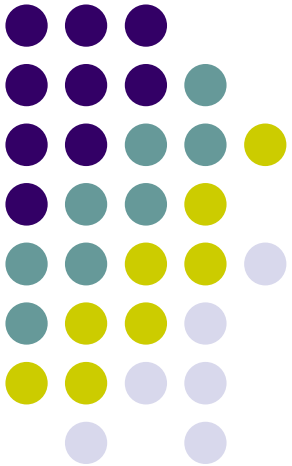
SDMX Information Model: An Introduction

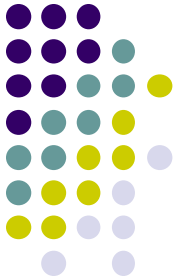
Workshop on Data and Metadata Sharing

Bangkok, 10-14 December 2018

Abdulla Gozalov, UNSD

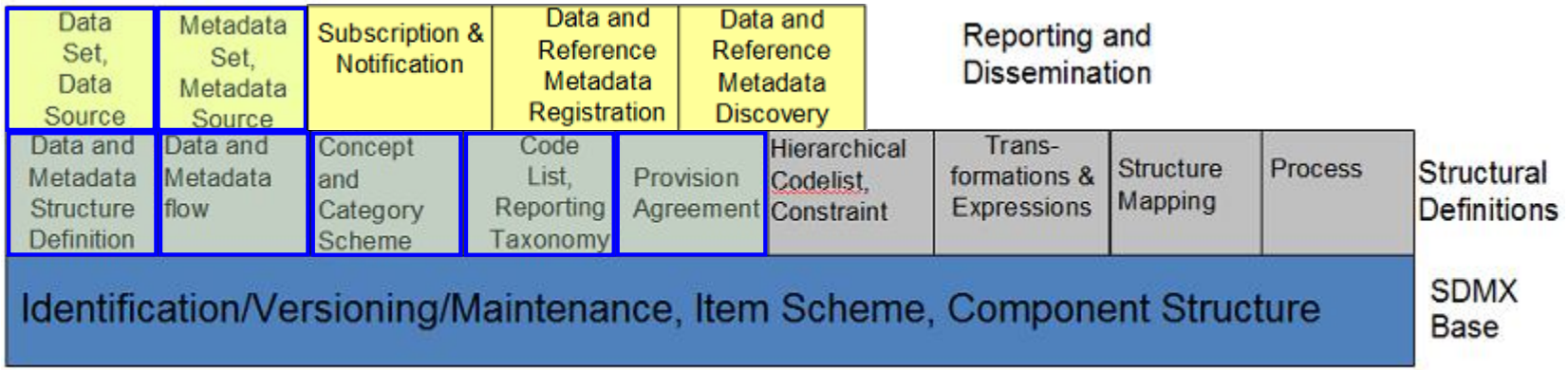


SDMX Information Model



- An abstract model, from which actual implementations are derived.
- Implemented in XML, GESMES, JSON, CSV
- Can be thought of as a number of packages arranged in 3 layers…

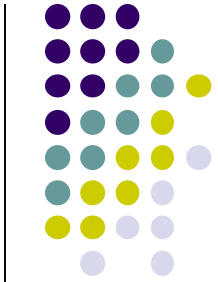
SDMX Information Model: Packages and Layers



- We will focus on:
  - Structural Definitions
  - Data and Metadata reporting

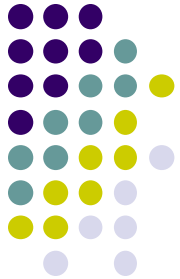
Source: SDMX SECTION 02 INFORMATION MODEL: UML CONCEPTUAL DESIGN

Structural vs Reference Metadata



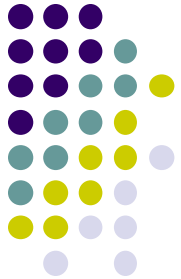
- Structural Metadata: Identifiers and Descriptors, e.g.
  - Data Structure Definition
  - Concept Scheme
  - Code
- Reference Metadata: Describes contents and quality of data, e.g.
  - Indicator definition
  - Comments and limitations

Data Structure Definition (DSD)



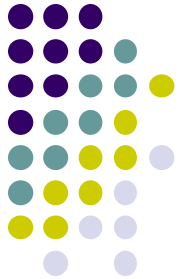
- Represents a data model used in exchange
- Defines dataset structure
- A DSD contains:
  - Concepts that pertain to the data
  - Code lists, which represent the concepts
  - Dimensional structure, which describes roles of the concepts
  - Groups, which define higher levels of aggregation.
- Also known as *Key Family*, but this term was discontinued in SDMX 2.1

Concept Scheme



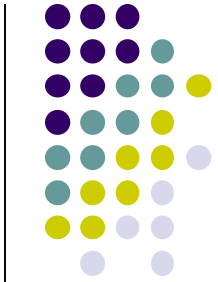
- “The descriptive information for an arrangement or division of concepts into groups based on characteristics, which the objects have in common.”
- Concept scheme places concepts into a maintainable unit.

Code Lists and Codes



- Code lists provide representation for concepts, in terms of Codes.
- Codes are language-independent and may include descriptions in multiple languages.
- Code lists must be harmonized among all data providers that will be involved in exchange.

Dimensional Structure

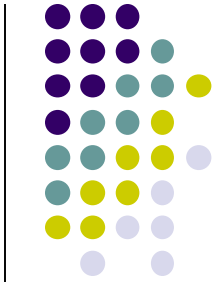


- Described as part of Data Structure Definition
- Lists concepts for:
  - Dimensions
  - Attributes
  - Measure(s)
- Links concepts to code lists
- Defines groups.
- Defines attribute attachment levels.

Groups

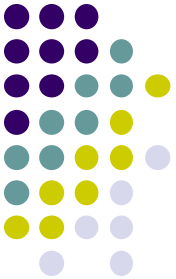
- In SDMX, groups define *partial keys* which can be used to attach information to.
- Attributes can be attached at observation, series, group, or dataset level. The parsimony principle calls for attributes to be attached to the highest applicable level.
  - But for practical purposes attributes are typically attached to the observation or time series
- Groups are not used in the SDG DSD, and are generally rarely used or supported

Time Series



- A set of observations of a particular variable, taken at different points in time.
- Observations that belong to the same time series, differ in their time dimension.
  - All other dimension values are identical.
  - Observation-level attributes may differ across observations of the same time series.

Time Series: Demonstration



1.1 Proportion of population below $1 (PPP) per day

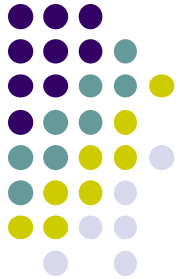
Series	1990	1992	1994	1996	1998	1999	2000	2002	2006	2007	2008	2009	2011
Rwanda													
MDG Population below $1 (PPP) per day, percentage Last updated: 02 Jul 2012							74.6[1,3]		72.1[1,3]				63.2[1,3]
State of Palestine													
MDG Population below $1 (PPP) per day, percentage Last updated: 02 Jul 2012										0.4[1,2,3]		0.0[1,2,3]	
Thailand													
MDG Population below $1 (PPP) per day, percentage Last updated: 02 Jul 2012	11.6[1,3]	8.6[1,3]	4.1[1,3]	2.5[1,3]	2.1[1,3]	3.2[1,3]	3.0[1,3]	1.6[1,3]	1.0[1,3]		0.4[1,3]	0.4[1,3]	

1.2 Poverty gap ratio

Series	1990	1992	1994	1996	1998	1999	2000	2002	2006	2007	2008	2009	2011
Rwanda													
MDG Poverty gap ratio at $1 a day (PPP), percentage Last updated: 02 Jul 2012							36.9[1,3]		34.8[1,3]				26.6[1,3]
State of Palestine													
MDG Poverty gap ratio at $1 a day (PPP), percentage Last updated: 02 Jul 2012										0.1[1,2,3]		0.0[1,2,3]	
Thailand													
MDG Poverty gap ratio at $1 a day (PPP), percentage Last updated: 02 Jul 2012	2.4[1,3]	1.6[1,3]	0.7[1,3]	0.4[1,3]	0.3[1,3]	0.5[1,3]	0.5[1,3]	0.3[1,3]	0.2[1,3]		0.0[1,3]	0.1[1,3]	

Footnotes	
1	Based on nominal per capita consumption averages and distributions estimated from household survey data.
2	Based on Purchasing Power Parity (PPP) dollars imputed using regression.
3	Source: http://iresearch.worldbank.org/PovcalNet/index.htm

Cross-Sectional Data



- A non-time dimension is chosen along which a set of observations is constructed.
  - E.g. for a survey or census the time is usually fixed and another dimension may be chosen to be reported at the observation level
- Used less frequently than time series representation

Time Series View vs Cross-Sectional View

2.1 Net enrolment ratio in primary education

	2009	2010	2011
Morocco			
Total net enrolment ratio in primary education, both sexes		94.1	96.2
Total net enrolment ratio in primary education, boys		95	96.8
Total net enrolment ratio in primary education, girls		93.3	95.6
State of Palestine			
Total net enrolment ratio in primary education, both sexes	88.2	89.2	
Total net enrolment ratio in primary education, boys	88.2	89.8	
Total net enrolment ratio in primary education, girls	88.2	88.5	
Uganda			
Total net enrolment ratio in primary education, both sexes	94.2	91	
Total net enrolment ratio in primary education, boys	93.1	89.7	
Total net enrolment ratio in primary education, girls	95.3	92.3	

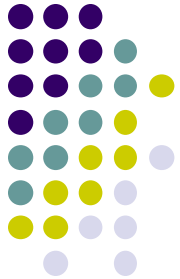
- The Sex dimension was chosen as the cross-sectional measure.

- Note that Time is still applicable.

2.1 Net enrolment ratio in primary education
2010

	Total	Boys	Girls
Morocco	94.1	95	93.3
State of Palestine	89.2	89.8	88.5
Uganda	91	89.7	92.3

Keys in SDMX



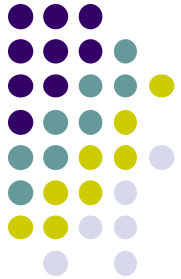
- **Series key** uniquely identifies a time series
  - Consists of all dimensions except **time**
- **Group key** uniquely identifies a group of time series
  - Consists of a subset of the series key

Dataset

- “…can be understood as a collection of similar data, sharing a structure, which covers a fixed period of time.”*
- A collection of time series or cross-sectional series
- Dataset serves as a container for series data in SDMX data messages.

*Source: Metadata Common Vocabulary

Exercise 3:
Encoding a time series

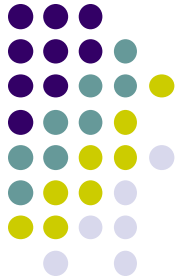


- Working with your table, identify each time series.
- For each time series, provide a valid value for each concept in its series key.

Metadata in SDMX

- Can be stored or exchanged separately from the object it describes, but be linked to it
- Can be indexed and searched
- Reported according to a defined structure

Metadata Structure Definition (MSD)



- MSD Defines:
  - The object type to which metadata can be associated
    - E.g. DSD, Dimension, Partial Key.
  - The components comprising the object identifier of the target object
    - E.g. the draft SDG MSD allows metadata to be attached to each indicator for each country
  - Concepts used to express metadata (“metadata attributes”).
    - E.g. Indicator Definition, Quality Management

Metadata Structure Definition and Metadata Set: an example

METADATA STRUCTURE DEFINITION

Target Identifier	Metadata Attributes
Component: **SERIES** (phenomenon to be measured)	Concept: **STAT_CONC_DEF** (Indicator Definition)
Component: **REF_AREA** (Reference Area)	Concept: **METHOD_COMP** (Method of Computation)

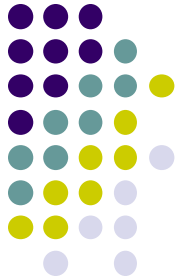
METADATA SET

SERIES=**SH_STA_BRTC** (Births attended by skilled health personnel)
REF_AREA=**KH** (Cambodia)

STAT_CONC_DEF="It refers to the proportion of deliveries that were attended by skilled health personnel including physicians, medical assistants, midwives and nurses but excluding traditional birth attendants."

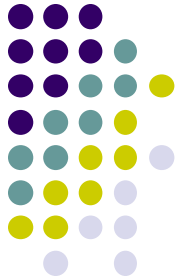
METHOD_COMP="The number of women aged 15-49 with a live birth attended by skilled health personnel (doctors, nurses or midwives) during delivery is expressed as a percentage of women aged 15-49 with a live birth in the same period. "

Dataflow and Metadataflow



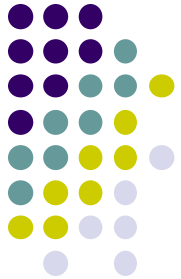
- Dataflow defines a “view” on a Data Structure Definition
  - Can be constrained to a subset of codes in any dimension
  - Can be categorized, i.e. can have *categories* attached
  - In its simplest form defines any data valid according to a DSD
- Similarly, Metadataflow defines a view on a Metadata Structure Definition.

Category and Category Scheme



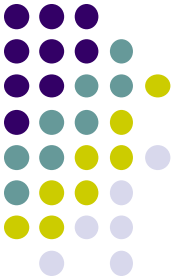
- Category is a way of classifying data for reporting or dissemination
  - Subject matter-domains are commonly implemented as Categories, such as “Demographic Statistics”, “Economic Statistics”
- Category Scheme groups Categories into a maintainable unit.

Data Provider and Provision Agreement



- Data Provider is an organization that produces and disseminates data and/or reference metadata.
- Provision Agreement links a Data Provider and a Data/Metadata Flow.
  - I.e. a Data Provider agrees to provide data as specified by a Dataflow.
- Like Dataflows, Provision Agreements can be categorized and constrained.

Content Constraints

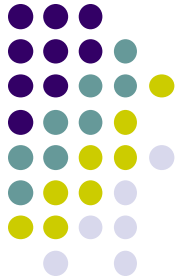


- Constraints can be used to define which combinations of codes are allowed
  - E.g. "*When* ***SERIES***=*'*Proportion of Women in Commune Councils*', **SEX** must be '*Female*'"
- Constraints can define more granular validation rules than a simple validation of codes
- Are often attached to the Dataflow but can also be attached to DSD, Provision Agreement, etc

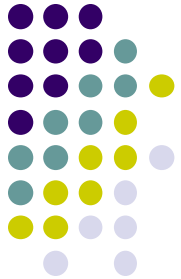
SDMX Messages

- Any SDMX-related information is exchanged in the form of documents called *messages*.
- An SDMX message can be sent in a number of standard formats including XML, JSON, CSV
- There are several types of SDMX messages, each serving a particular purpose, e.g.
  - **Structure** message is used to transmit structural information such as DSD, MSD, Concept Scheme, etc.
  - **GenericData**, **StructureSpecificData,** and other messages are used to send data.
- SDMX messages in the XML format are referred to as SDMX-ML messages.

Exercise 4: Developing a DSD



- Working with your table, develop a Data Structure Definition.



THANK YOU!