United Nations

Department of Economic and Social Affairs
Statistics

Introduction to SDMX data modeling

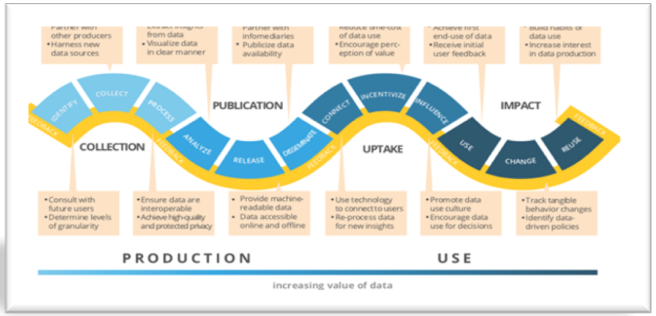
Deqing, China. 17-19 October 2019

Photo by Markus Spiske, unsplash

Interoperability

- Ability to seamlessly share, join, cross-analyse, exchange and re-use data produced from different sources, and at different times, to provide richer information for improved decision making
- It is a crucial characteristic of good quality data and of effective data management systems

Interoperability should be understood along the whole “data value chain”, from **collection** to **use**



Partner with other producers
• Harness new data sources
Extract insights from data
• Visualize data in clear manner
Partner with infomediaries
• Publicize data availability
Reduce time-cost of data use
• Encourage perception of value
• Achieve first end-use of data
• Receive initial user feedback
• Build habits of data use
• Increase interest in data production
FEEDBACK
IDENTIFY
COLLECT
PROCESS
ANALYZE
RELEASE
DISSEMINATE
CONNECT
INCENTIVIZE
INFLUENCE
USE
CHANGE
REUSE
COLLECTION
PUBLICATION
UPTAKE
IMPACT
• Consult with future users
• Determine levels of granularity
• Ensure data are interoperable
• Achieve high-quality and protected privacy
• Provide machine-readable data
• Data accessible online and offline
• Use technology to connect to users
• Re-process data for new insights
• Promote data use culture
• Encourage data use for decisions
• Track tangible behavior changes
• Identify data-driven policies
PRODUCTION
USE
increasing value of data

Interoperability and data modelling

- Interoperability is highly dependent on data and metadata modelling decisions and practices
- The same information content is often represented in variety of ways across different systems and organizations.
- There is usually no single “right” way of representing information
  - Some data structures are better suited for managing transactional processes (e.g., capturing data from a survey or maintaining a civil registration database)
  - Others are better suited for analyzing and communicating data to users (e.g., for the creation of data visualizations in a monitoring dashboard).

What is data modeling?

- A process focused on:

  1. Clearly and unambiguously **identifying things** that a dataset aims to capture
  2. Selecting the key properties that should be captured to **describe those things** in a meaningful way
  3. Deciding **how things relate** to each other
  4. Deciding how this information should be **formally codified**

→ *This is the essence of the Entity-Relationship model, which underlies most modern database management systems and applications*

Examples

- The content of a dataset may refer to **entities** such as “city”, “person”, or “activity”,
- These entities may be usefully described with **attributes** like “name”, “age”, or “industry”.
- In a specific application, it could be useful to capture **relationships** among entities and attributes, e.g., the fact that
  - one or more persons may live in a city,
  - a person may be employed in one or more types of activity …

Canonical data and metadata models

- Models that follow specific standardized patterns, making them highly **reusable** and conducive to data sharing.
- Provide a **common template** to which different datasets can be mapped
- Help develop a common understanding of how the various components of a dataset relate to each other and to the components of other datasets
- Reduce the number of transformations that user applications need to perform on their own to integrate the data from those sources

Standardization is not for free

- The underlying principle is **to hide from user the internal complexity of the operational data models** (e.g., which are optimized to avoid data redundancy and ensure data consistency validations), so they can concentrate on using data rather than spending time trying to understand the intricacies of internal data structures
- Data **providers need to take responsibility for mapping the data** from its original, operational structures, into commonly agreed presentations for dissemination and distribution purposes
- This may entail the need to undertake so-called "Extract-Transform-Load", or ETL, procedures, **hidden from the view of users**

The multi-dimensional ‘data cube’ model

- Presents all relevant data about a population of interest in a **simple**, **self-contained** tabular view
- Each data point is characterized by
  - **Measures**: Observed values on one or more variables interest
  - **Dimensions**: A set of uniquely identifying characteristics
  - **Attributes**: A set of additional characteristics that further describe it

Domains of dimensions, attributes and measures

- Each dimension, measure and attribute encapsulates **a concept**
- Concepts can be:
  - drawn from a code list (for e.g., “country ISO code”)
  - required to adhere to a specific data format (e.g., “YYYY” for years)
  - required to be contained within a specific range of values (e.g., “numerical values between 0 and 1”).
  - drawn from a type of values (e.g., “text”)

United Nations | Department of Economic and Social Affairs
Statistics

Developing a data model

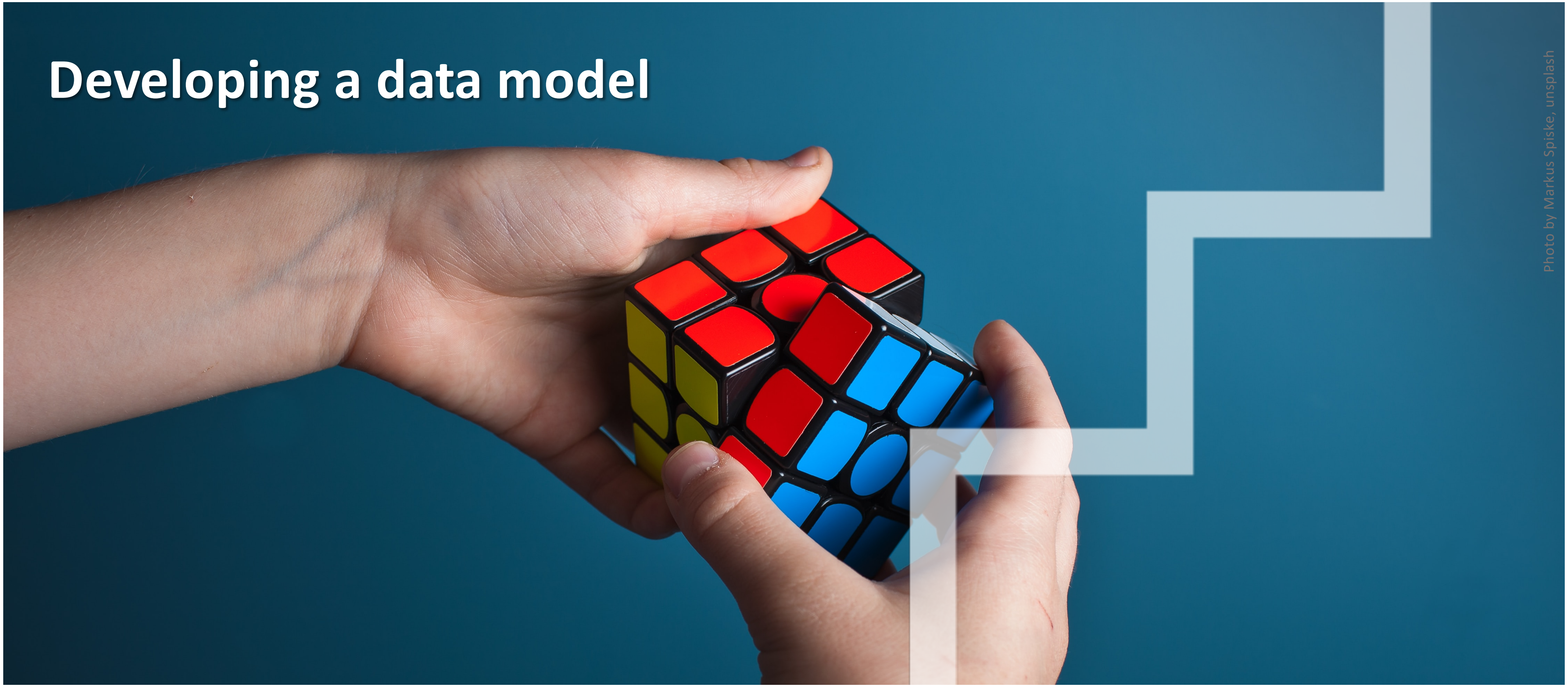
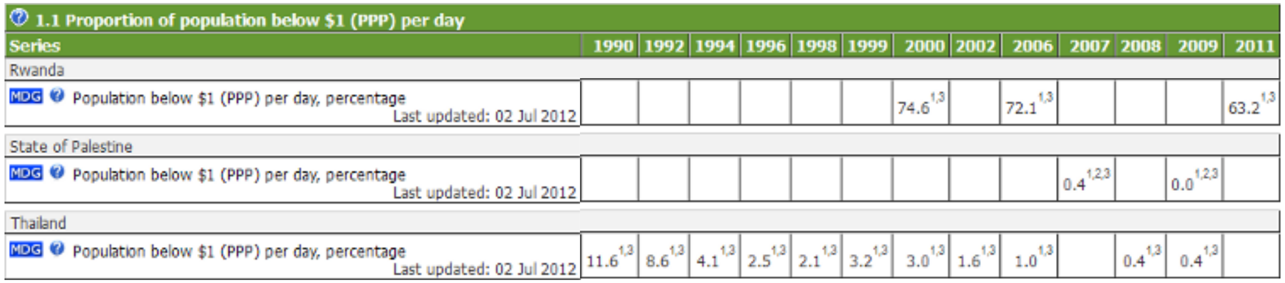


Photo by Markus Spiske, unsplash

Numbers vs Data



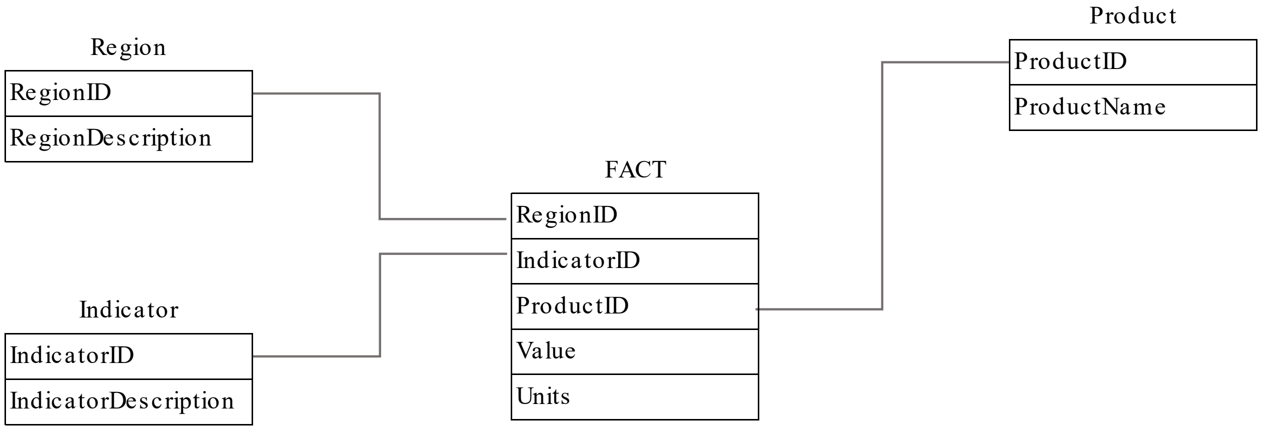
1.1 Proportion of population below $1 (PPP) per day

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]	

- Numbers by themselves are meaningless.
- To be usable, they must be properly described
- By adding descriptions to let users know what the figures actually represent, they become data

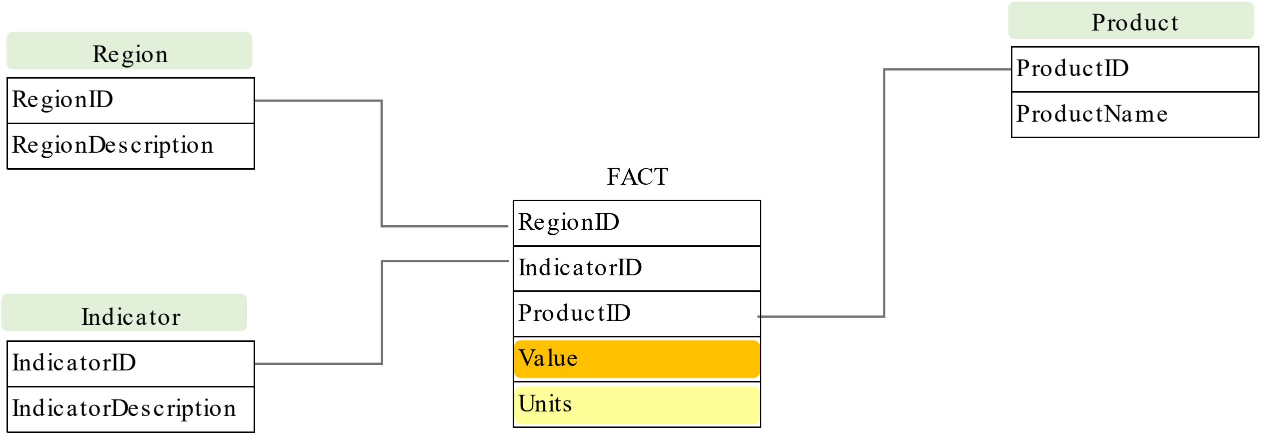
Developing a Data Model for Dissemination and Exchange

- This task is similar to developing a relational database
- In SDMX, a data model is represented by a **Data Structure Definition (DSD)**
- The “shape” of SDMX DSD is roughly similar to a “**star schema**”



Concept

- To design a DSD, we first need to find **concepts** that describe all relevant characteristics of our data



Region

RegionID

RegionDescription

Indicator

IndicatorID

IndicatorDescription

FACT

RegionID

IndicatorID

ProductID

Value

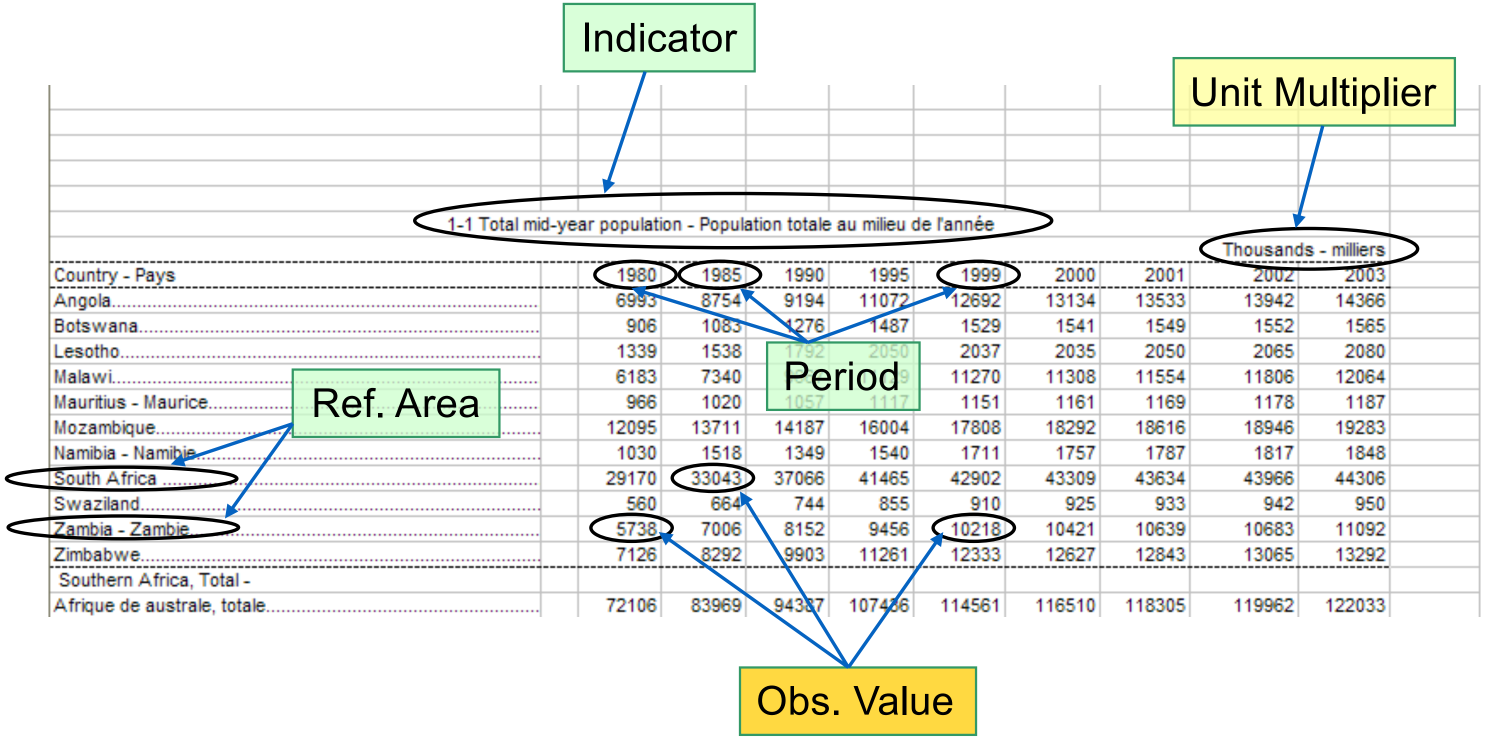
Units

Product

ProductID

ProductName

Identifying Concepts



1-1 Total mid-year population - Population totale au milieu de l'année

Thousands - milliers

Country - Pays	1980	1985	1990	1995	1999	2000	2001	2002	2003
Angola..	6993	8754	9194	11072	12692	13134	13533	13942	14366
Botswana....................................	906	1083	1276	1487	1529	1541	1549	1552	1565
Lesotho......................................	1339	1538	1792	2050	2037	2035	2050	2065	2080
Malawi..	6183	7340	[illegible]	[illegible]	11270	11308	11554	11806	12064
Mauritius - Maurice......................	966	1020	1057	1117	1151	1161	1169	1178	1187
Mozambique................................	12095	13711	14187	16004	17808	18292	18616	18946	19283
Namibia - Namibie........................	1030	1518	1349	1540	1711	1757	1787	1817	1848
South Africa	29170	33043	37066	41465	42902	43309	43634	43966	44306
Swaziland...................................	560	664	744	855	910	925	933	942	950
Zambia - Zambie..........................	5738	7006	8152	9456	10218	10421	10639	10683	11092
Zimbabwe....................................	7126	8292	9903	11261	12333	12627	12843	13065	13292
Southern Africa, Total -									
Afrique de australe, totale..............	72106	83969	94387	107436	114561	116510	118305	119962	122033

Dimension

- Which of the concepts are used to identify an observation?
  - Indicator
  - Reference area
  - Period
- When all 3 are known, we can unambiguously locate an observation in the table.
- In SDMX such concepts are called **dimensions**.
  - A dimension is similar in meaning to a database table’s primary key field.

Primary Measure

- Observation Value represents a concept that describes the actual values being transmitted.
- In SDMX, such a concept is called **Primary Measure**.
- Primary Measure is usually represented by concept **OBS_VALUE**.

Attribute

- In our example, **Unit Multiplier** represents additional information about observations.
- This concept is not used to identify a series or observation.
- Such concepts in SDMX are called **attributes**.
    - Similar to a database table’s non-primary key fields.

Dimension or Attribute?

- Concepts that **identify data**, should be made **dimensions**
- Concepts that provide **additional information**, should be made **attributes**

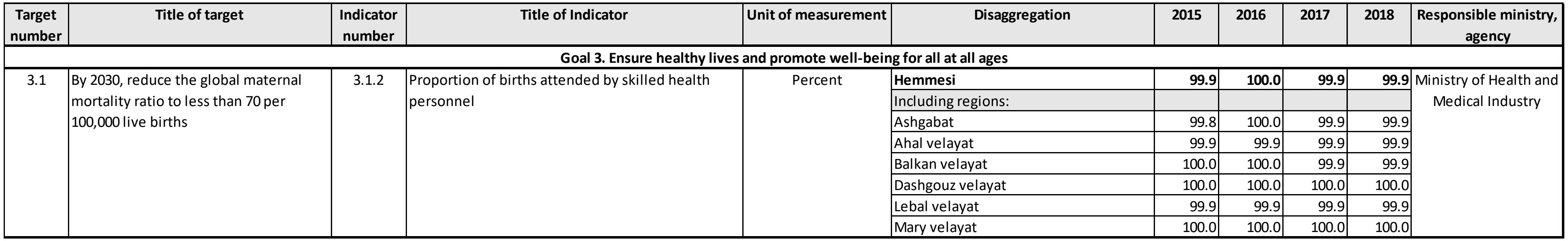
Special Dimensions

- **TIME** dimension provides the period of time to which the observation relates. If a DSD describes time series data, it must have one TIME dimension.
- **REFERENCE AREA** dimension describes the geographic location to which the observation refers (e.g., country, region, city, …)

→**Everything happens at a specific moment (or period) in time**

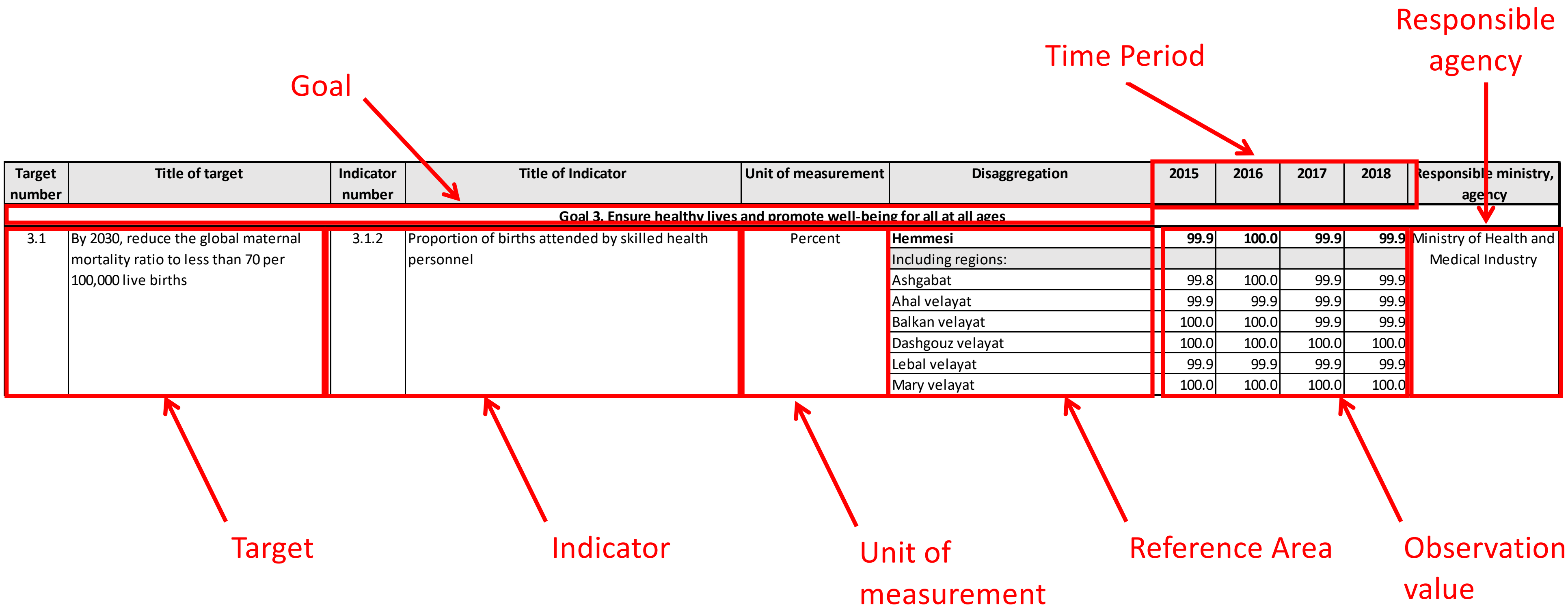
→**Everything happens somewhere**

Example



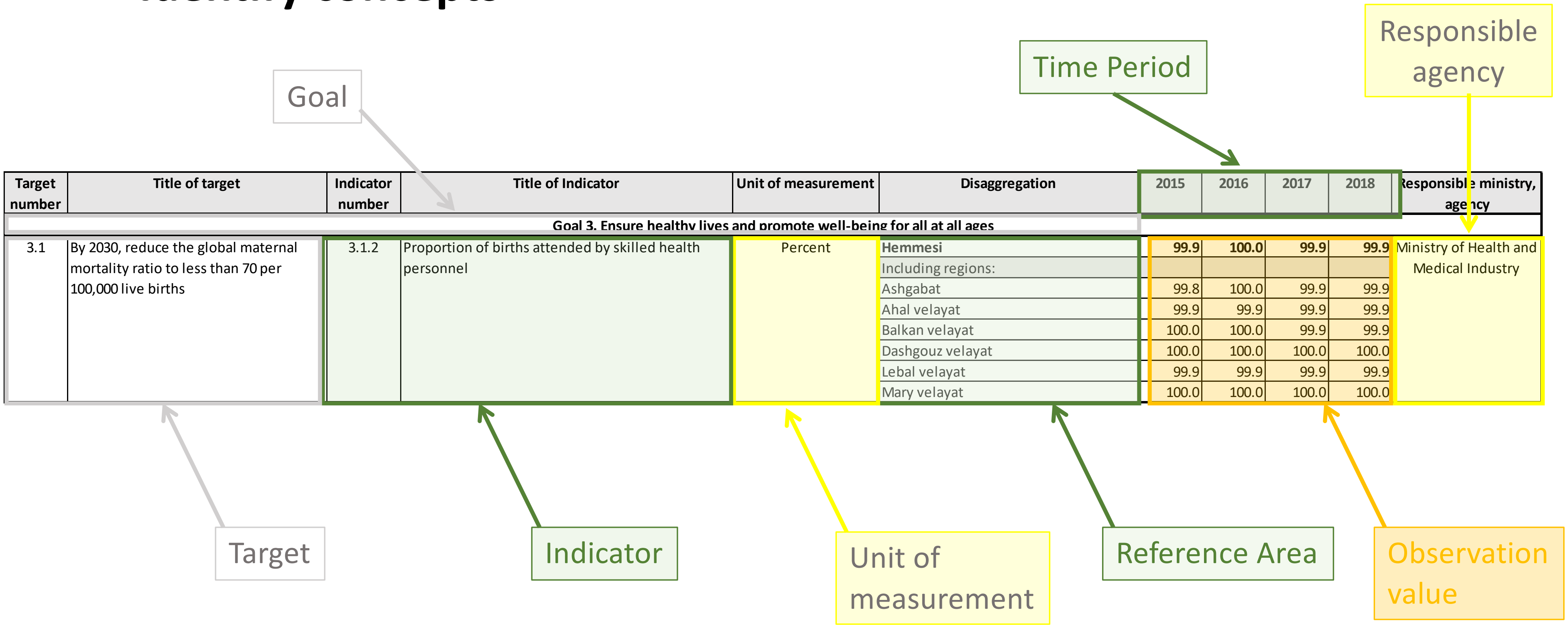
Target number	Title of target	Indicator number	Title of Indicator	Unit of measurement	Disaggregation	2015	2016	2017	2018	Responsible ministry, agency
Goal 3. Ensure healthy lives and promote well-being for all at all ages										
3.1	By 2030, reduce the global maternal mortality ratio to less than 70 per 100,000 live births	3.1.2	Proportion of births attended by skilled health personnel	Percent	**Hemmesi**	**99.9**	**100.0**	**99.9**	**99.9**	Ministry of Health and Medical Industry
					Including regions:					
					Ashgabat	99.8	100.0	99.9	99.9	
					Ahal velayat	99.9	99.9	99.9	99.9	
					Balkan velayat	100.0	100.0	99.9	99.9	
					Dashgouz velayat	100.0	100.0	100.0	100.0	
					Lebal velayat	99.9	99.9	99.9	99.9	
					Mary velayat	100.0	100.0	100.0	100.0	

Identify concepts


Goal
Time Period
Responsible agency
Target
Indicator
Unit of measurement
Reference Area
Observation value

Target number	Title of target	Indicator number	Title of Indicator	Unit of measurement	Disaggregation	2015	2016	2017	2018	Responsible ministry, agency
Goal 3. Ensure healthy lives and promote well-being for all at all ages										
3.1	By 2030, reduce the global maternal mortality ratio to less than 70 per 100,000 live births	3.1.2	Proportion of births attended by skilled health personnel	Percent	**Hemmesi**	**99.9**	**100.0**	**99.9**	**99.9**	Ministry of Health and Medical Industry
					Including regions:					
					Ashgabat	99.8	100.0	99.9	99.9	
					Ahal velayat	99.9	99.9	99.9	99.9	
					Balkan velayat	100.0	100.0	99.9	99.9	
					Dashgouz velayat	100.0	100.0	100.0	100.0	
					Lebal velayat	99.9	99.9	99.9	99.9	
					Mary velayat	100.0	100.0	100.0	100.0	

Identify concepts



Target number	Title of target	Indicator number	Title of Indicator	Unit of measurement	Disaggregation	2015	2016	2017	2018	Responsible ministry, agency
Goal 3. Ensure healthy lives and promote well-being for all at all ages										
3.1	By 2030, reduce the global maternal mortality ratio to less than 70 per 100,000 live births	3.1.2	Proportion of births attended by skilled health personnel	Percent	**Hemmesi**	**99.9**	**100.0**	**99.9**	**99.9**	Ministry of Health and Medical Industry
					Including regions:					
					Ashgabat	99.8	100.0	99.9	99.9	
					Ahal velayat	99.9	99.9	99.9	99.9	
					Balkan velayat	100.0	100.0	99.9	99.9	
					Dashgouz velayat	100.0	100.0	100.0	100.0	
					Lebal velayat	99.9	99.9	99.9	99.9	
					Mary velayat	100.0	100.0	100.0	100.0	

Goal

Time Period

Responsible agency

Target

Indicator

Unit of measurement

Reference Area

Observation value

Exercise 1: Identifying concepts

- Identify concepts in the table
- Mark each concept as:
  - Dimension
  - Primary Measure
  - Attribute
- Identify the Time Dimension (Reference Period)
- Identify the Reference Area Dimension

Representation

- When data are transferred, its descriptor concepts must have valid values.
- A concept can be
  - Coded
  - Un-coded with format
  - Un-coded free text

Code

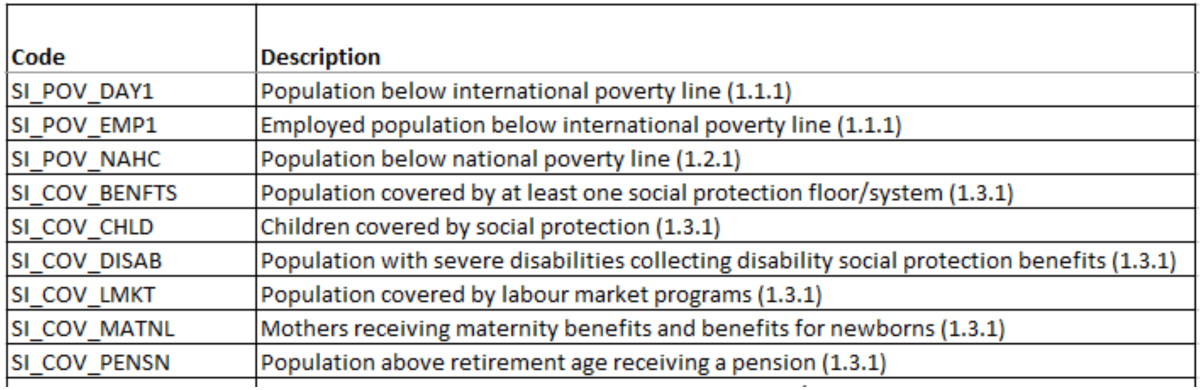
- A **language-independent** set of characters (letters, numbers or symbols) that **represents a concept** whose **meaning** is described in any number of languages

→ The use of codes allows natural language descriptions to be updated without disrupting data exchange

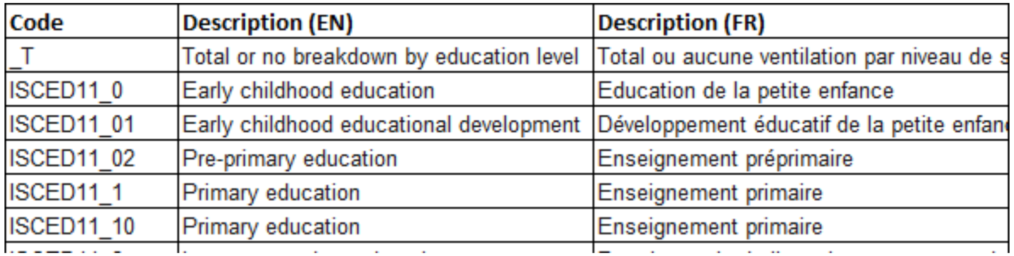
Code list

- A predefined “lookup” list that enumerates all possible values that coded concepts may take
- For example:
  - Sex code list: { M, F, _T }
  - Country code list: M49 or ISO-3-Alphanumeric country codes
  - Indicator code list, etc.
- Code lists are maintained as information units in their own right

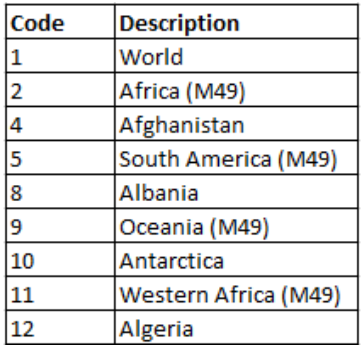
Code List: Some Examples



Code	Description
SI_POV_DAY1	Population below international poverty line (1.1.1)
SI_POV_EMP1	Employed population below international poverty line (1.1.1)
SI_POV_NAHC	Population below national poverty line (1.2.1)
SI_COV_BENFTS	Population covered by at least one social protection floor/system (1.3.1)
SI_COV_CHLD	Children covered by social protection (1.3.1)
SI_COV_DISAB	Population with severe disabilities collecting disability social protection benefits (1.3.1)
SI_COV_LMKT	Population covered by labour market programs (1.3.1)
SI_COV_MATNL	Mothers receiving maternity benefits and benefits for newborns (1.3.1)
SI_COV_PENSN	Population above retirement age receiving a pension (1.3.1)



Code	Description (EN)	Description (FR)
_T	Total or no breakdown by education level	Total ou aucune ventilation par niveau de s
ISCED11_0	Early childhood education	Education de la petite enfance
ISCED11_01	Early childhood educational development	Développement éducatif de la petite enfan
ISCED11_02	Pre-primary education	Enseignement préprimaire
ISCED11_1	Primary education	Enseignement primaire
ISCED11_10	Primary education	Enseignement primaire



Code	Description
1	World
2	Africa (M49)
4	Afghanistan
5	South America (M49)
8	Albania
9	Oceania (M49)
10	Antarctica
11	Western Africa (M49)
12	Algeria

Un-coded concepts

- Can be free-text: Any valid text can be used as a value for the concept.
  - E.g.: Footnote, Source…
- Can have their format specified
  - Postal code: 5 digits
  - Date: YYYY-MM-DD

Representation of concepts in SDMX

- **Dimensions** must be either coded or have their format specified.

→ Free text is not allowed.

- **Attributes** can be coded or un-coded; specifying the format of un-coded attributes is optional

Exercise 2: Representation

- Working with your model, determine representation for each concept
  - Coded, formatted, free-text
- Develop code lists and formats for your concepts
  - Use any approach for your codes

Importance of Data Models

- A data model, represented by the Data Structure Definition, defines what data can be encoded and transmitted.
- Flaws in a DSD may have significant adverse impact on data exchange
  - Missing concepts
  - Incorrect role of concepts
  - Un-optimized model

Data Structure Definition: Design Considerations

- Unambiguousness
  - Data must retain meaning outside usual context
  - *Do you need to supply country code with your data? (Yes)*

Source: **Guidelines for the Design of SDMX Data Structure Definitions**

Data Structure Definition: Design Considerations

- Density
  - Model should be such that data could be supplied for most or all of possible combinations of key values

Source: **Guidelines for the Design of SDMX Data Structure Definitions**

Data Structure Definition: Design Considerations

- Orthogonality
  - Concepts should be independent of each other

Source: **Guidelines for the Design of SDMX Data Structure Definitions**

Data Structure Definition: Design Considerations

- Parsimony
  - No redundant dimensions
  - Attributes attached at the highest possible level

Source: **Guidelines for the Design of SDMX Data Structure Definitions**

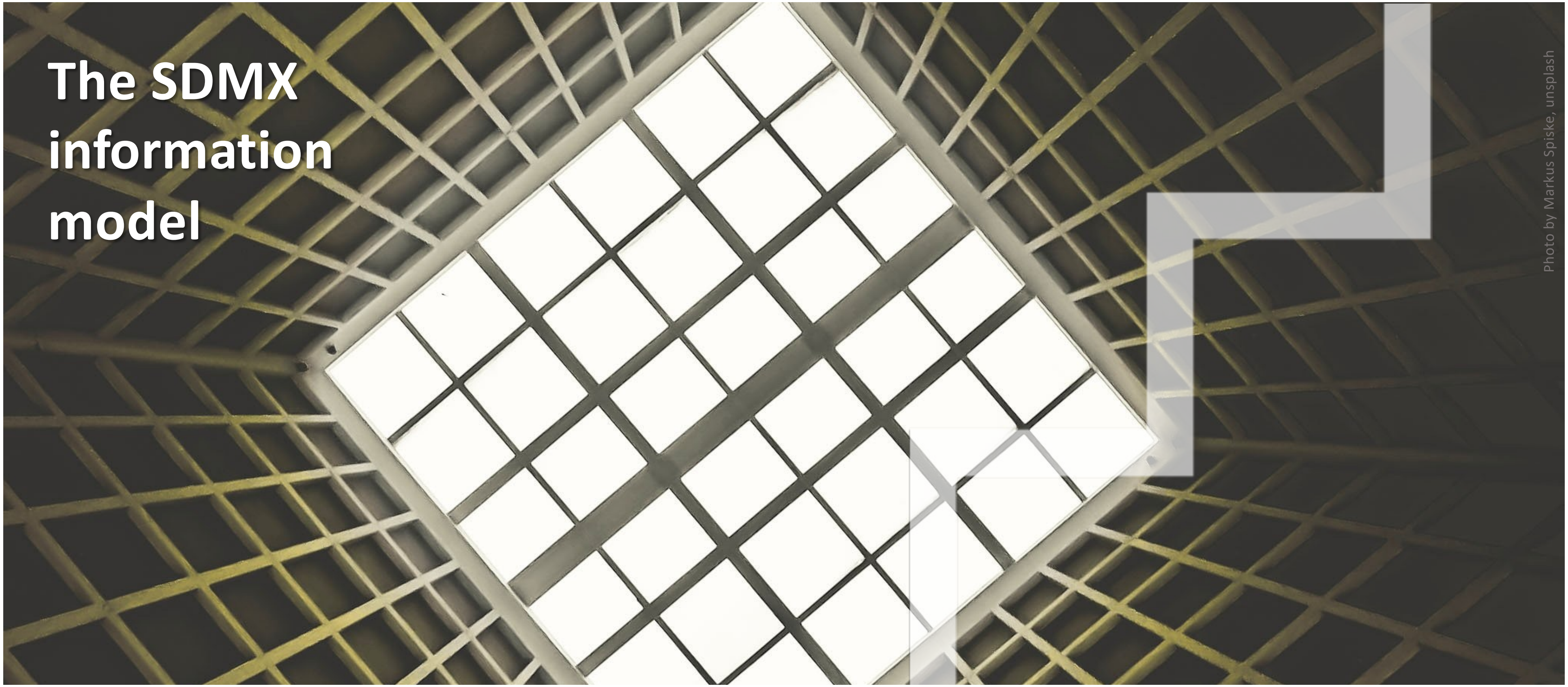
Data Structure Definition: Design Considerations

- “Mixed dimensions”
  - Can be used to minimize the number of dimensions
  - Can help avoid invalid combinations of key values
  - Should be used with caution!

DSD Design Tradeoffs: Simplicity vs Purity

- A *simple* model may increase maintenance costs
    - New codes may need to be added more frequently
    - Difficult to map and consume
- A *pure* model may increase the number of errors due its lower *density*
    - Some combinations of key values may be impossible in reality, but valid from the DSD point of view
- Splitting the *pure* model into multiple DSDs to improve *density* may increase maintenance costs
    - Multiple DSDs and other artefacts need to be maintained

United Nations | Department of Economic and Social Affairs
Statistics



The SDMX information model

Photo by Markus Spiske, unsplash

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 data exchange
- Specifies a standard structure for all datasets in a specific domain / data flow
- A DSD contains:
    - A list of the **concepts** that pertain to the dataset
    - Any **code lists** used to represent the concepts in the dataset
    - **Dimensional structure** (describing the role of each concept as dimension or attribute)
    - **Groups** that describe collections of individual observations within a dataset (e.g., a specific time series or a cross-section within a spatio-temporal data panel)

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

- Lists concepts for:
    - Dimensions
    - Attributes
    - Measure(s)
- Links concepts to code lists
- Defines groups
- Defines attachment levels of attributes

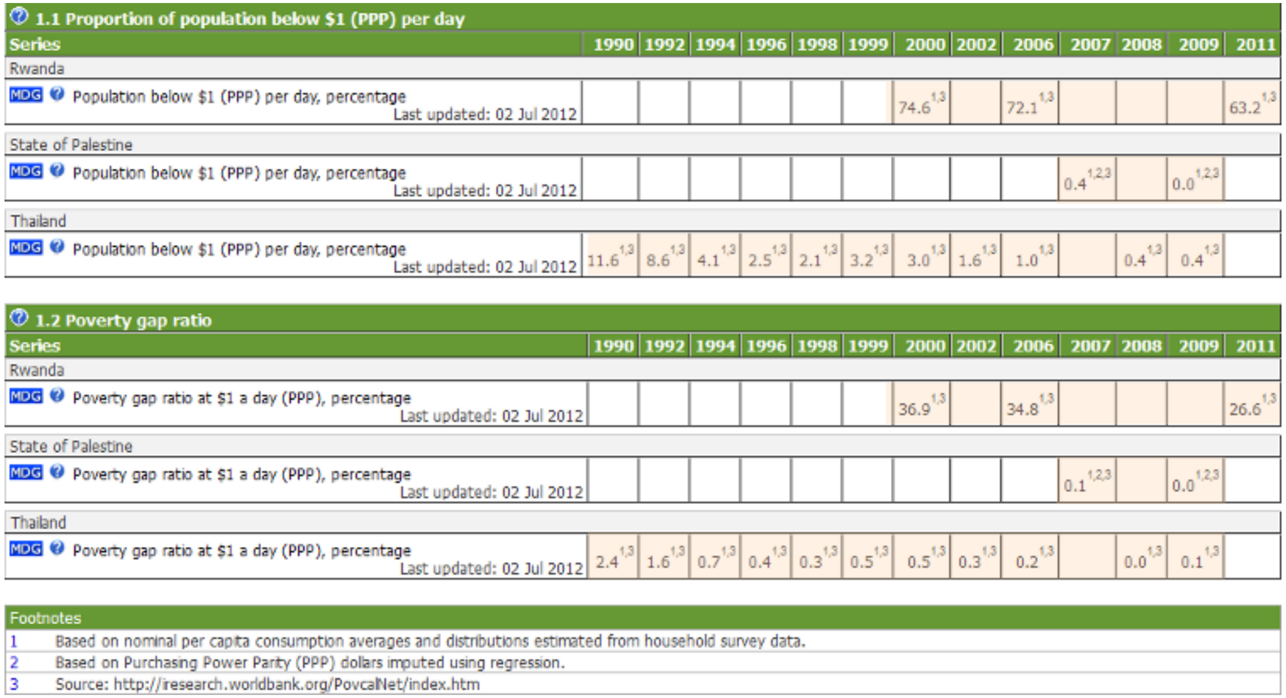
Groups

- Groups define *partial keys* that can be used to attach metadata to a subset of observations within a dataset
- 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 currently used in the SDG DSD

Time Series

- An ordered set of observations on the same variable, taken at different points in time.
- Observations that belong to the same time series, only 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

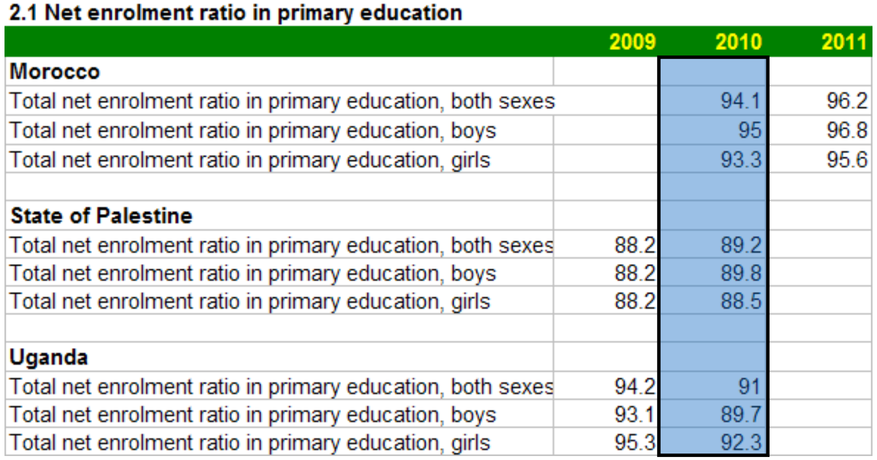
Series	1990	1992	1994	1996	1998	1999	2000	2002	2006	2007	2008	2009	2011
Rwanda													
MDG Poverty gap ratio at $1 a day (PPP), percentage Last updated: 02 Jul 2012							36.9[1,3]		34.8[1,3]				26.6[1,3]
State of Palestine													
MDG Poverty gap ratio at $1 a day (PPP), percentage Last updated: 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 set of observations that refer to the same time period, but take on different values for one or more non-time dimensions (such as geographic area or type of economic activity)
- E.g., all the data points in a specific survey or census usually refer to a fixed time period, and tabulations are organized along dimensions such as location, sex, industry, etc.

Time Series View vs Cross-Sectional View

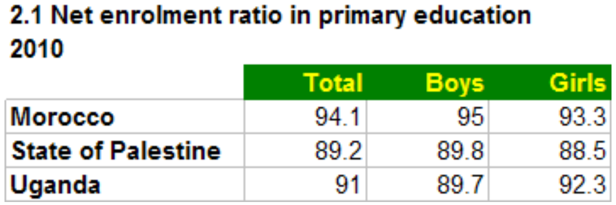


2.1 Net enrolment ratio in primary education

	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 country and sex dimensions were chosen as the cross-sectional dimensions.

Note that the time dimension is still applicable (but only takes a single, fixed value)



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 identify a time series
  - Consists of all dimensions except **TIME**
- **Group key** uniquely identifies a group
  - Consists of one or more dimensions (which may or may not include TIME)

Dataset

- A collection of logically related data, sharing a common structure pertaining to a specific set of geographic regions and periods of time
  - It can be thought of as a collection of one or more time series and/or one or more cross-sectional groups of observations
- A “dataset” serves, e.g., as a container for multiple time series being exchanged in a single data flow (e.g., in an SDMX message)

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.

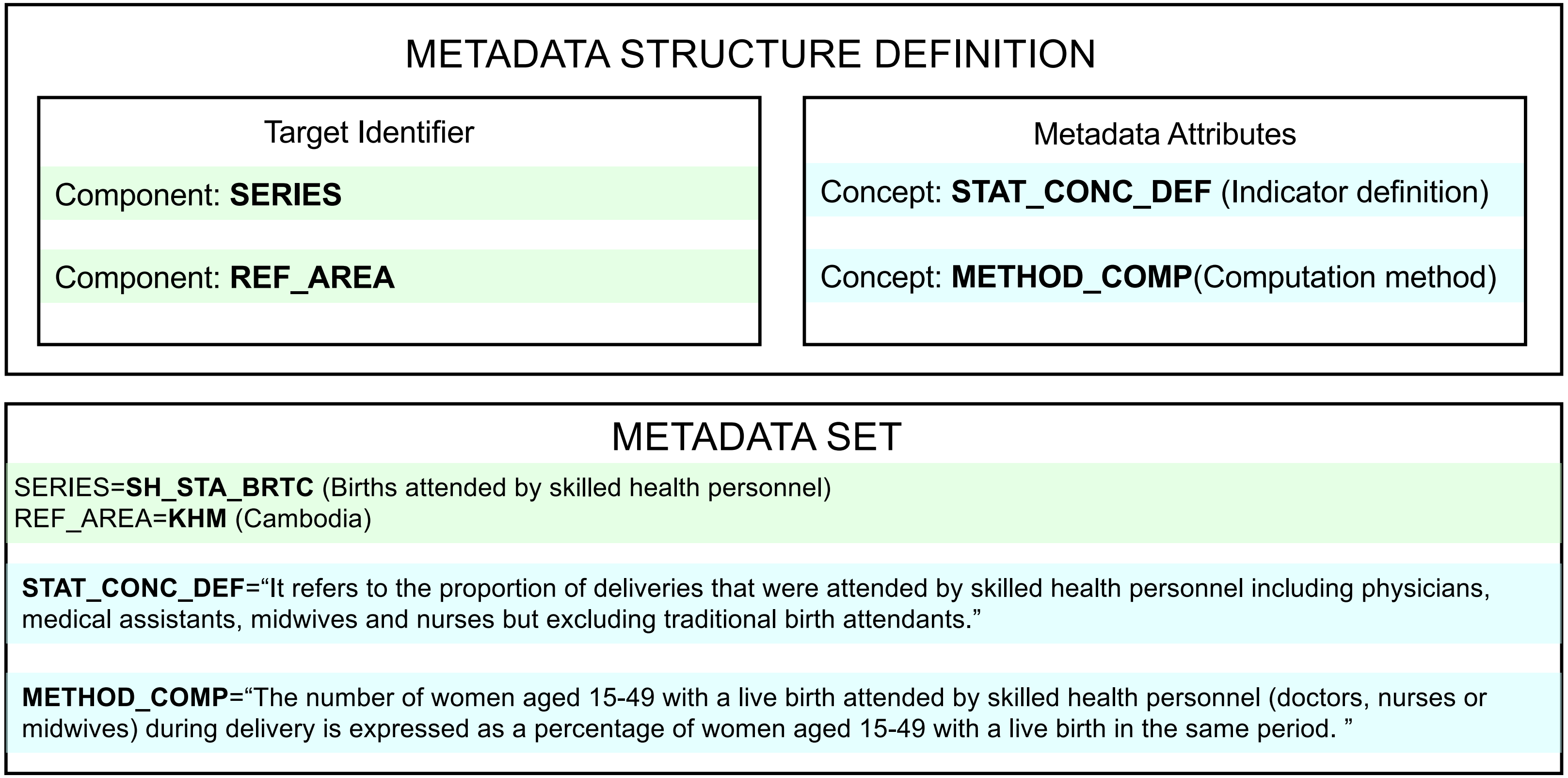
Metadata in SDMX

- 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 series for each country
    - Concepts used to express metadata (“**metadata attributes**”).
        - E.g. Indicator Definition, Quality Management

Metadata Structure Definition and Metadata Set



METADATA STRUCTURE DEFINITION

Target Identifier

Component: **SERIES**

Component: **REF_AREA**

Metadata Attributes

Concept: **STAT_CONC_DEF** (Indicator definition)

Concept: **METHOD_COMP**(Computation method)

METADATA SET

SERIES=**SH_STA_BRTC** (Births attended by skilled health personnel)
REF_AREA=**KHM** (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 documents called *messages*.
- An SDMX message can be sent various standard formats (e.g. XML, JSON, CSV)
- Different types of SDMX messages serve different purposes:
  - **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.

Keys in SDMX

- **Series key** uniquely identify a time series
  - Consists of all dimensions except **TIME**
- **Group key** uniquely identifies a group
  - Consists of one or more dimensions (which may or may not include TIME)

Dataset

- A collection of logically related data, sharing a common structure pertaining to a specific set of geographic regions and periods of time
  - It can be thought of as a collection of one or more time series and/or one or more cross-sectional groups of observations
- A "dataset" serves, e.g., as a container for multiple time series being exchanged in a single data flow (e.g., in an SDMX message)

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.

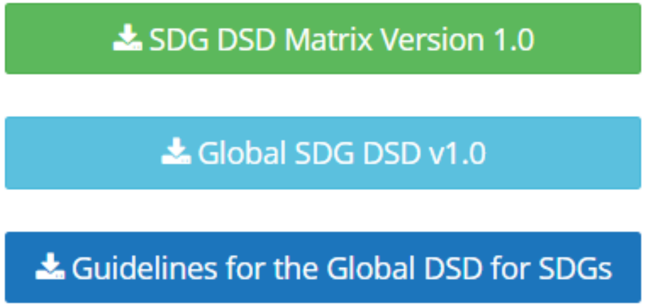
United Nations | Department of Economic and Social Affairs
Statistics

The SDG Data Structure Definition

1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
THE GLOBAL GOALS
For Sustainable Development

SDG Data Structure Definition

- Developed by the Working Group on SDMX for SDG Indicators, established by IAEG-SDGs in April 2016
- First version officially released on 14 June 2019


SDG DSD Matrix Version 1.0
Global SDG DSD v1.0
Guidelines for the Global DSD for SDGs

https://unstats.un.org/sdgs/iaeg-sdgs/sdmx-working-group/

SDG Data Structure Definition

- One single DSD is used for all SDG indicators
- Support for diverse indicators means not all dimensions are applicable in all cases
  - E.g. AGE is not applicable to indicator “Land area covered by forest”
  - Value **_T** (no breakdown) is used when an dimension is not applicable.

Dimension: Frequency (FREQ)

- “Indicates rate of recurrence at which observations occur (e.g. monthly, yearly, biannually, etc.).”
- By convention, SDGs DSD currently only supports annual frequency.
- Where the frequency is not annual (e.g. two-year average), detail should be provided in the TIME_DETAIL attribute.

Dimension: REPORTING_TYPE

- Used to distinguish between National, Regional, Global Reporting
- Countries to use value **N** (national reporting)
- Regional organizations to use value **R** (regional reporting)
- Custodian agencies to use value **G** (Global reporting)

Dimension: Series (SERIES)

- Used to represent "sub-indicators"
  - A single indicator can have multiple series
  - Not to be confused with SDMX time series (each series can have multiple time series, i.e., multiple disaggregation with observations organized over time)
- Example: Indicator 5.5.1, "Proportion of seats held by women in (a) national parliaments and (b) local governments" has 4 series:
  - SG_GEN_PARL Proportion of seats held by women in national parliaments
  - SG_GEN_PARLN Number of seats held by women in national parliaments
  - SG_GEN_PARLNT Number of seats in national parliaments
  - SG_GEN_LOCG Proportion of seats held by women in local governments

Dimension: Reference Area (REF_AREA)

- Country or geographic area to which the measured statistical phenomenon relates
- It is envisaged that countries will report national-level values but may wish to extend the code list with its sub-national areas for dissemination

Dimension: Sex (SEX)

- Gender condition: male or female. This dimension applies only if data can be disaggregated by sex.
- Use **_T** where not applicable
- For gender indicators must be set to **F** as applicable
  - E.g. for series *Proportion of seats held by women in national parliaments*

Dimension: Age (AGE)

- “Age - or age range - of the individuals the observation refers to.”
- Use **_T** where not applicable

Dimension: Urban/Rural location (URBANISATION)

- Has 3 codes
    - _T (Total)
    - _U (Urban)
    - _R (Rural)
- Use _T where not applicable

Dimension: INCOME_WEALTH_QUANTILE

- Used for disaggregating the data by income or wealth quintile of the population
- In the future can be extended to cover decile, percentile, etc
- Use _T where not applicable

Dimension: Education Level (EDUCATION_LEV)

- “Highest level of an educational programme the person has successfully completed.”
- Supports top categories of ISCED11 and ISCED97, as well as custom SDG codes
- Use _T where not applicable

Dimension: OCCUPATION

- “Job or position held by an individual who performs a set of tasks and duties.”
- Supports top categories of ISCO-08, ISCO-98, ISCO-68
- Use _T where not applicable

Dimension: Disability Status (DISABILITY STATUS)

- Used to break down SDG indicators by disability
- At the moment, only used to distinguish between persons with a disability, and persons without a disability
- Use _T where not applicable

Dimension: Economic Activity (ACTIVITY)

- "High-level grouping of economic activities based on the types of goods and services produced."
- Consists of top-level ISIC categories.
- Use **_T** where not applicable.

Dimension: Product Type (PRODUCT)

- Product or commodity code
- Combines SDG-specific entries from several classifications including CPC, Material Flows, and non-standard
- Use **_T** where not applicable

Dimension: Custom Breakdown (CUST_BREAKDOWN)

- Special dimension introduced to facilitate non-standard breakdowns, primarily in national context
- At the moment empty but in the future will be populated with generic codes (e.g. CODE1, CODE2, etc), to which data providers will assign meaning in their own context
- Used in conjunction with attribute CUST_BREAKDOWN_LB, which transmits description of the custom code.
- Use **_T** where not applicable

Dimension: COMPOSITE_BREAKDOWN

- Mixed dimension: represents several merged code lists
    - E.g. by International Organizations, Hazard Type etc
- Used for breakdowns that are only used in 1 or 2 indicators, in order to avoid creating too many dimensions
- Use _T where not applicable

Time Dimension: TIME_PERIOD

- The observation corresponds to a specific point in time … or a period…”
- The convention for SDGs is to always provide a four-digit year in the TIME_PERIOD concept. Further info must be placed in TIME_DETAIL, and structured period information in TIME_COVERAGE.

Primary Measure: Observation value (OBS_VALUE)

- Used to convey the value of a variable at a period of time
- Should be a floating-point number

Attribute: Observation Status (OBS_STATUS)

- “Information on the quality of a value or an unusual or missing value”
  - E.g. can be used to indicate a break in series
- Mandatory observation-level attribute

Attribute: Unit Multiplier (UNIT_MULT)

- Exponent in base 10 specified so that multiplying the observation numeric values by **10^UNIT_MULT** gives a value expressed in the unit of measure
- If the observation value is in millions, unit multiplier is 6; if in billions, 9, and so on. Where the number is simple units, use 0.
- Mandatory observation-level attribute

Attribute: Unit of Measure (UNIT_MEASURE)

- Unit in which the data values are expressed
- It may not be obvious which is the correct unit in some cases. Coding guidelines are available and will be further developed.
- Mandatory time series-level attribute

Attribute: Time Period Details (TIME_DETAIL)

- “When TIME_PERIOD refers to a date range, this attribute is used to provide metadata on the actual range the observation refers to (e.g. for period ‘2001-2003’ TIME_PERIOD would be 2002 but the actual dates --2001-2003-- would be expressed here).”
- Optional observation-level free-text attribute

Attribute: TIME_COVERAGE

- ISO8601 representation of the actual time interval to which the observation refers
- While TIME_PERIOD should always be expressed as a year, and TIME_DETAIL is free-text with additional information,   TIME_COVERAGE can optionally be used to provide the exact interval in a structured format
- Optional observation-level attribute.

Attribute: Base Period (BASE_PER)

- Period of time used as the base of an index number, or to which a constant series refers
- Where a base period applies, it is expected to always be set to a year
- Typically, used for constant prices, as in “2005 USD dollar”
- Optional observation-level attribute.

Attribute: Nature of data points (NATURE)

- Information on the production and dissemination of the data
- Expresses whether a data point has been produced and disseminated by the country, estimated by international agencies, etc.
- Normally set to C (Country Data) in national reporting
- Optional observation-level attribute

Attribute: Source details (SOURCE_DETAIL)

- Provides additional textual information on the data source, e.g. a specific survey that was used to generate the indicator.
- Optional observation-level free-text attribute.

Attributes: UPPER_BOUND and LOWER_BOUND

- Where the observation value represents a point estimate, can be used to convey the Upper and Lower bounds
  - In MDG DSD, separate series codes had to be created for upper and lower bounds
- Optional observation-level attributes

Attributes: Footnotes (COMMENT_OBS and COMMENT_TS)

- “Additional information on specific aspects of each observation, such as how the observation was computed/estimated or details that could affect the comparability of this data point with others in a time series.”
- Attribute COMMENT_OBS is used for observation-level footnotes, and COMMENT_TS for time series-level footnotes. Both are optional.

Attribute: GEO_INFO_URL

- Provides web address of a geoinformation file. Used in conjunction with attribute GEO_INFO_TYPE.
- Optional time series-level attribute.

Attribute: GEO_INFO_TYPE

- Specifies type of geoinformation file provided in attribute GEO_INFO_URL.
- Optional time series-level attribute.

SDG DSD: Mappings

- Due to its support for heterogeneous indicators, it’s not always obvious which values should be used in some dimensions
- What should be SEX in indicator “Births attended by skilled personnel”:
  - Not Applicable? Total? Female?

SDG DSD: Mappings

- Inconsistent mappings lead to duplications and other anomalies
- Coding guidelines will be developed and enforced through content constraints
- The use of a single code for no breakdown (e.g. for Total and Not Applicable) simplifies the mappings.

United Nations | DESA Statistics Division

Thank you.