Guidelines for the Global Data Structure Definition for Sustainable Development Goals Indicators

DSD Version: 1.0

Guidelines version: 1.0

Date: June 2019

1. Introduction

In September 2015, United Nations Member States adopted the 2030 Agenda for Sustainable Development and tasked the United Nations Statistical Commission as a functional commission of the UN Economic and Social Council to develop the global indicator framework. The overarching principle of the 2030 Agenda for Sustainable Development is that no one should be left behind. "Data which is high-quality, accessible, timely, reliable and disaggregated by income, sex, age, race, ethnicity, migration status, disability and geographic location and other characteristics relevant in national contexts" is called (A/RES/70/1).

In March 2015 at its forty-sixth session, the United Nations Statistical Commission created an Inter-agency and Expert Group on SDG Indicators (IAEG-SDGs), which is composed of representatives from a regionally-balanced group of Member States and includes regional and international agencies as well as other key stakeholders, such as civil society, academia and the private sector, as observers. The IAEG-SDGs was tasked with providing a proposal for a global indicator framework (and associated global and universal indicators) for the follow up and review of the 2030 Agenda to be considered by the Statistical Commission at its forty-seventh session in March 2016. At the forty-seventh session of the Commission, the Global indicator framework was agreed upon by the Member States.

To ensure timely and efficient collection, validation, and dissemination of SDG indicators, a data exchange format needs to be agreed upon and used by SDG data providers. This will enable the automation of data exchange while simplifying and improving data validation and dissemination.

Statistical Data and Metadata Exchange (SDMX) is a standard sponsored by seven international organizations (BIS, ECB, Eurostat, OECD, IMF, UN, and WB). It was endorsed by the United Nations Statistical Commission in 2008 as a preferred standard for data exchange, and was approved as an ISO standard (ISO/IS 17369:2013). It has been successfully used for data exchange and dissemination in areas such as Macro-Economic Statistics, International Merchandise Trade, and others including MDG indicators.

Previously, the Inter-Agency and Expert Group for MDG Indicators (IAEG-MDGs) established an SDMX Task Team, which produced a Data Structure Definition and a draft Metadata Structure Definition for MDG Indicators, i.e., commonly agreed formats for the exchange of MDG indicators. These structures have been used by several international agencies to submit MDG indicators to UNSD for their inclusion in the MDG database and report, and were also used by UNSD to establish exchange of development indicator data and metadata with 15 National Statistical Offices.

To facilitate the development of SDMX-based data and metadata exchange formats for SDG Indicators, Inter-agency Expert Group on SDG Indicators approved the Terms of Reference of the Working Group on SDMX for SDGs Indicators ("SDMX-SDGs Working Group") at its third meeting in Mexico City in April 2016. The overall objective of the Working Group is to design a solution for the exchange and dissemination of SDG Indicators based on SDMX standard.

2. Concepts and Code Lists

This section describes all concepts in the SDG Concept Scheme. The concepts are also available in the *SDG DSD Matrix* spreadsheet.

2.1. Dimensions

All dimensions in the SDG data model are coded and have an associated code list. See the *Data Model* worksheet in the *SDG Data Matrix* spreadsheet for additional detail.

Frequency (FREQ); code list: CL_FREQ

Indicates rate of recurrence at which observations occur. By convention, all SDG indicators should be provided with the Annual frequency. Where the frequency is not annual (e.g. two-year average), detail should be provided in the *TIME_DETAIL* attribute.

Reporting type (REPORTING_TYPE); code list: CL_REPORTING_TYPE

Refers to the type of data provider; used to distinguish between National, Regional, Global Reporting. The value should be set based on the data provider regardless of the origin of the figure. For example, if custodian agency reports a figure originally reported by the country, it should still set the value to Global, since the figure is part of the global dataset. National government agencies should always set the value of this concept to National. See also attribute *NATURE*.

Series (SERIES); code list: CL_SERIES

SDG indicator or series. A single SDG indicator has one or more series associated with it.

Reference Area (REF_AREA); code list: CL_AREA

Country or other geographic area to which the measured statistical phenomenon relates. The code list combines reference area codes that follow the ISO-3166 and M49 classifications.

Sex (SEX); code list: CL_SEX

The state of being male or female.

Age (AGE); code list: CL_AGE

Age - or age range - of the individuals the observation refers to.

Degree of urbanisation (URBANISATION); code list : CL_URBANISATION

Differentiates between urban and rural areas.

Income or wealth quantile (INCOME_WEALTH_QUANTILE); code list: CL_QUANTILE

Refers to income or wealth quintile of the population. In the future, may be extended to cover decile, percentile, etc.

Education Level (EDUCATION_LEV); code list: CL_EDUCATION_LEV

Highest level of an educational programme the person has successfully completed. Based on the ISCED classification.

Occupation (OCCUPATION); code list: CL_OCCUPATION

Job or position held by an individual who performs a set of tasks and duties. Based on the ISCO classification.

Disability status (DISABILITY_STATUS); code list: CL_DISABILITY

Refers to a person's disability status: with or without disability.

Economic Activity (ACTIVITY); code list; CL_ACTIVITY

High-level grouping of economic activities based on the types of goods and services produced. Based on the ISIC classification.

Product type (PRODUCT); code list; CL_PRODUCT

Product or commodity type. Combines SDG-specific entries from several classifications including CPC, Material Flows, and non-standard codes.

Custom breakdown (CUST_BREAKDOWN); code list: CL_CUST_BREAKDOWN

A "wildcard" dimension used to facilitate non-standard breakdowns, which has a list of generic codes (C01, C02, etc). Used in conjunction with the *CUST_BREAKDOWN_LB* attribute, which transmits the label for the code used in the *CUST_BREAKDOWN* dimension. Should be used in situations where the classification is volatile and may vary from year to year, or other situations where a non-standard breakdown is required.

Composite breakdown (COMPOSITE_BREAKDOWN); code list: CL_COMP_BREAKDOWN

A dimension that is used to implement multiple seldom-used breakdowns. The code list represents several merged breakdowns. Of these, only one breakdown at a time can be used to disaggregate a series.

Time period (TIME_PERIOD)

Timespan or point in time to which the observation actually refers. The convention in SDGs is to only provide a four-digit calendar year in the *TIME_PERIOD* concept. Further information can be provided in the *TIME_DETAIL* attribute, and structured time period information in the *TIME_COVERAGE* attribute.

2.2. Attributes

Observation Status (OBS_STATUS); code list: CL_OBS_STATUS

Information on the quality of a value or an unusual or missing value. Mandatory observation-level attribute.

Unit Multiplier (UNIT_MULT); code list : CL_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. Use 3 for thousands, 6 for millions, 9 for billions, and so on. For simple units, use 0. Mandatory observation-level attribute.

Unit of Measure (UNIT_MEASURE); code list: CL_UNIT_MEASURE

Unit in which the data values are expressed. Mandatory time series-level attribute.

Nature of data points (NATURE); code list: CL_NATURE

Information on the production and dissemination of the data (e.g.: if the figure has been produced and disseminated by the country, estimated by international agencies, etc.). <u>Normally set to C (Country Data) in national reporting</u>. Optional observation-level attribute.

Observation-level footnotes (COMMENT_OBS)

Descriptive text which can be attached to the observation. 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. Mandatory observation-level attribute.

Time series-level footnotes (COMMENT_TS)

Descriptive text which can be attached to the time series. Additional information on specific aspects of the time series or each observation in the time series. Mandatory time series-level attribute.

Structured time coverage (TIME_COVERAGE)

ISO8601 representation of the actual time interval to which the observation refers. Optional observation-level attribute.

Upper bound (UPPER_BOUND)

Where the observation value represents a point estimate, can be used to convey the upper bound value. Optional observation-level attribute.

Lower bound (LOWER_BOUND)

Where the observation value represents a point estimate, can be used to convey the lower bound value. Optional observation-level attribute.

Base Period (BASE_PER)

Period of time used as the base of an index number, or to which a constant series refers. Should be set to a calendar year. Typically used for constant prices, as in "2005 USD dollar" – then the unit of measure should be set to "Constant USD", and the base period to "2005". Optional observation-level 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

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 attribute.

Location of a geoinformation file (GEO_INFO_URL)

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

Type of geoinformation file (GEO_INFO_TYPE)

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

Date of the most recent change of the series' data (DATA_LAST_UPDATE)

The date that the series' data was last updated in the disseminating database. Optional time series-level attribute.

3. Coding conventions

3.1. Use of the _T code

In all code lists of the SDG DSD, the _T code is used to indicate the absence of a breakdown. This covers situations where the value represents a Total (e.g. *Total Sex*, *Total Age*), or where the breakdown is not statistically applicable (e.g. *Sex* for indicator *Land Area Covered By Forest*). In cases where a dimension is not statistically applicable for a particular indicator, _T is the only acceptable dimension value for this indicator.

3.2. Predefined codes for gender and age-specific indicators and similar situations

The code F (Female) must be used in the *Sex* dimension where a series is women-specific. For example, F must be used in series such as

- *Proportion of ever-partnered women and girls subjected to physical and sexual violence by a current or former intimate partner in the previous 12 months*
- *Proportion of elected seats held by women in deliberative bodies of local government*
- *Number of seats held by women in national parliaments*, and others.

Care must be taken mapping these series to the SDG DSD, since these series are not disaggregated by sex, and Sex is often not provided for them in dissemination databases. It may be necessary to override the mapping for such series as the default mapping may be Sex = Total.

At the same time, the code F should only be used in series where the statistical subject is persons of Female sex. Where this is not the case, the value _T should be used. For example, the code _T should be used in series such as

- *Legal frameworks that promote, enforce and monitor gender equality (percentage of achievement, 0 - 100) -- Area 1: overarching legal frameworks and public life.*
- *Proportion of countries where the legal framework (including customary law) guarantees women's equal rights to land ownership and/or control*
- *Proportion of births attended by skilled health personnel,* and others of similar nature.

Similarly, an appropriate age group must be used in age-specific indicators, e.g.

- *Infant mortality rate*: age should be set to Y0 (under 1 years old)
- *Proportion of children aged 36-59 months who are developmentally on track in at least three of the following domains: literacy-numeracy, physical development, social-emotional development, and learning*: age should be set to M36T59 (36 to 59 months old)

For series *Proportion of women married or in a union of reproductive age (aged 15-49 years) who have their need for family planning satisfied with modern methods*, Sex should be set to F and Age should be set to Y15T49. Similarly, both Sex and Age should be set as appropriate in other series restricted by age and sex.

In the case of Unit of Measure, care should be taken to ensure that the unit is aligned with the series. For example, acceptable units of measure for series **_EN_H2O_WBAMBQ_** _Proportion of bodies of water with good ambient water quality_ are **RO** and **PT**.

See also Series-Concept Map below.

3.3. Series-Concept Map

The full series-concept mapping matrix is currently under development and will be available shortly.

4. Example Coding

Below are some examples of time series encoding as per the DSD. These are purely for illustrative purposes.

Global dataset; Employed population below international poverty line; age 15+; Ethiopia

FREQ	REPORTING_ TYPE	REF_ AREA	SERIES	AGE	SEX	URBANI SATION	INCOME_ WEALTH_ QUANTILE	EDUCATION_ LEV	OCCUPA TION	DISABILITY_ STATUS	ACTIVITY	PRODUCT	COMPOSITE_ BREAKDOWN	CUST_ BREAKDOWN
A	G	ET	SI_POV_EMP1	Y_GE15	_T	_T	_T	_T	_T	_T	_T	_T	_T	_T

Global dataset; Proportion of ever-partnered women and girls subjected to physical and sexual violence by a current or former intimate partner in the previous 12 months; age 18-49; the Netherlands

FREQ	REPORTING_ TYPE	REF_ AREA	SERIES	AGE	SEX	URBANI SATION	INCOME_ WEALTH_ QUANTILE	EDUCATION_ LEV	OCCUPA TION	DISABILITY_ STATUS	ACTIVITY	PRODUCT	COMPOSITE_ BREAKDOWN	CUST_ BREAKDOWN
A	G	528	VC_VAW_MARR	Y18T49	F	_T	_T	_T	_T	_T	_T	_T	_T	_T

Use of Composite Breakdown
Global dataset, Passenger volume (passenger kilometres), Mode of Transport: Air, Eastern Asia, exc. Japan and China (MDG)

FREQ	REPORTING_ TYPE	REF_ AREA	SERIES	AGE	SEX	URBANI SATION	INCOME_ WEALTH_ QUANTILE	EDUCATION_ LEV	OCCUPA TION	DISABILITY_ STATUS	ACTIVITY	PRODUCT	COMPOSITE_ BREAKDOWN	CUST_ BREAKDOWN
A	G	223	IS_RDP_PFVOL	_T	_T	_T	_T	_T	_T	_T	_T	_T	MOT_AIR	_T

5. DSD Matrix

5.1. Overview

The DSD Matrix is spreadsheet representation of the SDG Data Structure Definition and provides a summary of its concepts and code lists. The spreadsheet contains the following worksheets:

- **Data Model** – contains the Concept Scheme, concept roles, and references to code lists used by the concepts.
- **Code list sheets** – describe the contents of the DSD code lists.
- **Series-Concept Map** (pending) – will contain detailed information on admissible codes per SDG series.

5.2. Data Model

The Data Model worksheet contains a spreadsheet representation of the SDG Concept Scheme and the DSD. The worksheet contains additional information for each concept:

- Concept role in the SDG DSD
- Attribute type and attachment level, if applicable
- Concept value type
- Code list referenced, if any
- Code list maintenance agency.

5.3. Code list sheets

These are spreadsheet representation of the code lists. For each code list, its corresponding worksheet contains the full list of codes and their descriptions. Currently, the descriptions are generally available in the English language; where descriptions in other languages are available, these are also provided. The code list sheets also contain comments and additional information, such as indicator references for the series.

5.4. Series-Concept Map

This worksheet is currently under development. When available, it will display a detailed matrix of admissible codes, in each code list, per SDG series. The worksheet is a spreadsheet representation of SDG Content Constraints (keyset constraints), as well as constraints on attributes.

6. SDMX Artefacts

Official releases of the SDG DSD and related artefacts will be published at the SDMX Global Registry (http://registry.sdmx.org).

7. Governance

7.1. Ownership

The SDG Data Structure Definition and all associated artefacts and related materials are owned by the Interagency and Expert Group on SDG Indicators (IAEG-SDGs).

7.2. Maintenance

The DSD and associated artefacts are developed and maintained by the SDMX-SDGs Working Group, composed of representatives of 12 countries and 10 international agencies, with the United Nations Statistics Division acting as the Secretariat of the Working Group.

Maintenance Agency for SDG SDMX artefacts is specified as follows:

- Concept Scheme: IAEG-SDGs
- Code Lists: SDMX for cross-domain code lists, IAEG-SDGs for SDG-specific code lists
- DSD and associated artefacts: IAEG-SDGs

Physical maintenance of the SDG SDMX artefacts is carried out by UNSD on behalf of the SDMX-SDGs Working Group.

7.3. Maintenance Schedule

The DSD is updated and published up to 4 times per year, as necessary. These releases coincide with releases of the global SDG database.

Minor updates are carried out through sub-annual releases. These include those that do not break backward compatibility:

- Additions of codes to existing code lists
- Additions of conditional (non-mandatory) attributes.

Major updates are carried out through annual releases of the DSD, which coincide with the publication of the SDG Report. The major updates include those that render the DSD not backwards-compatible:

- Additions of dimensions
- Additions of mandatory attributes.