Metadata for European Microdata

EUROSTAT

Luca Gramaglia

Luca.Gramaglia@ec.europa.eu

Microdata project

- Project “Microdata for European microdata”
- Currently metadata which goes with the microdata are based on a document which
  - is annual (static) and does not describe systematically the changes over time;
  - describes the data as received by Eurostat, not as released to the users
  - is domain specific
  - is usually in the form of pdf documents or word/excel files
- Purpose: Eurostat aims at creating an effective and efficient metadata system that solves the issues

Microdata project result

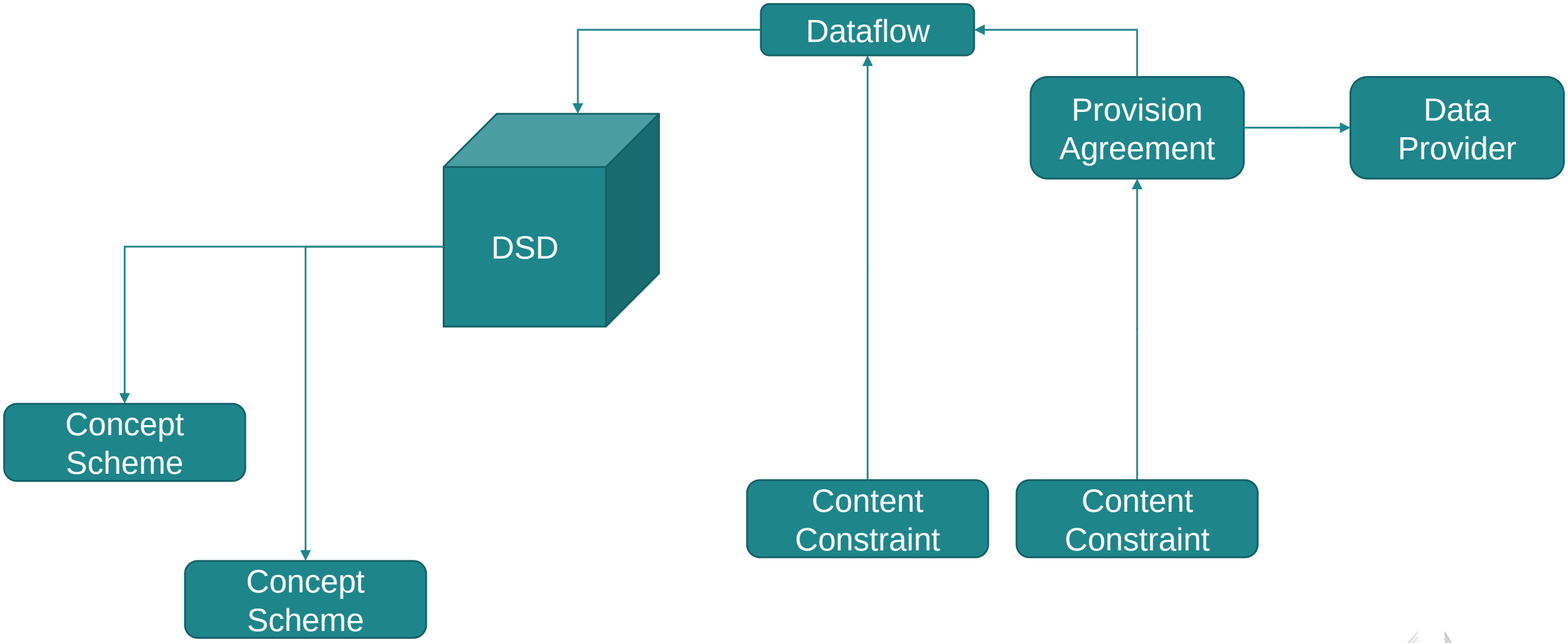
Ultimate objectives of the project:

- **Metabase** (repository + services/APIs), which makes metadata accessible to the public
- The system should allow to explore generic concepts and domain specific variables across the following dimensions:
  - domain or survey
  - time
  - level of anonymisation
  - version
  - country

DSDs for Metabase

- EU Statistics on Income and Living Conditions (EU-SILC)
- Labour Force Survey (LFS)
- Continuing Vocational Training Survey (CVTS)
- European Health Interview Survey (EHIS)
- Harmonised European Time Use Survey (HETUS)
- Integrated Farm Statistics (IFS)

Overview of Proposed Data Model



Modelling microdata SDMX version 2.1

- When modelling aggregate data we usually employ several Dimensions and Attributes to describe on Primary Measure (observation value)
- Conceptually, in microdata we have more than one measure. Although we could also use multiple measures already in v 2.1:
  - Less or no support from IT tools for multiple measures (e.g.in the Euro SDMX Converter or SDMX-RI)
  - No backwards-compatibility between versions 2.0 and 2.1 when multiple measures are used
  - Measures cannot be declared as optional or mandatory

Questionnaire into SDMX

	Question	Answer	Value
SURV_YEAR	YEAR OF THE SURVEY	2022	2022
PERS_ID	PERSONAL ID FOR THE SURVEY	1234	1234
WEIGHT	PERSONAL WEIGHT	5	5
X1	CONTACT WITH RESPONDENT	1 Person contacted	1
X2	INTERVIEW RESULT	1 Interview completed and accepted for database	1
…	…		
X10	MONTH OF INTERVIEW	3	3
	TOTAL DURATION OF INTERVIEW	2	2
B14	Have you ever worked during your lifetime? By work, we mean paid employment or unpaid work only if performed at an enterprise owned by a family member.	1 Yes	1
B15(SV9) REF/DNK=9	(optional). Do you work part-time or full-time? If you have more than one job, please think about job where you usually work the most hours. IF OPTIONAL NOT INCLUDED IN THE SURVEY BY COUNTRY = 97 NOT COLLECTED		
B16(SV10) REF/DNK=9	Do you have a fixed-term contract or a permanent job? If you have more than one job, please think about job where you usually work the most hours. (READ OUT)		
B21	What is your legal marital status?	1 Never been married/in a civil partnership	1
B22	(optional). How old were you when you got married for the first time?		

European Commission

Microdata

D	D	M	A	A	A	A	A	A	A	A	A
SURV_YEAR	PERS_ID	OBS_VALUE	AGE	INCOME	SEX	INCOME_STATUS	B13	B14	B15	B16	B21
2022	1234		27	100	M	A	2	1			1
2022	1235		71	100	F	A	2	2	1	1	3
2022	1236		45	200	F	A	1	2	8		1
2022	1238		18	100	F	E	2	1			1
2022	1258		76	100	F	E	1	1			2
2022	1298		55	100	M	A	2	2	8		2

 Dimension

 Measure

 Attribute

Challenges – modelling

- As everything is modelled as an attribute in SDMX 2.1, the conceptual difference between attributes and measures is lost in our model.
- Impossibility to indicate that a Flag/flags refer to a specific measure.
  - If we would need flag to measure (in our case attribute) We cannot link / associate an attribute to a measure specifically: for example, information on how the “income” variable is collected cannot be linked to the “income” measure
- Observation value is not in use, but we have to keep it

Challenges – modelling

- With SDMX 3.0, several of the modelling challenges could be addressed
- We would be able to use many Measures to store the value instead of attributes.
- Attributes could be attached to specific measures (new ‘MeasureRelationship’ property for attributes)
- The representation of measures in microdata often includes specific values with special meaning. For example, the expected values for ‘income’ can be 0 to 1000000000, with -1 indicating ‘not stated’. SDMX 3.0 allows to include information about these values with specific meaning (so-called Sentinel values)

v2.1 modelling vs v3.0 modelling

D	D	M	A	A	A	A	A	A	A	A	A
SURV_YEAR	PERS_ID	OBS_VALUE	AGE	INCOME	SEX	INCOME_STATUS	B12	B14	B15	B16	B21
2022	1234		27	100	M	A	2	1			1

D	D	M	M	M	A	M	M	A	M	M	3
SURV_YEAR	PERS_ID	AGE	INCOME	SEX	INCOME_STATUS	B12	B14	B14_A	B16	B21	1
											1
2022	1234	27	100	M	A	2	1			1	2
2022	1235	71	100	F	A	2	2	1	1	3	2
2022	1236	45	200	F	A	1	2	8		1	
2022	1238	18	100	F	E	2	1			1	
2022	1258	76	100	F	E	1	1			2	
2022	1298	55	100	M	A	2	2	8		2	

Challenges – browsing

- The SDMX API and most SDMX tools are “structure-centric” – the entry point for browsing is the data structure
- Researchers working on microdata tend to me variable-centric – they focus on once concept and want to see in which micro-datasets it is present and how it has evolved over time.
- This issue can be solved at software level, but most existing SDMX software does not take this approach to browsing structural metadata.

Challenges – attaching reference metadata

- Additional useful information could be attached to variables or metadataflows in our model. For example, detailed variable definitions, methodological notes on the sources or on the statistical disclosure control mechanisms applied, etc.
- This information could be attached in several ways to the SDMX model applied for microdata:
  - As annotations to the affected objects
  - Or as reference metadata linked to the objects
- Exploring these different possibilities in the context of the implementation of SDMX 3.0 is an area of active research

Thank you!

European Commission