United Nations

Department of Economic and Social Affairs
Statistics

# The SDMX information model

Photo by Markus Spiske, unsplash

# Figures vs data

**Number of touristic establishments in Italy, annual data**

Number Tourism establishments Italy Annual data

| Indicator<br>Time | A100<br>Hotels and similar | B010<br>Tourist Campsites | B020<br>Holiday dwellings |
|---|---|---|---|
| **2002A00** | 33411 | 2374 | 61479 |
| **2003A00** | 33480 | 2530 | 58526 |
| **2004A00** | 33518 | 2529 | 56586 |
| **2005A00** | 33527 | 2411 | 68385 |
| **2006A00** | 33768 | 2510 | 68376 |
| **2007A00** | 34058 | 2587 | 61810 |

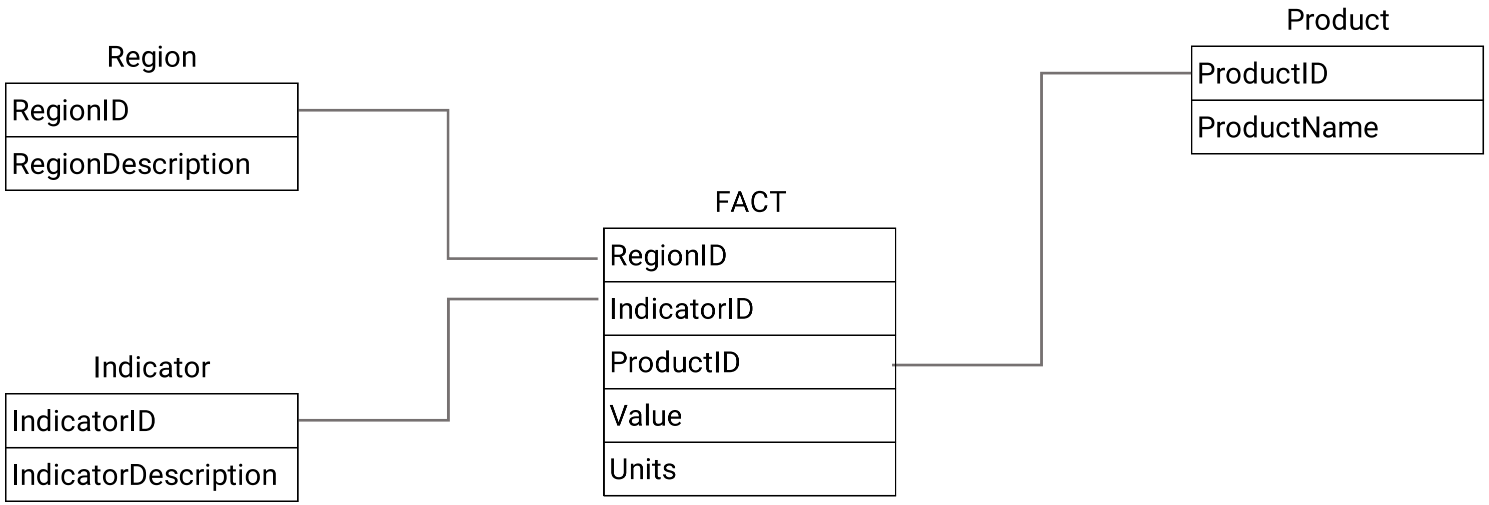
- Figures by themselves are meaningless.
- For data to be usable, it must be properly described. The descriptions let users know what the data actually represents.

# Developing a Data Model for Data Exchange

- Data model is developed to provide descriptions for all relevant characteristics of the data to be exchanged
- In some aspects similar to developing a relational database
- In SDMX, data model is represented by a Data Structure Definition (DSD).
- The “shape” of SDMX DSD is roughly similar to star schema.
- To design a DSD, we first need to find *concepts* that identify and describe our 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**”



| Region |
|---|
| RegionID |
| RegionDescription |

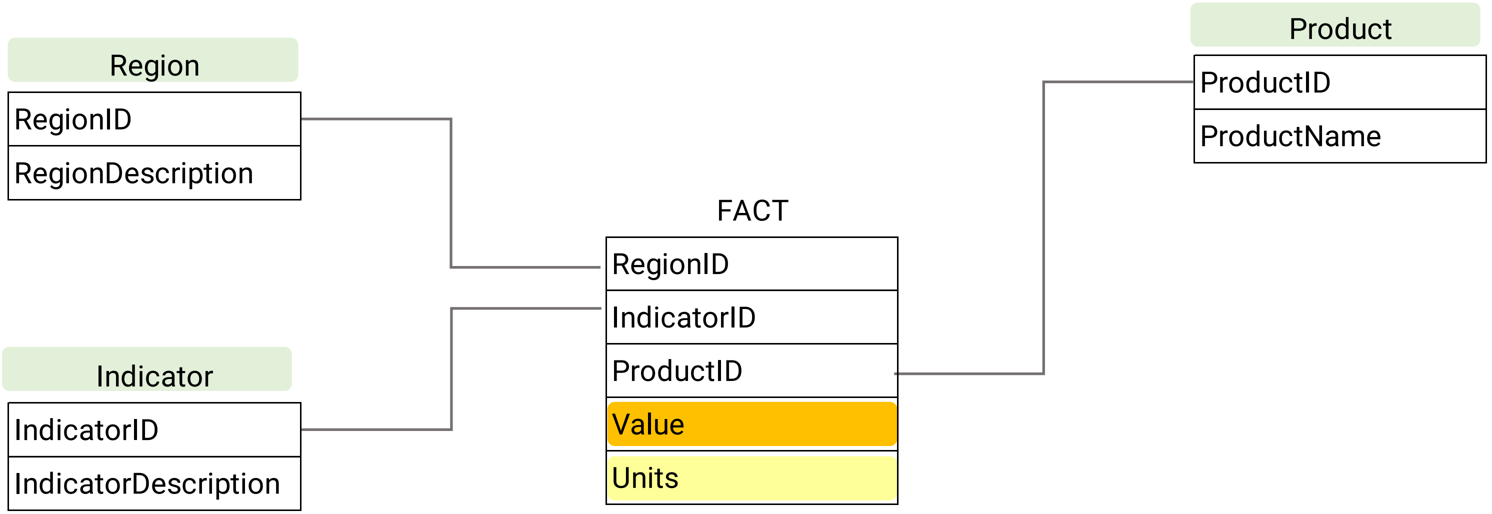
| Indicator |
|---|
| IndicatorID |
| IndicatorDescription |

| FACT |
|---|
| RegionID |
| IndicatorID |
| ProductID |
| Value |
| Units |

| Product |
|---|
| ProductID |
| ProductName |

# Concept

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



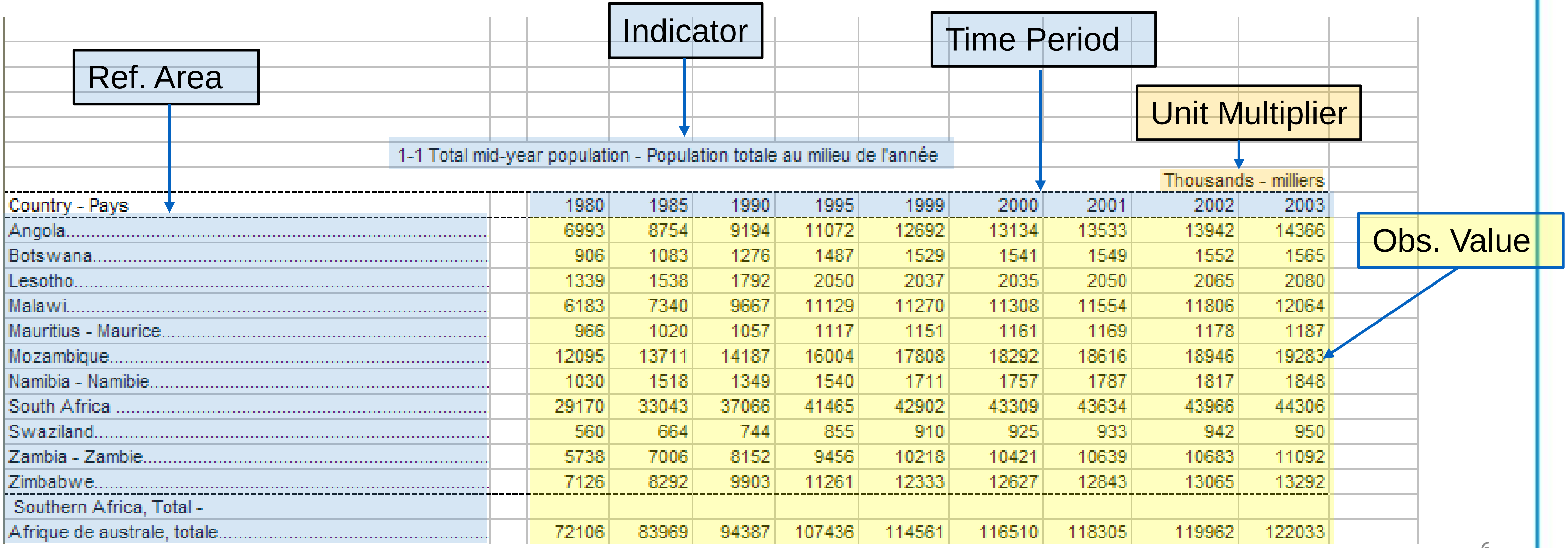
| Region |
|---|
| RegionID |
| RegionDescription |

| 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 | 9667 | 11129 | 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 -<br>Afrique de australe, totale | 72106 | 83969 | 94387 | 107436 | 114561 | 116510 | 118305 | 119962 | 122033 |

# SDMX Concept Scheme

- “Set of Concepts that are used in a Data Structure Definition or Metadata Structure Definition.”*
- Concept scheme places concepts into a maintainable unit.

| Concept name | Concept ID |
|---|---|
| Indicator | INDICATOR |
| Reference area | REF_AREA |
| Time period | TIME_PERIOD |
| Unit multiplier | UNIT_MULT |
| Observation value | OBS_VALUE |

* Source: **SDMX Glossary**

# Dimension

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

# 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 are called **attributes**.
  - Not to be confused with XML attributes!
  - Similar to a database table’s non-primary key fields.

# 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**

# 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 with ID **OBS_VALUE**.

# Dimension or Attribute?

- Choosing the role of a concept has profound implications on the structure of data.
- Concepts that identify data, should be made dimensions. Concepts that provide additional information about data, should be made attributes.
- If a concept is a dimension, it is possible to have time series that are different only in the value of this concept.
  - E.g. if Unit of Measure is a dimension, it is possible to have separate series for “T” and “T/HA” or, more controversially, “KG” and “T”

# Dimension or Attribute? (2)

| Cambodia | | |
|---|---|---|
| Fixed and Mobile telephone subscriptions | 2013 | 20.6 million |
| Fixed and Mobile telephone subscriptions | 2012 | 19.7 million |
| Fixed and Mobile telephone subscriptions | 2013 | 140.9 per 100 pop. |

Unit of measure as a dimension…

| Ref.Area | Indicator | Time Period | Unit | Unit Mult. | Obs. Value |
|---|---|---|---|---|---|
| Cambodia | Fixed and Mobile telephone subscriptions | 2013 | Number | Millions | 20.6 |
| Cambodia | Fixed and Mobile telephone subscriptions | 2012 | Number | Millions | 19.7 |
| Cambodia | Fixed and Mobile telephone subscriptions | 2013 | Per 100 pop. | Units | 140.9 |

# Dimension or Attribute? (3)

Unit of measure as an attribute…

Violation!

| Ref.Area | Indicator | Time Period | Unit | Unit Mult. | Obs. Value |
|---|---|---|---|---|---|
| Cambodia | Fixed and Mobile telephone subscriptions | 2013 | Number | Millions | 20.6 |
| Cambodia | Fixed and Mobile telephone subscriptions | 2012 | Number | Millions | 19.7 |
| Cambodia | Fixed and Mobile telephone subscriptions | 2013 | Per 100 pop. | Units | 140.9 |

- The dataset above is invalid: duplicate observation
- The two values above are only different in their attributes

# Dimension or Attribute? (4)

Unit of measure as an attribute…

👍

| Ref.Area | Indicator | Time Period | Unit | Unit Mult. | Obs. Value |
|---|---|---|---|---|---|
| Cambodia | Fixed and Mobile telephone subscriptions | 2013 | Number | Millions | 20.6 |
| Cambodia | Fixed and Mobile telephone subscriptions | 2012 | Number | Millions | 19.7 |
| Cambodia | Fixed and Mobile telephone subscriptions per 100 population | 2013 | Per 100 pop. | Units | 140.9 |

- Now there is no violation because every row has a unique key
- The Unit concept is still useful

# Attribute attachment

- In SDMX 2.0, attributes can be attached at observation, time series, group, or dataset level.
- In SDMX 2.1, attributes can be attached at observation, dimension(s), group, or dataset.
  - When attribute is attached to all dimensions except time, it is effectively attached to time series
- For practical purposes attributes are often attached at observation or time series.

# Data model so far…

| Concept | ID | Role | Attachment |
|---|---|---|---|
| Indicator | INDICATOR | Dimension | |
| Reference area | REF_AREA | Dimension | |
| Time period | TIME_PERIOD | Dimension | |
| Unit multiplier | UNIT_MULT | Attribute | Time series |
| Observation value | OBS_VALUE | P.Measure | |

# 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

- DSD defines a range of valid values for each concept.
- When data are transferred, each of its descriptor concepts must have valid values.
- A concept can be
  - Coded
  - Un-coded with format
  - Un-coded free text

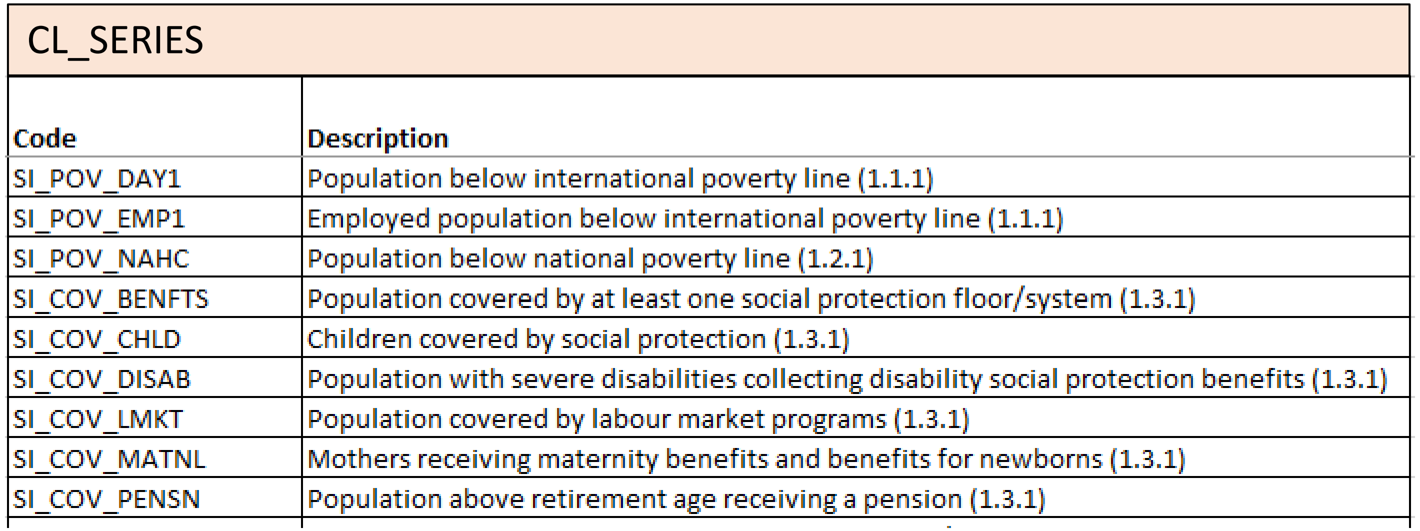
# Code

- “A language-independent set of letters, numbers or symbols that represent a concept whose meaning is described in a natural language.”
- A sequence of characters that can be associated with a descriptions in any number of languages.
    - Descriptions can be updated without disrupting mappings or other components of data exchange.

# Code List

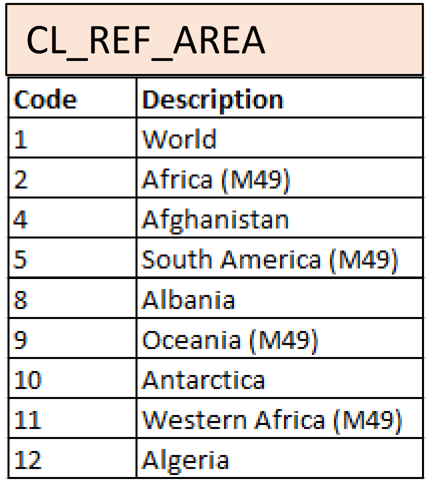
- “A predefined list from which some statistical coded concepts take their values.”
- A code list is a collection of codes maintained as a unit.
- A code list enumerates all possible values for a concept or set of concepts
  - Sex code list
  - Country code list
  - Indicator code list, etc

# Code List: Some Examples



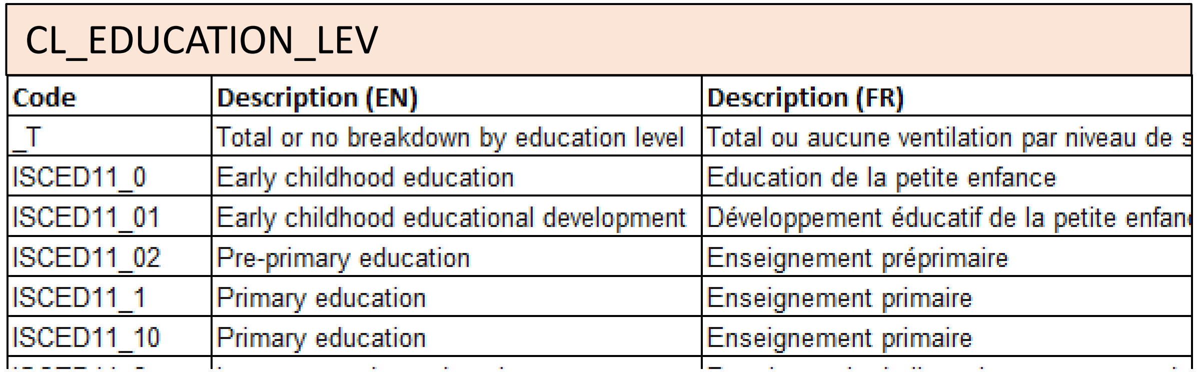
CL_SERIES

| 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) |



CL_REF_AREA

| 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 |



CL_EDUCATION_LEV

| 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 |

# SDMX Concepts and Code lists

- Code lists provide a 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.

## Un-coded Concepts

- Can be free-text: Any valid text can be used as a value for the concept.
  - Footnote
- Can have their format specified
  - Postal code: 5 digits
  - Last update: date/time

# 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; format may optionally be specified.

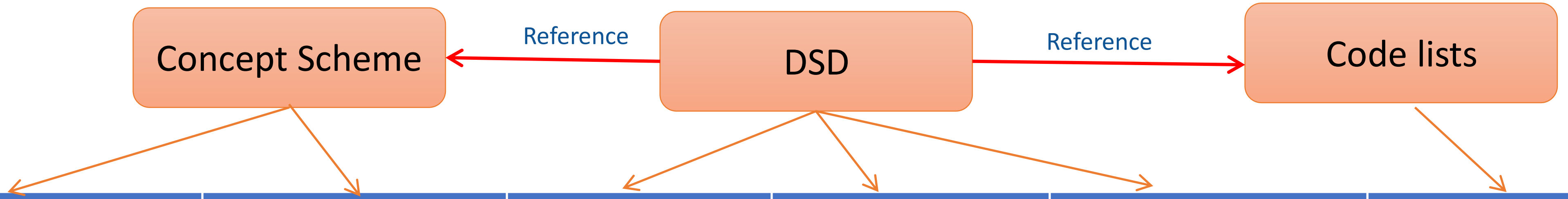
# Data model so far…

| Concept | ID | Role | Attachment | Representation |
|---|---|---|---|---|
| Indicator | INDICATOR | Dimension | | CL_INDICATOR |
| Reference area | REF_AREA | Dimension | | CL_REF_AREA |
| Time period | TIME_PERIOD | Dimension | | Date/time (YYYY) |
| Unit multiplier | UNIT_MULT | Attribute | Time series | CL_UNIT_MULT |
| Observation value | OBS_VALUE | Pr. Measure | | Floating point number |

# Exercise 2: Representation

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

# Data Structure Definition: summary



| Concept | Concept ID | Role | Attachment | Representation | Code list ID |
|---|---|---|---|---|---|
| Indicator | INDICATOR | Dimension | | Code List | CL_INDICATOR |
| Reference area | REF_AREA | Dimension | | Code List | CL_REF_AREA |
| Time period | TIME_PERIOD | Dimension | | YYYY | |
| Unit multiplier | UNIT_MULT | Attribute | Time series | Code List | CL_UNIT_MULT |
| Obs. value | OBS_VALUE | Pr. Measure | | Number | |

# Importance of Data Model

- Data model, represented by DSD, 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

- Parsimony
    - No redundant dimensions
    - Attributes attached at the highest possible level
- Simplicity
    - “Mixed dimensions” are used to minimize the number of dimensions
    - Can help avoid invalid combinations of key values
    - Should be used with caution
    - Opposite of “purity”

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

# Data Structure Definition: Design Considerations (2)

- Unambiguousness
  - Data must retain meaning outside usual context
  - Do you supply country code with your data?
- Density
  - Model should be such that data could be supplied for most or all of possible combinations of key values
  - Related to simplicity
- Orthogonality
  - Meaning of the value of concepts should be independent of each other
  - Helps avoid ambiguity

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

# DSD Design Tradeoffs: Simplicity vs Purity

- A *simple* model may increase maintenance costs
  - Codes frequently need to be added
  - Difficult to map and consume
- A *pure* model may increase the number of errors due its lower *density*
  - Some combinations of key values are 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

# Dataset

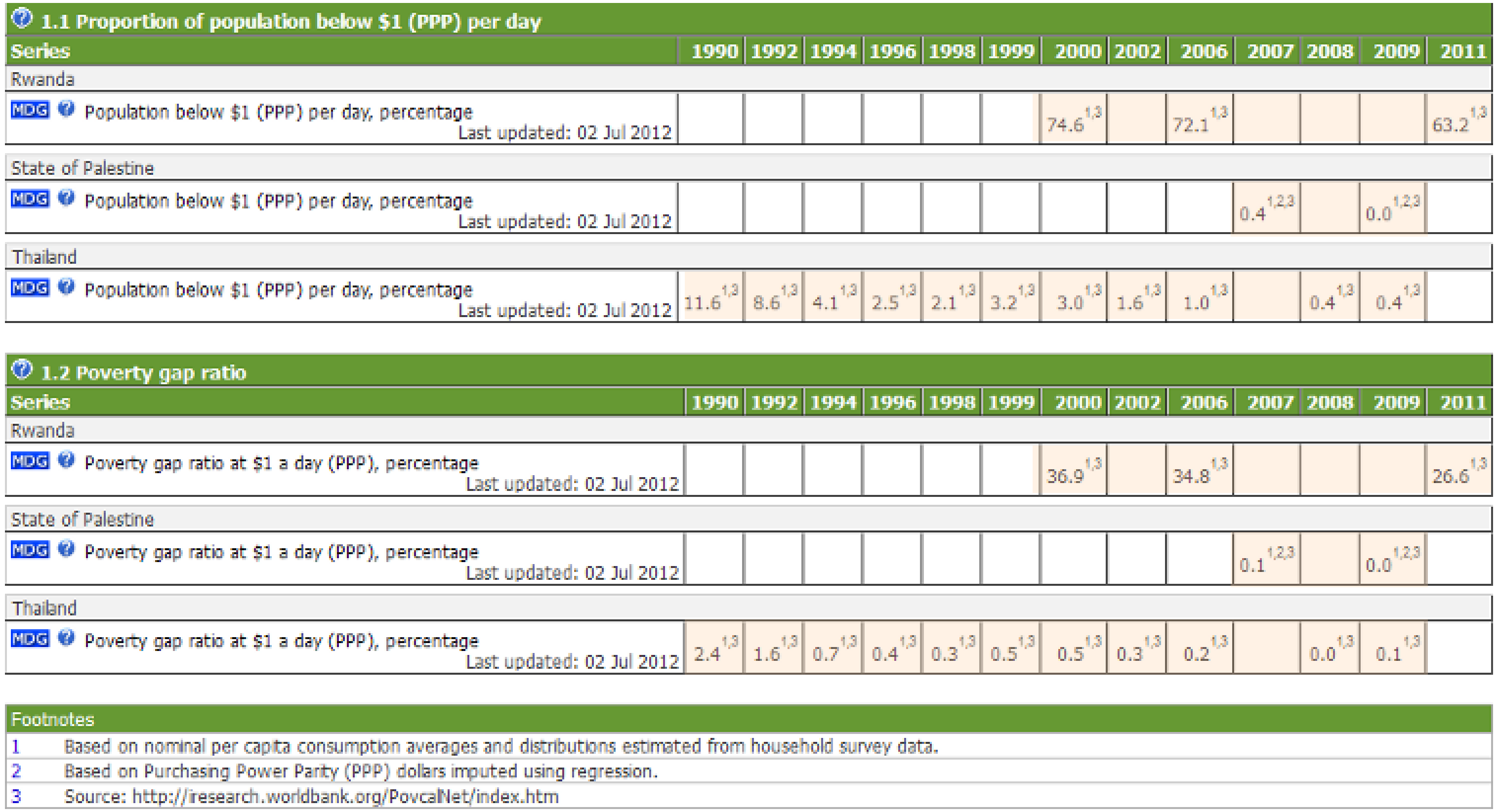
- Organised collection of data defined by a Data Structure Definition (DSD)*
- A dataset is structured in accordance with *one* DSD
- Serves as a container for time-series or cross-sectional series in SDMX data messages.

*Source: **SDMX Glossary**

# 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<br>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<br>Last updated: 02 Jul 2012 | | | | | | | | | | 0.4[1,2,3] | | 0.0[1,2,3] | |
| Thailand | | | | | | | | | | | | | |
| MDG Population below $1 (PPP) per day, percentage<br>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**

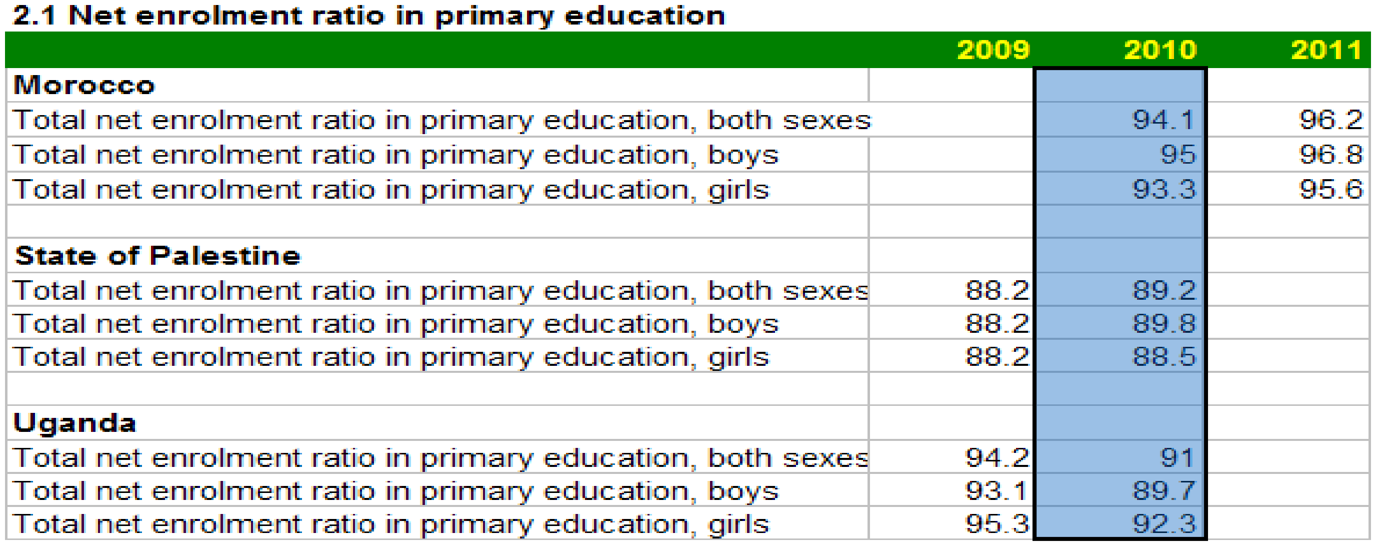
| Series | 1990 | 1992 | 1994 | 1996 | 1998 | 1999 | 2000 | 2002 | 2006 | 2007 | 2008 | 2009 | 2011 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rwanda | | | | | | | | | | | | | |
| MDG Poverty gap ratio at $1 a day (PPP), percentage<br>Last updated: 02 Jul 2012 | | | | | | | 36.9[1,3] | | 34.8[1,3] | | | | 26.6[1,3] |
| State of Palestine | | | | | | | | | | | | | |
| MDG Poverty gap ratio at $1 a day (PPP), percentage<br>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<br>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 |

## Non-Time Series Data
## (a.k.a. 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

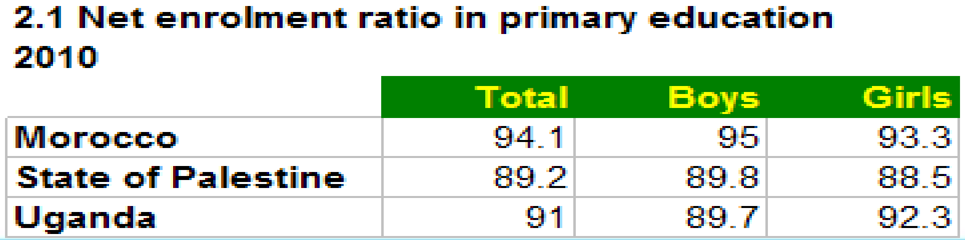


2.1 Net enrolment ratio in primary education

| | 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 series
    - In the case of 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

# Exercise 3:
# Encoding a time series

- Working with your table, determine the total number of time series.
- For the first 5 time series, provide a valid value for each concept in its series key.

## Structural and Reference Metadata

- Structural Metadata: Identifiers and Descriptors, e.g.
  - Data Structure Definition
  - Concept Scheme
  - Code

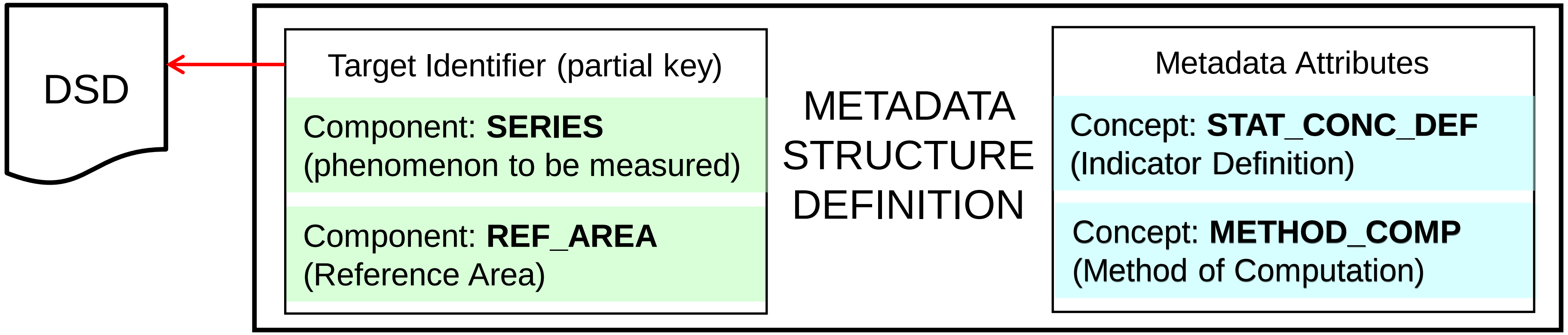
} What we have covered so far

- Reference Metadata: Describes contents and quality of data, e.g.
  - Indicator definition
  - Comments and limitations

# Reference 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 and Metadata Set: an example



DSD

METADATA STRUCTURE DEFINITION

Target Identifier (partial key)

Component: **SERIES** (phenomenon to be measured)

Component: **REF_AREA** (Reference Area)

Metadata Attributes

Concept: **STAT_CONC_DEF** (Indicator Definition)

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. ”

# Metadata Structure Definition (MSD)

- MSD Defines:
  - The object type to which reference 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

# 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.

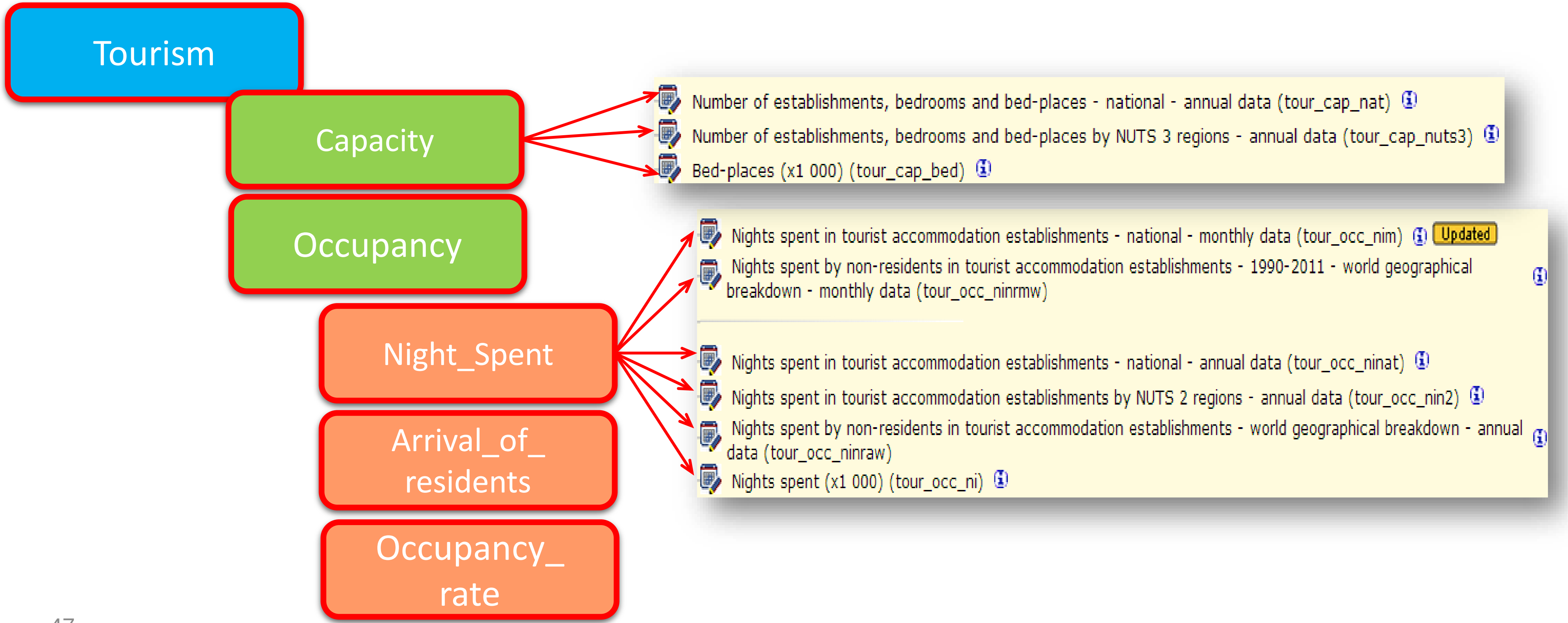
# Content Constraints

- 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

# 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.

# Dataflows - classification



Number of establishments, bedrooms and bed-places - national - annual data (tour_cap_nat)

Number of establishments, bedrooms and bed-places by NUTS 3 regions - annual data (tour_cap_nuts3)

Bed-places (x1 000) (tour_cap_bed)

Nights spent in tourist accommodation establishments - national - monthly data (tour_occ_nim) Updated

Nights spent by non-residents in tourist accommodation establishments - 1990-2011 - world geographical breakdown - monthly data (tour_occ_ninrmw)

Nights spent in tourist accommodation establishments - national - annual data (tour_occ_ninat)

Nights spent in tourist accommodation establishments by NUTS 2 regions - annual data (tour_occ_nin2)

Nights spent by non-residents in tourist accommodation establishments - world geographical breakdown - annual data (tour_occ_ninraw)

Nights spent (x1 000) (tour_occ_ni)

## 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.