



Conference on Data Quality for International Organisations
Newport, Wales (UK), 27-28 April 2006

The use of SDMX standards for retrieving metadata from different providers

Marco Pellegrino, Eurostat
marco.pellegrino@cec.eu.int



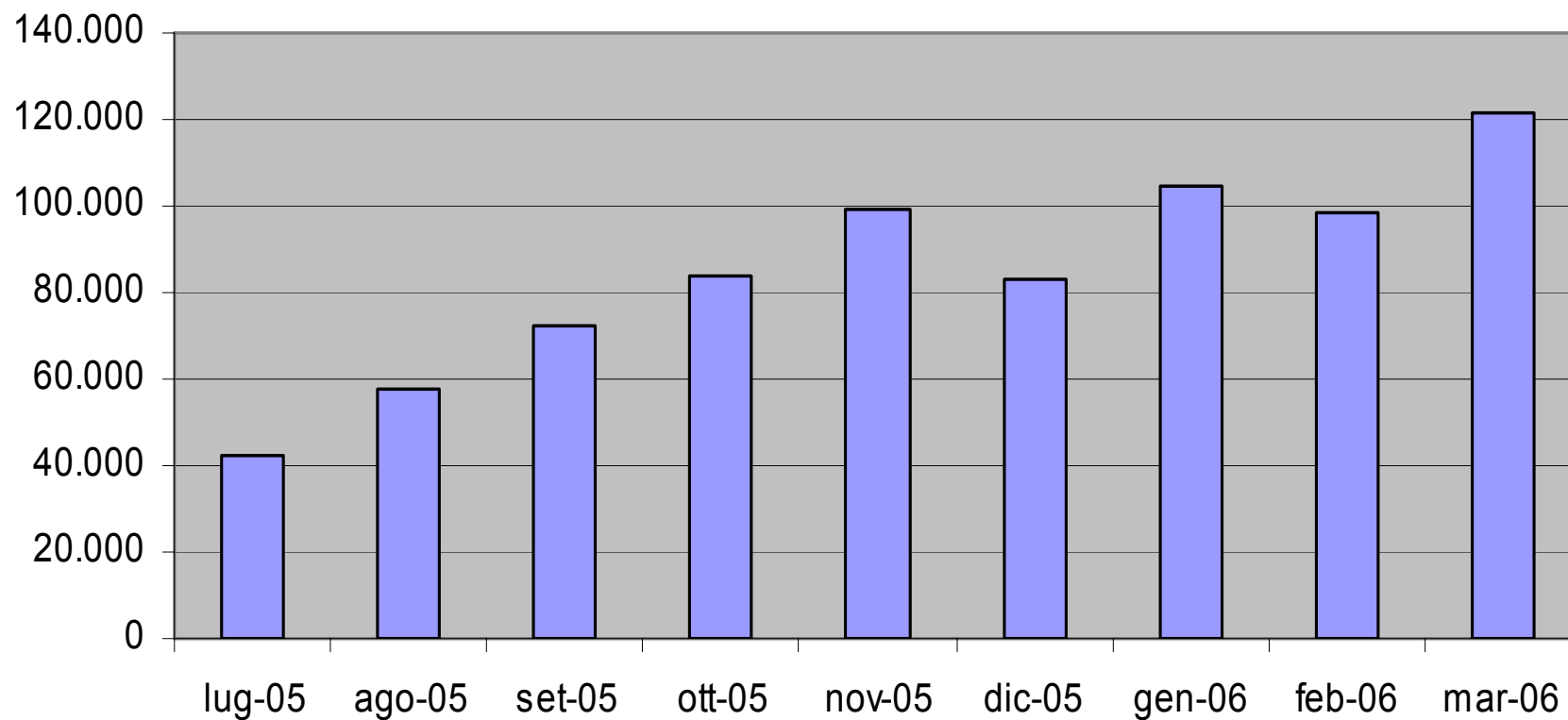
Eurostat dissemination policy

In Eurostat, since October 2004, metadata are...

- ❑ Standardised across domains
- ❑ Linked to a common terminology (MCV)
- ❑ Targeted at generalist readers; extensions for expert readers
- ❑ Attached at different levels of the data cube
- ❑ Monitored and regularly updated
- ❑ Open to public consultation (accountability: for each domain or sub-domain, one contact unit is identified as responsible for maintaining the information and for users' feedback)

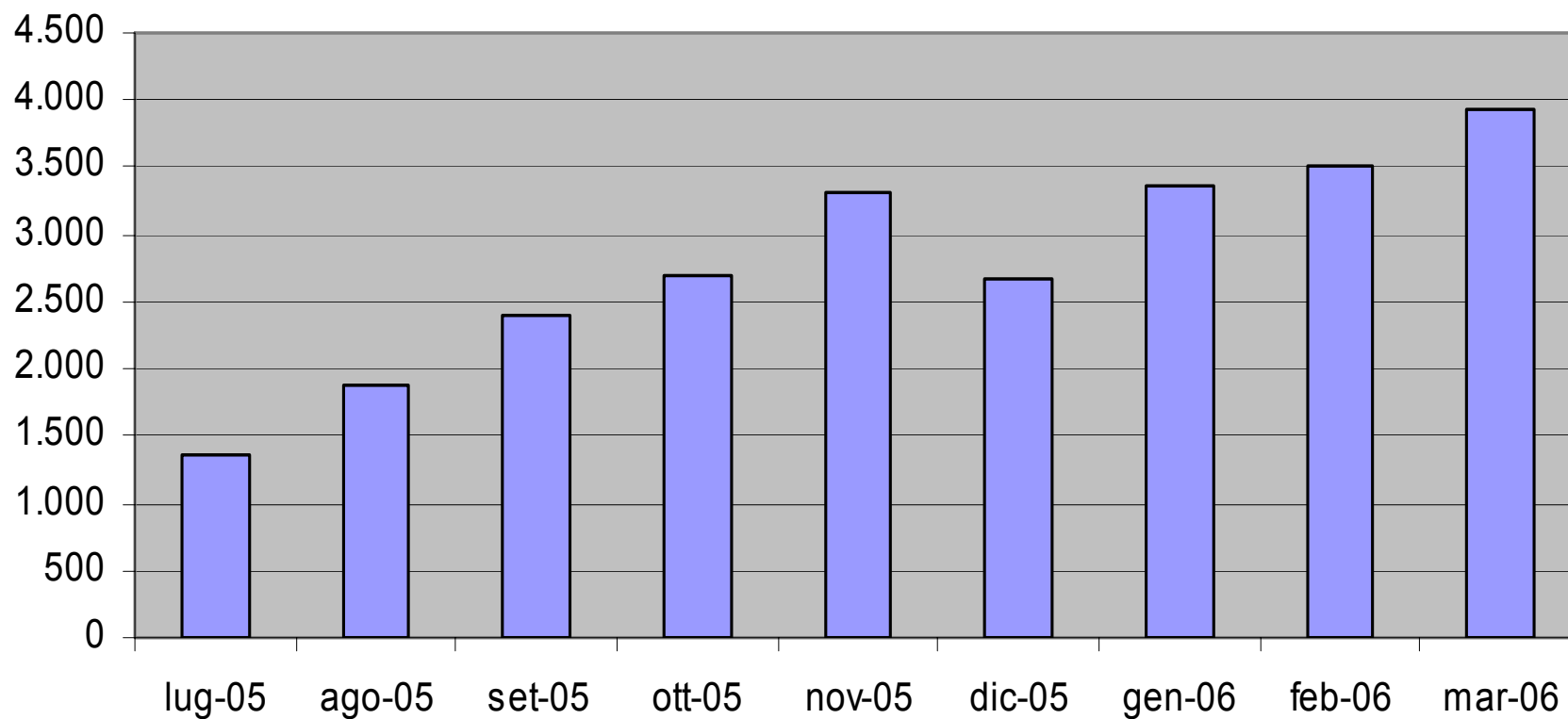


Eurostat: Internet monthly consultation of reference metadata files





Eurostat: Internet daily consultation of reference metadata files





Strengths and Weaknesses

Strengths

- ❑ Unprecedented discipline
- ❑ Comparability across domains and over time
- ❑ Higher integration with some national metadata
- ❑ Visibility and external pressure

Weaknesses

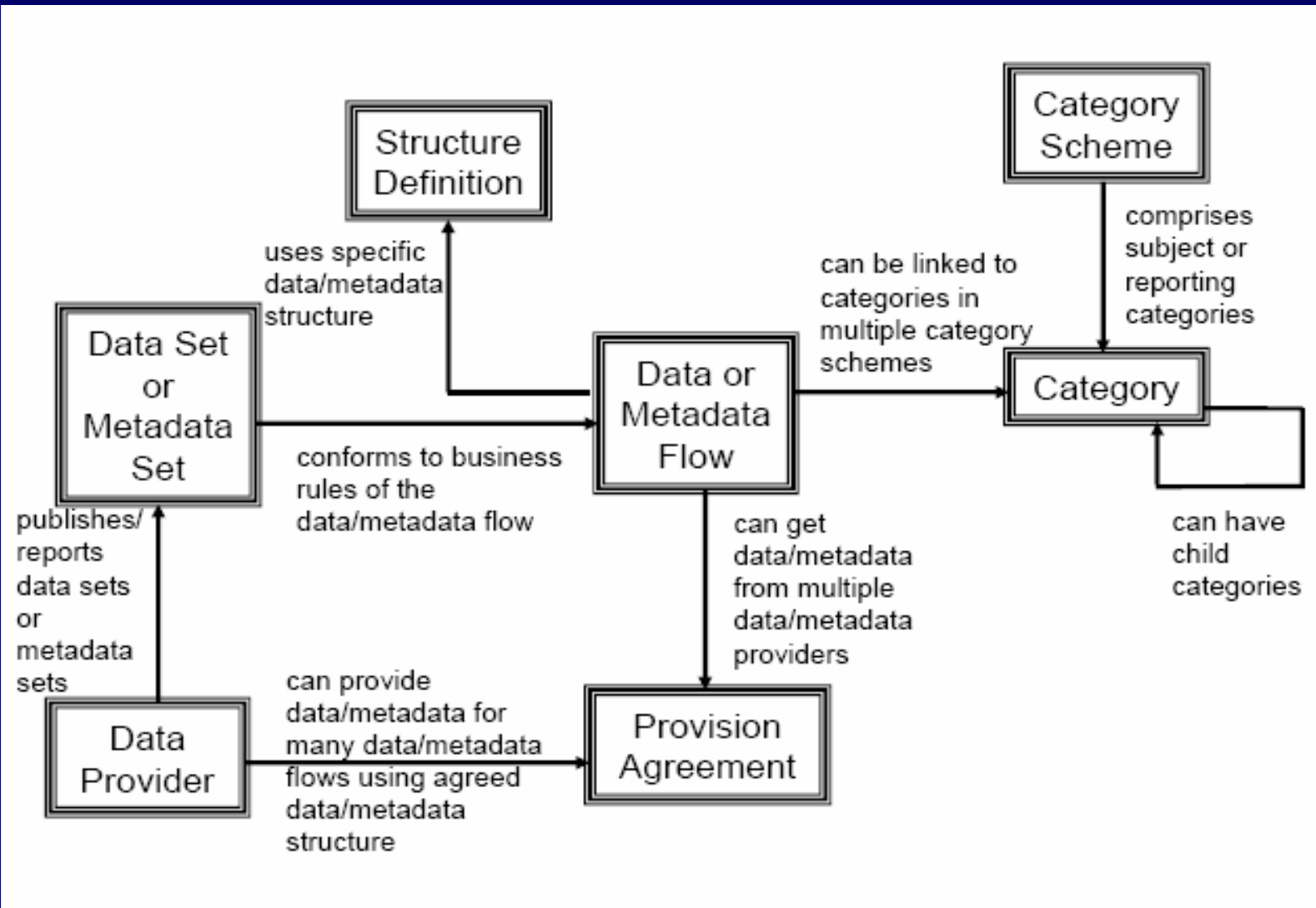
- ❑ Time-consuming monitoring and updating procedures
- ❑ High training and documentation needs (underestimated?)
- ❑ Incomplete format (first self-assessment)

Eurostat: Metadata Format for Free Dissemination

Base Page	Methodology
General information Geographical area Statistical domain Contact information	Concepts, definitions and classifications Statistical concept Definition of indicators Classification system used
Data Data description Time coverage Periodicity Timeliness	Scope/coverage of the data Geographical coverage Statistical units Statistical population
Access Dissemination of release calendar Release procedures	Accounting conventions Reference period Base period Recording of transactions
Integrity (practices and procedures) Rules on compilation and confidentiality Access to data before release Commentaries on the occasion of data release Revision and changes in methodologies	Nature of the basic data Data sources used Type of survey Techniques of data collection
Quality References to detailed methodology and sources Related data bases and information Quality framework and quality reports	Compilation practices Compilation of European aggregates Adjustments Data validation Revision policy
Dissemination formats (news releases, publications, on-line, databases)	Other aspects Special warnings



Data and Metadata Reporting





SDMX Standards

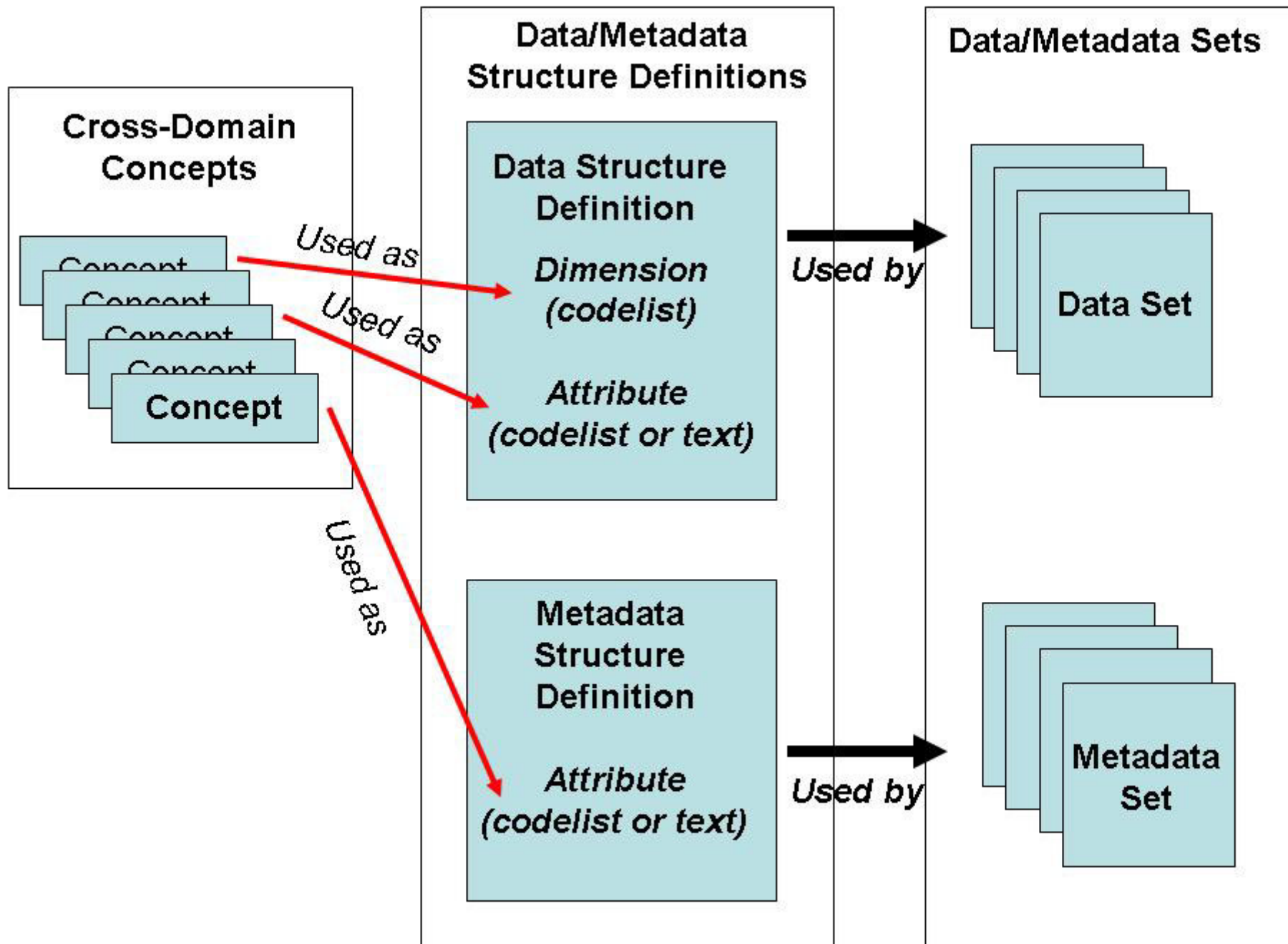
Two distinct but complementary sets of standards:

- ❑ **Technical standards (Version 2.0)**
 - **ISO 17369**
 - **Provide a data model for exchanging a dataset using a “Data Structure Definition”**
- ❑ **Content guidelines**
 - **A “Metadata Structure Definition” is used for exchanging metadata, independently of a specific set of data**



SDMX Content Guidelines

- Standardisation of concepts
 - => Cross-domain Metadata Concepts
- Standardisation of terminology
 - => Metadata Common Vocabulary
- Standardisation of Data/Metadata Structures
 - => Responsibility of domain groups
categorisation using a Statistical Domain List





Data Structure Definition (DSD)

- ❑ Identifies concepts that are dimensions (to identify series)
- ❑ Identifies concepts that are attributes (to describe series)
- ❑ Provides code lists and representations for the concepts
- ❑ Gives an attachment level for the concepts, based on the packaging structure (Data Set, Series, Observation)



Data Structure Definition (DSD)

Dimensions - Key			
Type	Concept	Representation/Code list	
Dimension (role is Frequency)	FREQ	Code List: CL_FREQ	
Dimension (role is Time)	TIME	Date/Time	
Dimension	DEMOGRAPHIC_TYPE	Code List: CL_DEMOG_TYPE	
Dimension	REGION	Code List: CL_REGION	
Dimension	AGE_RANGE	Code List: CL_AGE_RANGE	
Dimension	MEASURE_TYPE	Code List: CL_MEASURE_TYPE	
Measure			
OBS_VALUE (role is Primary Measure)			
Attributes			
Concept	Assignment Status	Assignment Level	Representation/Code List
OBS_STATUS	Mandatory	Observation	Code List: CL_OBS_STATUS
TITLE	Mandatory	Data Set	Text
SOURCE	Conditional	Data Set	Text
PUBLICATION_DATE	Conditional	Data Set	Text



Metadata Structure Definition (MDS D)

The Metadata Structure Definition identifies the structures to which metadata can be attached and defines the allowable content of metadata.

It is common for such type of metadata to be shareable amongst many elements to which it relates, so that it is often stored in a separate metadata repository where it is referenced from the element to which it relates. In SDMX, these metadata are called **Reference Metadata**.

Reference Metadata are content metadata that gives more information about the object so as to make its interpretation more meaningful.

The Metadata Structure Definition must:

1. Identify precisely what metadata is relevant (in terms of concepts), rules for its usage (e.g. mandatory/conditional) and value domain (e.g. code lists).
2. Define precisely the context of the metadata in terms of the object type to which it is relevant, and the precise identity of the object.



Reference Metadata may be attached at different levels

- **Dataset** (a whole domain or collection of data/metadata)
- **Dimension** (ex. country, indicator, version, frequency, time)
- **Dimension member** (discrete point for a specific dimension, ex. GDP)
- **Coordinates** (a combination of one or more dimension members each from a different dimension, ex. “flash estimate 1st quarter 2006 GDP at current prices for Germany”)
- **Coordinate levels** – Refers to where metadata is attached in relation to the data.
 1. Dimension level: highest coordinate level with one dimension member (ex. Germania)
 2. Intermediate level: combination of dimension members excluding other coordinate levels (GDP, Germany, estimate)
 3. Sibling level (from GESMES): all dimensions fixed except for "Frequency" and Time" (GDP, current prices, Germany, estimate)
 4. Series level: all dimensions fixed except "Time" (GDP, quarterly, current prices, Germany, estimate)
 5. Observation level: all dimensions fixed



The SDMX metamodel allows reference metadata to be:

- ❑ Exchanged without the need to embed it within the object that it is describing
- ❑ Stored separately from the object that it describes, yet be linked to it (for example, a metadata repository can support the dissemination of metadata resulting from requests generated by systems or services that have access to the object to which the metadata is linked)
- ❑ Indexed to aid searching (example: a registry service can process a metadata report and extract structural information that allows it to catalogue the metadata in a way that will enable users to query for it)
- ❑ Reported according to a defined structure and scheme

High-level reference metadata concepts

1. Contact
2. Metadata Update
3. Institutional framework
4. Confidentiality
5. Data presentation
6. Frequency, periodicity
7. Release calendar
8. Simultaneous release
9. Dissemination formats
10. Supplementary data
11. Accessibility
12. Timeliness and punctuality
13. Transparency
14. Quality management
15. Accuracy, reliability
16. Comparability, coherence
17. Relevance
18. Professionalism and ethics
19. Statistical concepts
20. Classification
21. Scope/coverage
22. Accounting conventions
23. Source data
24. Statistical processing
25. Validation
26. Revision policy

EUROSTAT –SDMX CROSS-DOMAIN CONCEPTS MAPPING

Eurostat mapping
to SDMX cross-
domain concepts

EUROSTAT DISSEMINATION METADATA CONCEPTS		MAPPING TO CURRENT DRAFT OF SDMX CROSS- DOMAIN CONCEPTS
Top level	Child level	
Metadata Update	Last certified without update	Date of update
	Last update of content	Date of update
Contact	Organisation	Contact
	Address	Contact
	Contact name or service	Contact
	e-mail address	Contact
Data coverage	Short description of data domain	Data presentation
	Data breakdown and main variables	Data presentation
	Units of measure	Data presentation
Periodicity	Periodicity of compilation	Frequency and periodicity
	Database frequency	Frequency and periodicity
Timeliness and punctuality	Timeliness	Timeliness and punctuality
	Punctuality	Timeliness and punctuality
Transparency of practices	Legal acts, reporting requirements	Institutional framework
	Rules on confidentiality	Institutional framework
	Internal access	Transparency
	Commentary on the occasion of release	Transparency
Accessibility	Notification of changes in methodology	Transparency
	Release calendar	Release calendar
	Simultaneous release	Simultaneous release
	Dissemination formats	Dissemination formats
Quality cross-checks	Documentation on methodology	Accessibility of documentation
	Related data and quality cross-checks	[No direct concordance]
Accuracy and reliability	References to quality reports	[No direct concordance]
	Overall accuracy assessment	Accuracy
	Quality checks before release	Accuracy
Comparability and coherence	Comparability over time	Comparability and coherence
	Comparability over space	Comparability and coherence
	Comparability with related sources	Comparability and coherence
	Comparability between datasets	Comparability and coherence
	Breaks in time series	Comparability and coherence
Relevance	Rate of available statistics (user needs)	Relevance
	Intended audience and purpose	Relevance
	Supplementary data	Supplementary data
Statistical concepts and classifications	Statistical concept	Statistical concept
	Definition of indicators	Statistical concept
	Classification system	Classification systems
	Conformity with official standards	Classification systems
	Classification coverage	Classification systems
Scope of the data	Reference area / geopolitical entity	Scope/coverage
	Time coverage	Scope/coverage
	Statistical unit	Scope/coverage
	Statistical population	Scope/coverage
Accounting conventions	Reference period	Accounting conventions
	Base period	Accounting conventions
Nature of basic data	Basis for recording	Accounting conventions
	Data source used	Source data
	Type of survey	Source data
Compilation practices	Methods of data collection	Source data
	Compilation	Statistical processing
	Adjustments and weights	Statistical processing
	Data validation	Statistical processing
Other	Revision policy and practice	Revision policy and practice
	Warnings on re-use and limitations	[No direct concordance]



XML Schema

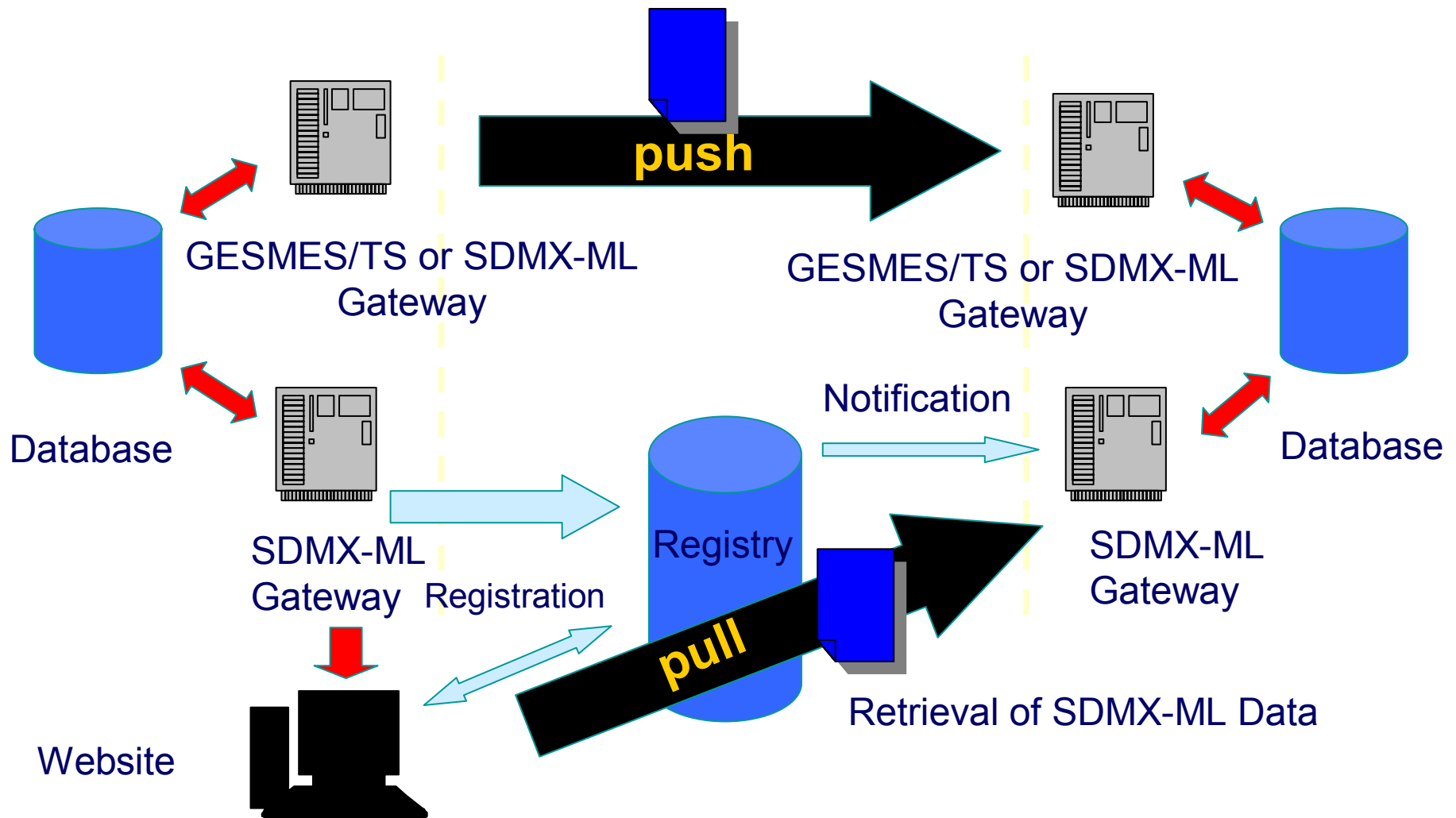
The purpose of an XML Schema is to define the building blocks of an XML document:

- elements that can appear in a document
- attributes that can appear in a document
- which elements are child elements
- the order of child elements
- the number of child elements
- whether an element is empty or can include text
- data types for elements and attributes
- default and fixed values for elements and attributes

SDMX-XML + XSL Transformation

Data/Metadata Reporting & Registry-Based Dissemination

Transmission of SDMX-EDI or XML messages





A few interesting quotations

Nothing is more practical than a good theory

*I never worry about the future. It comes soon enough
(Albert Einstein)*

*Reasonable people adapt themselves to the world
Unreasonable people attempt to adapt the world to themselves
All progress, therefore, depends on unreasonable people
(George Bernard Shaw)*