# Developing a Data Model

Workshop on Data and Metadata Sharing

Bangkok, 10-14 December 2018

Abdulla Gozalov, UNSD

# Figures 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<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] | |

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

# Developing a Data Model for SDMX Exchange

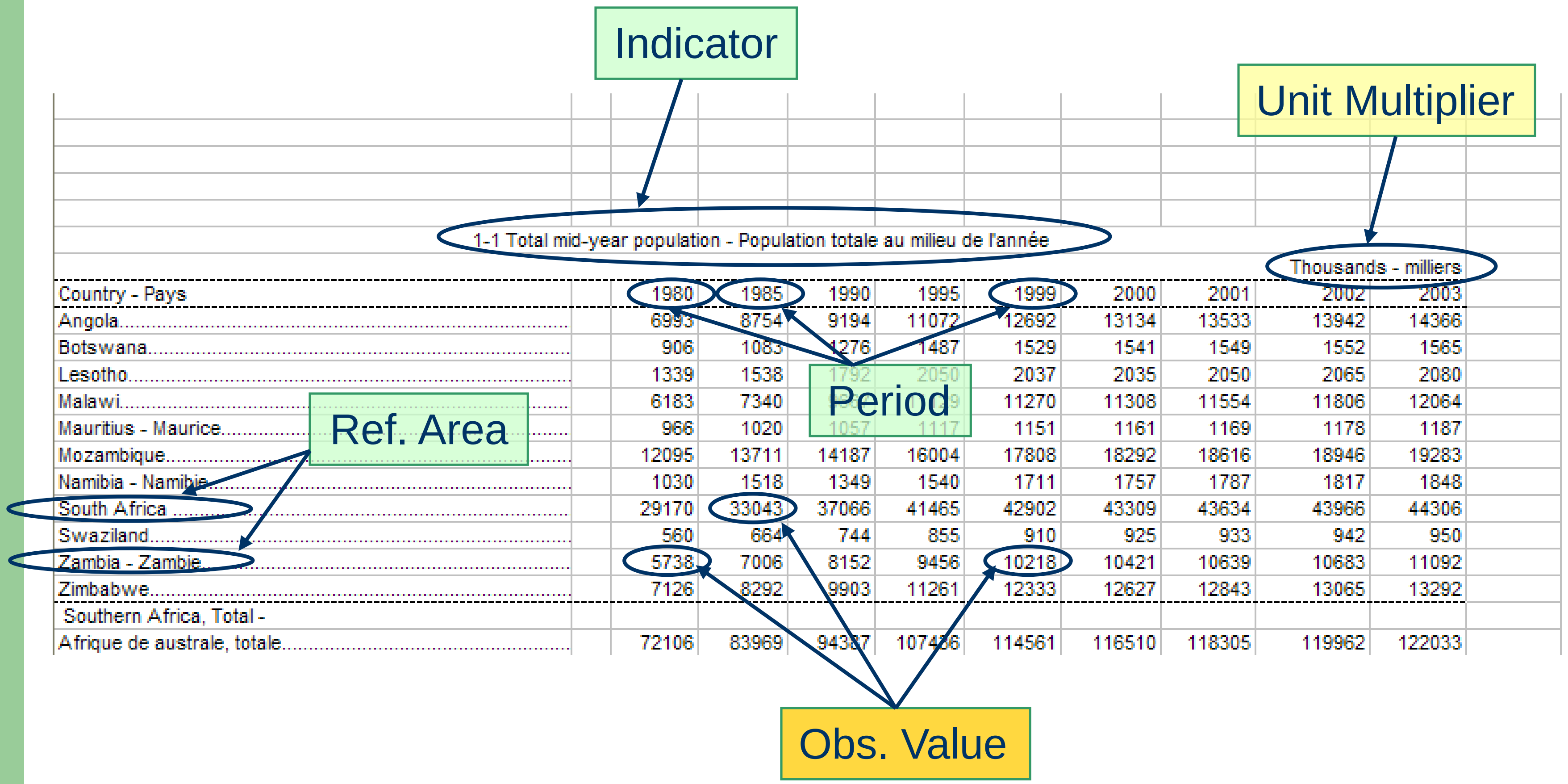
- In some aspects similar to a developing a relational database
- In SDMX, data model is represented by a Data Structure Definition.
  - The “shape” of SDMX DSD is roughly similar to star schema.
- To design a DSD, we first need to find *concepts* that identify and describe our data.

# Concept

- “A unit of knowledge created by a unique combination of characteristics”*
- Each concept describes something about the data.
- Concepts should express all relevant data characteristics.

* Source: **Metadata Common Vocabulary**

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

# 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”

# Special Dimensions

- **TIME** dimension provides observation time. If a DSD describes time series data, it must have one TIME dimension.
- **FREQUENCY** dimension describes interval between observations. If there is a TIME dimension, one other dimension must be marked as FREQUENCY dimension.

# Exercise 1: Identifying concepts

- Identify concepts in the table
- Mark each concept as:
  - Dimension
  - Time Dimension
  - Primary Measure (i.e. observation value)
  - Attribute

# 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 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

| 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.
  - Footnote
- Can have their format specified
  - Postal code: 5 digits

# 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.

# 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 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