadata between organisations. Their use is encouraged where possible across statistical domains in the following three areas:

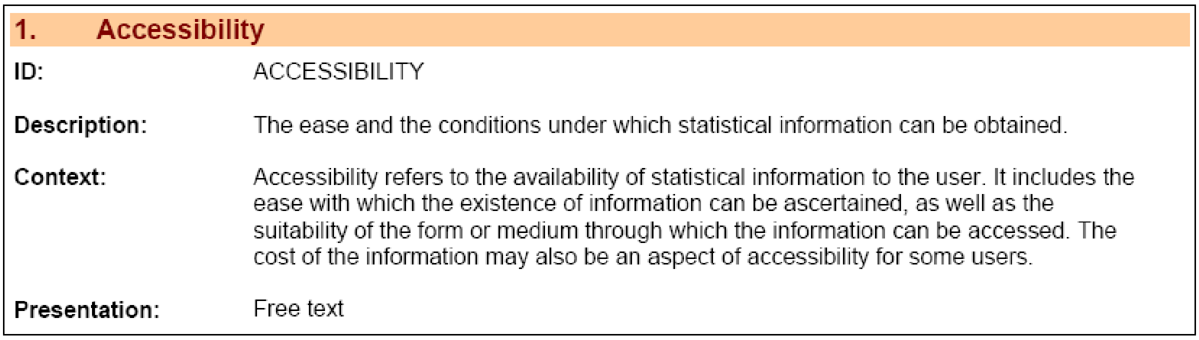
- Cross-domain concepts.
- Statistical subject-matter domains.
- A Metadata common vocabulary.

The three areas are shortly presented in next sub-chapters.

3.3.1 Cross-domain concepts

This guideline contains a list of statistical concepts, inter alia related to statistical processes and data quality. This list is based on the concepts used by the contributing international organisations. It is not exhaustive and it is expected to grow in the future.

The concepts can be used at the data as well as at the metadata sides. Each concept has a unique identifier (ID), a description, the context in which the concept may be used and its presentation in the SDMX standard. Figure 4 is an example of cross-domain concepts.



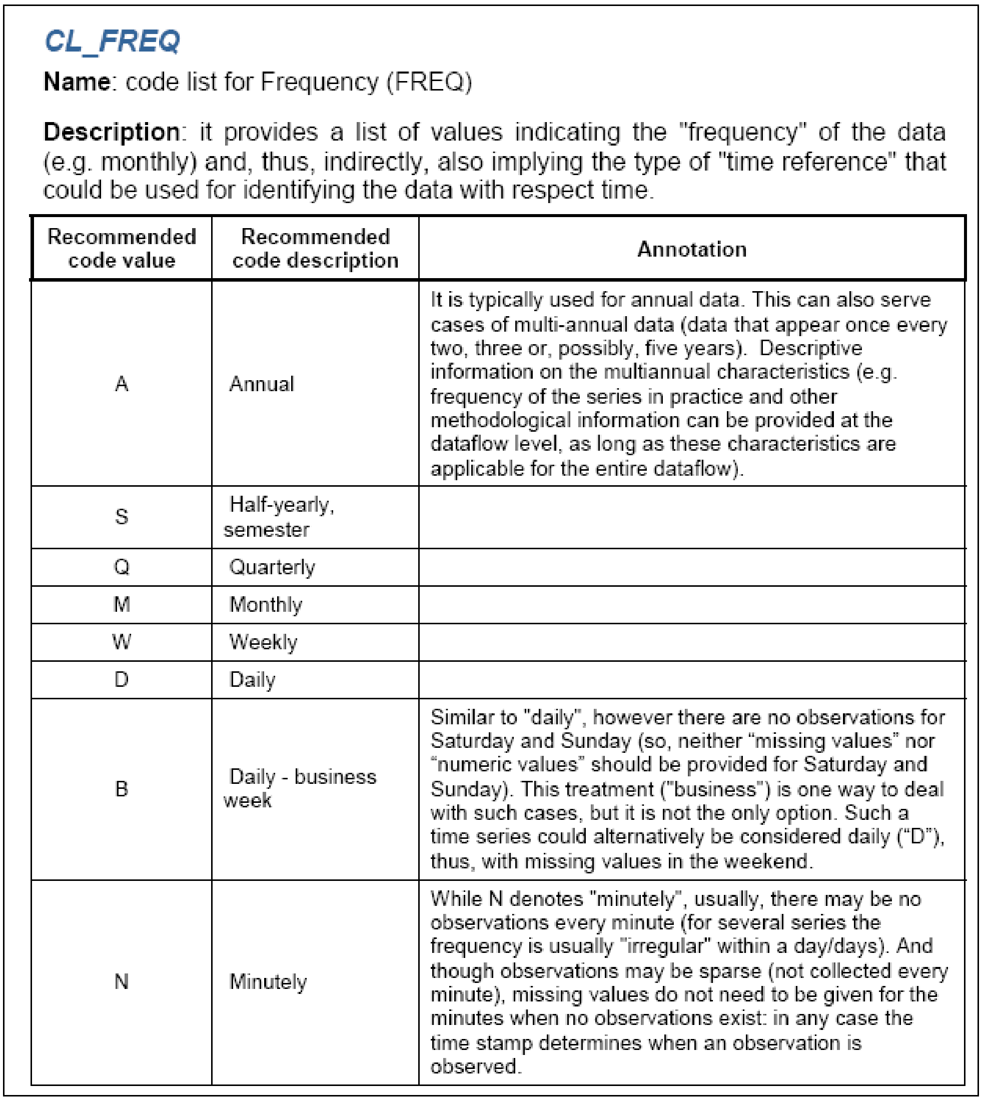
1. Accessibility

ID:	ACCESSIBILITY
Description:	The ease and the conditions under which statistical information can be obtained.
Context:	Accessibility refers to the availability of statistical information to the user. It includes the ease with which the existence of information can be ascertained, as well as the suitability of the form or medium through which the information can be accessed. The cost of the information may also be an aspect of accessibility for some users.
Presentation:	Free text

Figure 4 – Example of cross-domain concepts[1]

Cross-domain concepts, like all concepts, may have a 'CODE LIST' as presentation. This means that they have a limited set of possible values enumerated in code lists. Those code lists are called cross-domain code lists. Figure 5 is an example of cross-domain code-list for the frequency. Code lists have a general description, a list of codes, their descriptions and annotations. Annotations provide additional information on the codes.

[1] Reference source: www.sdmx.org (extraction performed in March 09).



CL_FREQ

Name: code list for Frequency (FREQ)

Description: it provides a list of values indicating the "frequency" of the data (e.g. monthly) and, thus, indirectly, also implying the type of "time reference" that could be used for identifying the data with respect time.

Recommended code value	Recommended code description	Annotation
A	Annual	It is typically used for annual data. This can also serve cases of multi-annual data (data that appear once every two, three or, possibly, five years). Descriptive information on the multiannual characteristics (e.g. frequency of the series in practice and other methodological information can be provided at the dataflow level, as long as these characteristics are applicable for the entire dataflow).
S	Half-yearly, semester	
Q	Quarterly	
M	Monthly	
W	Weekly	
D	Daily	
B	Daily - business week	Similar to "daily", however there are no observations for Saturday and Sunday (so, neither "missing values" nor "numeric values" should be provided for Saturday and Sunday). This treatment ("business") is one way to deal with such cases, but it is not the only option. Such a time series could alternatively be considered daily ("D"), thus, with missing values in the weekend.
N	Minutely	While N denotes "minutely", usually, there may be no observations every minute (for several series the frequency is usually "irregular" within a day/days). And though observations may be sparse (not collected every minute), missing values do not need to be given for the minutes when no observations exist: in any case the time stamp determines when an observation is observed.

Figure 5 – Example of cross-domain code list[2]

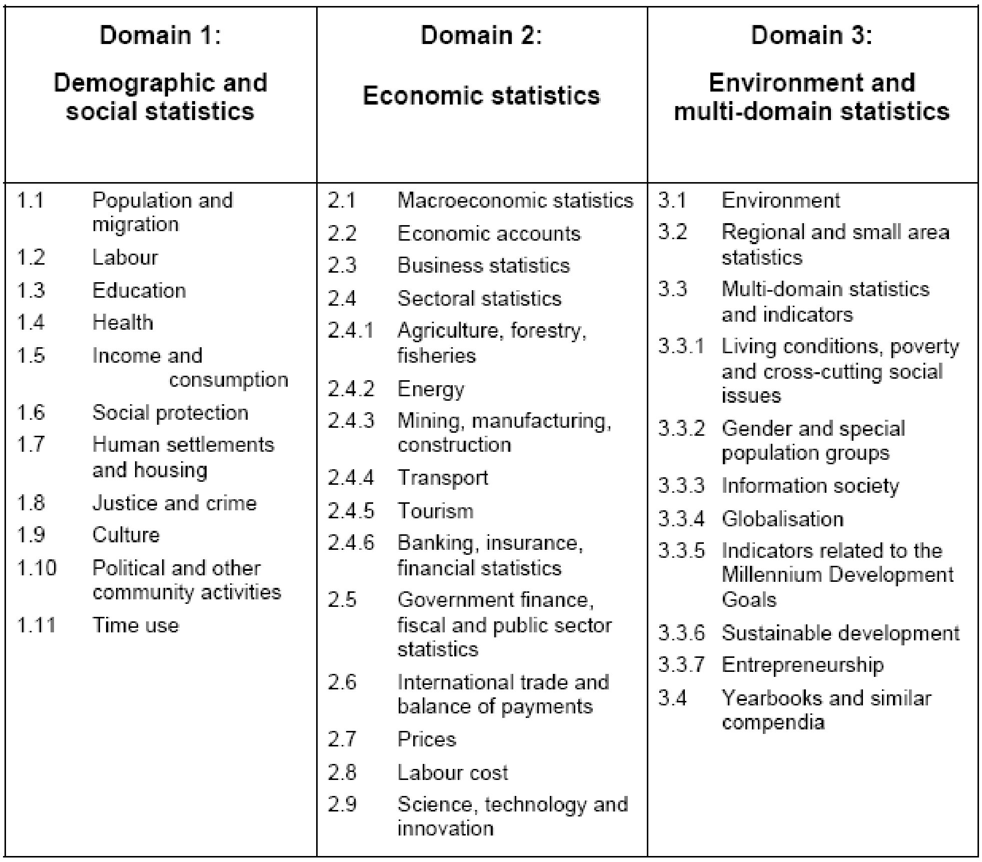
3.3.2 Statistical subject-matter domains

'Statistical subject-matter domains' is a high level classification - based on the work of the United Nations Economic Commission for Europe (UNECE) - of statistical areas. The classification offers a starting point for organising the exchange of statistical data and metadata.

A statistical subject-matter domain refers to a statistical activity that has common characteristics with respect to variables, concepts and methodologies for data collection and the whole statistical data compilation process. Examples of statistical domains are price statistics, national accounts, environment statistics or education statistics.

The SDMX statistical subject-matter domain list is intended to cover the universe of statistical information handled by a large number of international organisations and national agencies, often referred to as official statistics. Official statistics constitute the basic information system of a society. Figure 6 lists the subject-matter domains.

[2] Reference source: www.sdmx.org (extraction performed in March 09).



Domain 1: Demographic and social statistics	Domain 2: Economic statistics	Domain 3: Environment and multi-domain statistics
1.1 Population and migration 1.2 Labour 1.3 Education 1.4 Health 1.5 Income and consumption 1.6 Social protection 1.7 Human settlements and housing 1.8 Justice and crime 1.9 Culture 1.10 Political and other community activities 1.11 Time use	2.1 Macroeconomic statistics 2.2 Economic accounts 2.3 Business statistics 2.4 Sectoral statistics 2.4.1 Agriculture, forestry, fisheries 2.4.2 Energy 2.4.3 Mining, manufacturing, construction 2.4.4 Transport 2.4.5 Tourism 2.4.6 Banking, insurance, financial statistics 2.5 Government finance, fiscal and public sector statistics 2.6 International trade and balance of payments 2.7 Prices 2.8 Labour cost 2.9 Science, technology and innovation	3.1 Environment 3.2 Regional and small area statistics 3.3 Multi-domain statistics and indicators 3.3.1 Living conditions, poverty and cross-cutting social issues 3.3.2 Gender and special population groups 3.3.3 Information society 3.3.4 Globalisation 3.3.5 Indicators related to the Millennium Development Goals 3.3.6 Sustainable development 3.3.7 Entrepreneurship 3.4 Yearbooks and similar compendia

Figure 6 – SDMX list of statistical subject-matter domains- overview[3]

The list of Statistical Subject-Matter Domains has three functions:

- As a standard scheme against which similar domain lists of national and international organisations can be mapped to facilitate the exchange of data and metadata.
- As an identifier framework for registering and searching statistical data on SDMX registries, the architecture of which has been developed in SDMX Technical Standards Version 2.0.
- As a navigation aide for identification and organisation of corresponding 'domain groups' playing an active role in the use of SDMX technical standards and content oriented guidelines for the exchange of statistics and related metadata.

3.3.3 Metadata Common Vocabulary

The Metadata Common Vocabulary (MCV) contains concepts and related definitions used in structural and reference[4] metadata of international organisations and national data producing

[3] Reference source: www.sdmx.org (extraction performed in March 09).

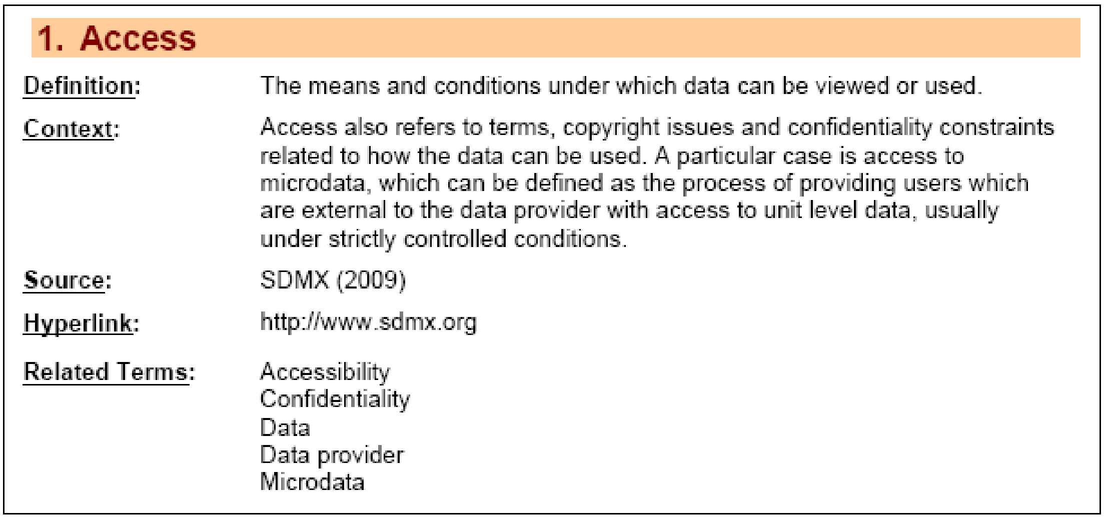
agencies. The MCV is a vocabulary that recommends a common terminology to be used in order to facilitate communication and understanding.

The MCV is closely linked to the cross-domain concepts as it also contains all these concepts, stating their definitions and context descriptions.

The MCV covers a selected range of metadata concepts:

- **General metadata concepts** - mostly derived from ISO, UNECE and UN documents, useful for providing a general context to metadata (for example: classification, metadata registry, statistical metadata, statistical production).
- **Metadata terms describing statistical methodologies and data quality** (for example: frequency, data collection method, data revision, source, adjustment, accuracy, or the metadata and quality frameworks adopted by international organisations.
- **Terms referring specifically to data and metadata exchange** (for example: bilateral exchange or gateway exchange).

Figure 7 depicts an example of MCV.



1. Access

Definition: The means and conditions under which data can be viewed or used.

Context: Access also refers to terms, copyright issues and confidentiality constraints related to how the data can be used. A particular case is access to microdata, which can be defined as the process of providing users which are external to the data provider with access to unit level data, usually under strictly controlled conditions.

Source: SDMX (2009)

Hyperlink: http://www.sdmx.org

Related Terms: Accessibility
Confidentiality
Data
Data provider
Microdata

Figure 7 – Example of MCV[5]

3.3.4 What does interoperability mean?

Interoperability is a property referring to the ability of diverse systems and organisations to work together (inter-operate). The term is often used in a technical systems engineering sense, or alternatively in a broad sense, taking into account social, political, and organisational factors that impact system to system performance. There are 2 levels of interoperability:

[4] 'Structural' metadata define the structure of statistical data and metadata whereas 'reference' metadata describe the actual content of metadata.

[5] Reference source: www.sdmx.org (extraction performed in March 09).

- **Syntactic Interoperability**: If two or more systems are capable of communicating and exchanging data, they are exhibiting syntactic interoperability. Specified data formats, communication protocols and the like are fundamental. In general, XML or EDIFACT standards provide syntactic interoperability. Syntactical interoperability is required for any attempts of further interoperability.

- **Semantic Interoperability**: Beyond the ability of two or more computer systems to exchange information, semantic interoperability is the ability to automatically interpret the information exchanged meaningfully and accurately in order to produce useful results as defined by the end users of both systems. To achieve semantic interoperability, both sides must defer to a common information exchange reference model. The content of the information exchange requests are unambiguously defined: what is sent is the same as what is understood.

3.4 IT architecture for data exchange

3.4.1 Standard formats for the exchange of data and metadata

Based on the common information model, the SDMX standards include data exchange formats based on XML (SDMX-ML) and EDIFACT (SDMX-EDI, which is identical to GESMES/TS and which has been widely used since the 1990s).

The advantage of the XML syntax is that it is a widely-used open standard, which can be processed with a wide range of IT applications, including free and/or open-source software. The EDIFACT syntax is more specialised (e.g. appropriate for representing large databases, due to its compact format) and is usually processed with proprietary applications.

The use of SDMX-ML, which supports the full Information model of SDMX 2.0, provides additional benefits.

3.4.2 Architectures for data exchange

SDMX, besides describing and specifying technical standards (the Information Model, message formats for data and metadata, Registry service definitions), comprises an IT architecture to be used for the efficient exchange and sharing of statistics.

For this purpose, SDMX identifies three basic process patterns (bilateral, gateway and data-sharing) and two modes (push and pull) regarding the exchange of statistical data and metadata.

In the data-sharing model a group of partners agree on providing access to their data according to standard processes, formats and technologies.

In the pull mode, the data consumer retrieves the data from the provider's web server. The data may be made available for download in an SDMX-conformant file, or they may be retrieved from a database in response to an SDMX-conformant query, via a web service running on the provider's server. In both cases, the data are made available to any organisation requiring them, in formats which ensure that data are consistently described by appropriate metadata, whose meaning is common to all parties in the exchange.

Data sharing using the pull mode is well adapted to the database-driven and data hub architectures. Both architectures provide the best benefits for the data producers because they can lessen the burden of publishing the data to multiple counterparties.

In both architectures, it is necessary to implement a notification mechanism, providing provisioning metadata in order to alert collecting organisations that data and metadata sets are made available by data providers. Details about the online mechanism for getting data (for example, a queryable online database or a simple URL) and constraints regarding the allowable content of the data sets that will be provided.

At the heart of a data-sharing architecture there is often an SDMX Registry. This is a central location where structural and provisioning metadata can be found. In fact all the users/applications that need to access data can query the registry in order to know what data sets and metadata sets are available from data providers, and how to access them.

3.4.2.1 Push and Pull mode

SDMX supports two complementary modes for data exchange and data sharing:

The 'push' mode (Figure 8) - data is transmitted from one organisation to another. This mode means that the data provider takes action to send the data to the organisation collecting the data. This can take place using different means, such as e-mail or file transfer. These are the traditional modes of data collection, carried out by international organisations for many years.

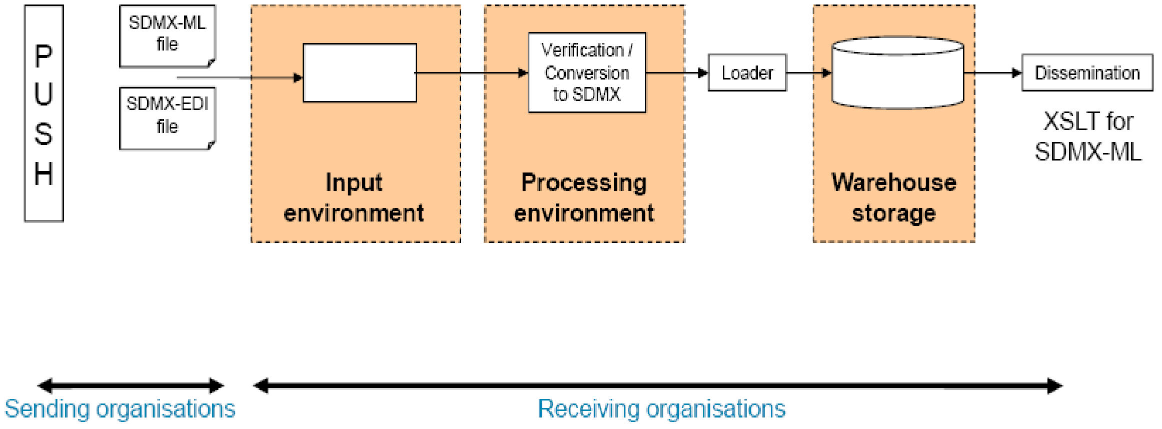


Figure 8 – Push mode

The 'pull' mode (Figure 9) implies that the data provider makes the data available for the users via Internet technology:

- For download in a SDMX-conformant file;
- As a result from SDMX-conformant query to a web service linked to a database on the provider's side.

In both cases, the data are made available to any organisation requiring them, in formats which ensure that data are consistently described by appropriate metadata, whose meaning is common to all parties in the exchange.

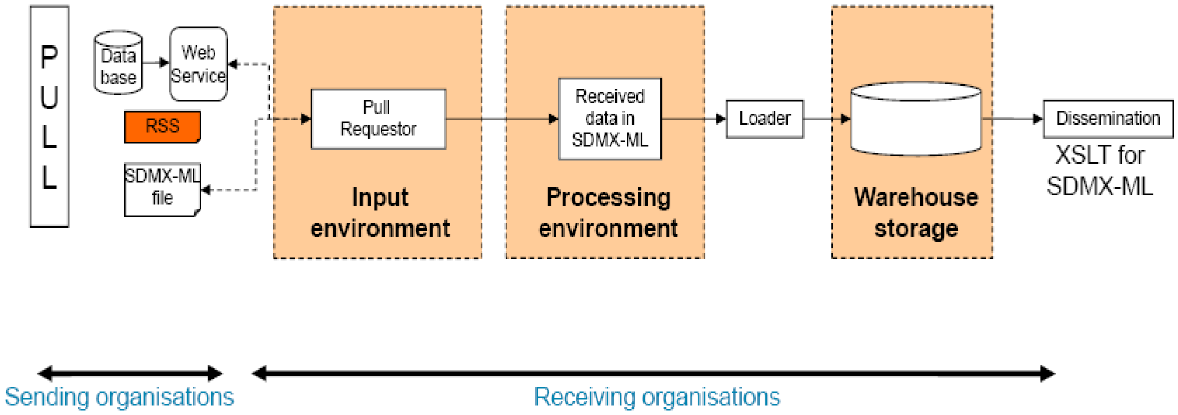


Figure 9 – Pull mode

3.4.2.2 Database driven architecture and Hub concept (architecture)

The database-driven architecture is implemented by those collecting organisations that periodically need to fetch the data and to load them in their database. In general a batch process is used in order to automate the flow in which a whole or a partial dataset, including incremental updating, is used.

The pull approach within a database-driven architecture includes the following steps based on a provision agreement:

1) When new data is available, the data provider creates an SDMX-ML file containing the new data set **OR** provide a web service (WS) that builds SDMX-ML messages upon request. Notification to data consumers about the new data and the details on how to obtain them can be performed with an RSS web feed.

2) The data collector Pull Requestor reads the new RSS feed entry (or receives the information on the new data by other means. He can now retrieve the SDMX-ML file from the specified URL OR use the 'Query Message' included in the RSS feed to query the data provider's web service.

Figure 10 represents the database-driven architecture.

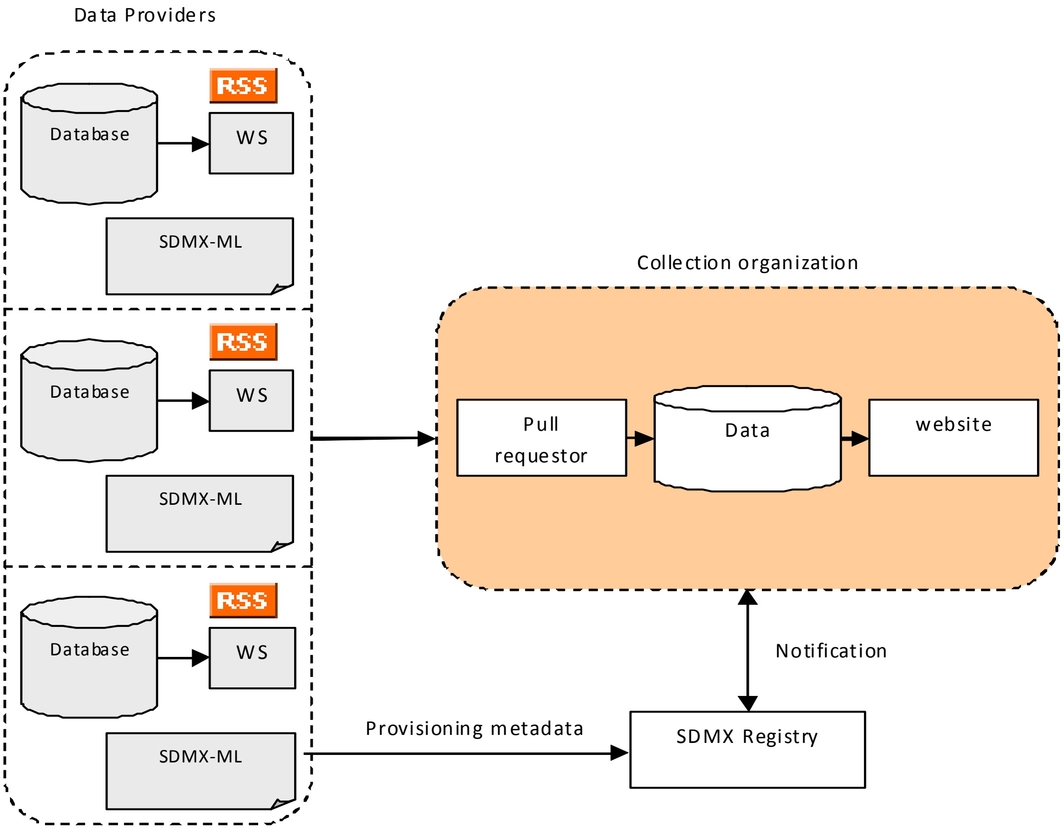


Figure 10 - Database driven architecture[6]

[6] Reference source: www.sdmx.org (extraction performed in March 09).

SDMX also supports the ‘Data Hub’ concept/architecture, where users obtain data from a central hub which itself automatically assembles the required dataset by querying other data sources.

Data providers can notify the hub of new sets of data and corresponding structural metadata (measures, dimension, code lists, etc.) and make data available directly from their systems through querying means.

Data users can browse the hub to define a dataset of interest via the above structural metadata and retrieve the desired dataset.

From the data management point of view, the hub is also based on a specific datasets, which are - contrary to the database driven architecture - not kept locally at the central hub system. Instead the following process operates:

1) A user identifies a dataset through the web interface of the central hub using the structural metadata, and requests it.
2) The central hub translates the user request in one or more queries and sends them to the related data providers’ systems.
3) Data providers’ systems process the query and send the result to the central hub in standard format.
4) The central hub puts together all the results originated by all interested data providers’ systems and presents them in a human readable format.

Figure 11 represents the Data Hub architecture.

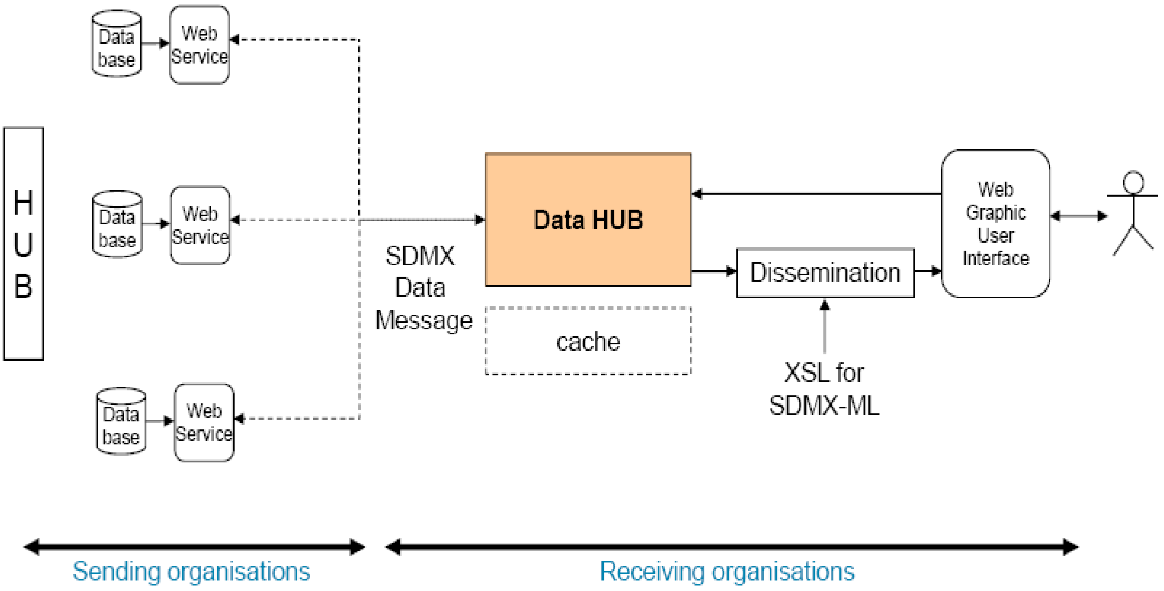


Figure 11 – Data Hub concept

3.4.3 SDMX registry

The SDMX IT architecture also comprises SDMX registries. In fact, a SDMX Registry plays an important role in this architecture; it can be seen as a central application which is accessible to other programs over the Internet (or an Intranet or Extranet) to provide information needed to facilitate statistical tasks.

SDMX has developed specific registry standards in order to enable statistical organisations to perform efficient data and metadata sharing. In general terms, an SDMX registry is essentially an application which can accept SDMX query messages and return the locations of SDMX-compliant information, which may include data as well as structural and reference metadata.

SDMX Registry functions include to provide information about:

- What data sets and metadata sets are available, and where they are located.
- How the data sets and metadata sets are provided: how often they are updated, what their contents are, how they can be accessed, and similar questions.
- The structure of data sets and metadata sets, answering questions like: What code lists do they use? What concepts are involved?

In addition, it allows applications to sign up (or subscribe) for notifications, so that when a data set or metadata set of interest becomes available, the application will be automatically alerted.

Thus it can be seen as the index of a distributed database or metadata repository which is made up of all the data provider's data sets and reference metadata sets within a statistical community.

3.5 IT tools

3.5.1 Introduction

This chapter explains how the freely available SDMX IT tools can be used in the implementation of SDMX standards in local IT systems.

Adoption of the SDMX technical standards and guidelines by organisations involved in statistical data and metadata exchange will need to result in actual implementation of the standards in the local statistical IT systems.

Developing specific SDMX-based IT functionality is a task, common to all organisations in this context. As a natural consequence efficiency gains can be achieved by offering to share one's developments with each other. As part of its advocacy activities SDMX promotes the provision of freely available tools and open source software products that offer functionality needed for this implementation of the SDMX standards.

Various types of tools for different purposes and target audiences are available: demonstration tools, production modules and production applications.

3.5.2 Types of Tools

SDMX demonstration tools help statistical as well as IT experts to understand the basic principles of SDMX and allow them to execute (on a small scale) the different types of functionality required for setting up an 'SDMX capable' statistical processing system. They are meant for educational and demonstration purposes and are often used in SDMX courses.

- A 'DSD Builder Tool', for example, can be used to take the first steps in understanding DSDs when practicing how to define data structures for data that are be processed in the institution. Loading freely available public DSDs into the tool will allow one to analyse the relationships between different data structures.
- Similar examples of other functionalities are the conversion between different SDMX technical formats or the creation of actual SDMX data and metadata files.

SDMX production modules provide specific functionality required in SDMX-based statistical processing at an 'industrial strength' level. This means that they could be directly integrated into a production application to perform a specific task in the processing workflow.

Examples are the 'SDMX checker suite' developed by the ECB and the 'SDMX converter' offered by Eurostat. The SDMX production modules are targeted to IT experts who need to implement SDMX functionality. The modules can considerably shorten the required development time for an individual organisation implementing SDMX in its local systems.

SDMX production applications are a set of production modules, usually offered by one organisation, that cover a wider range of SDMX statistical processing functionality in an integrated way. They have been developed by that organisation for its own internal use and it can thus be expected that the application is actually used for production purposes in that organisation. The target audience are business and IT experts tasked to implement an SDMX production system in an organisation. They will want to evaluate such applications from both the business and the IT 'fit' for their own workflow and application environment.

3.5.3 Availability of SDMX Tools

The existing SDMX tools have been developed (or commissioned) by SDMX sponsors and other organisations actively involved in implementing SDMX and in general made available under open source licences.

The growing SDMX user community is encouraged to evaluate and test the already existing tools and modules and to contribute and cooperate for its further development. Those wishing to contribute their developments under an open source licence are invited to contact the SDMX Secretariat[7].

The idea behind the concept: Starting from the freely available open source tools and modules could result in cases of even closer cooperation between different organisations in the form of shared development of statistical applications. This may lead to the harmonisation of not only of the statistical data exchange, but also of the statistical processing applied in different organisations.

Up-to-date Information about available tools can be found via the Tools page on the SDMX website www.sdmx.org, from which the user is redirected to a site from which they can actually be downloaded and additional information is available. A useful source is also the web site www.osor.eu (Open Source Observatory and Repository)

[7] secretariat@sdmx.org

Table 10 presents a selection of currently available SDMX tools:

SDMX Tools
SDMX Registry (Eurostat) - SDMX-compliant Metadata Registry with a web based user interface and a web service for interacting with partners, thus the components of the SDMX registry are: • Database • Web service • Web user interface
Data Structure Wizard (Eurostat) Desktop application designed to work with SDMX-compliant registries for editing and viewing SDMX structural metadata objects; Create, Edit and View Data Structure Definitions (DSD builder tool); Import and Export related artefacts; Operates offline or online (= connected to the SDMX registry).
SDMX Converter (Eurostat) Converts between different data file formats of SDMX 2.0 standard as well as GESMES and CSV formats.
SDMX Visualisation Tools (Eurostat) General-purpose visualisation tools for SDMX data files.
Cycle Clock - Visualise business cycles (Eurostat) Visualises business cycles by analysing time series of financial and economic statistical data from SDMX files.
ECB Visualisation Framework Libraries (API) that can be used to build tools to visualise statistical data and metadata expressed in SDMX-ML.
ECB Checker A tool for reading and checking SDMX-EDI data files.
Data Structure Definition (Key Family) Database Tool (Metadata technology) Load SDMX files in databases (MSAccess).
Metadata Structure Definition Editor (Metadata technology) Create, Edit and View Metadata Structure Definitions.
SDMX Transformations Package (Metadata technology) Validate SDMX files; Convert files from/to SDMX.
Excel SDMX Authoring Tool (Metadata technology) Visualise the content of SDMX files

Table 10 – Selection of currently available SDMX tools

4 Learn more about SDMX

The main Sources about SDMX are represented by:

The Official website: www.sdmx.org

- SDMX User guide: Version 2009.1;
- SDMX Standard Specifications: 2.0 and 1.0;
- Content-Oriented Guidelines: 2009;
- SDMX tools section;
- Overview of SDMX reference implementations;
- Announcement of conferences & other SDMX related events;
- Links to further web sources, for example to SDMX partners or the web site www.osor.eu (Open Source Observatory and Repository)

5 Glossary

Table 11 presents the list of concepts and acronyms with their definition.

Concept	Definition
DSD	Data Structure Definition
ECB	European Central Bank
EDIFACT	Electronic Data Interchange for Administration, Commerce and Transport
ESMS	Euro SDMX Metadata Structure
GESMES/TS	GESMES Time Series data exchange message
ISO	International Organisation for Standardisation
IT	Information Technology
MSD	Metadata Structure Definition
MCV	Metadata Common Vocabulary
NACE	National Classification of Economic Activities
RSS	Really Simple Syndication (*also used Rich Site Summary*) - family of Web feed formats to publish frequently updated information
SDMX	Statistical Data and Metadata eXchange.
SDMX-EDI	SDMX Electronic Data Interchange - EDIFACT format for exchange of SDMX-structured data and metadata
SDMX-IM	SDMX Information Model
SDMX-ML	SDMX Markup Language - XML format for the exchange of SDMX-structured data and metadata
UML	Unified Modelling Language
UN	United Nations
UNECE	United Nations Economic Commission for Europe
XML	EXtensible Markup Language

Table 11 - Glossary