

SDMX GuIDElines

Guidelines for Representing Methodological Changes in Data Structure Definitions

Version 1.0

Contents

[Document History 2](#_Toc5199210)

[Introduction 3](#_Toc5199211)

[Problem Definition 3](#_Toc5199212)

[Implementation Options for Methodological Changes 5](#_Toc5199213)

[Example UNSD Case 4: 5](#_Toc5199214)

[Example UNSD Case 5: 5](#_Toc5199215)

[1. Represent known changes in the dimensionality 6](#_Toc5199216)

[2. Represent unknown changes in the dimensionality 7](#_Toc5199217)

[3. Working with a series break attribute 9](#_Toc5199218)

# Document History

|  |  |  |
| --- | --- | --- |
| **Version** | **Date** | **Comment** |
| 1.0 | 3/4/2019 | Initial version |

# Introduction

When designing SDMX artefacts for an implementation project, one major design choice is the dimensionality of the Data Structure Definition(s), that is, which and how many dimensions are used to uniquely identify the relevant time series. Various trade-offs related to this design choice, such as between DSD complexity and parsimony[[1]](#footnote-1), are discussed in the Modelling Guidelines.[[2]](#footnote-2)

One aspect that is mentioned in the Modelling Guidelines, but not elaborated in detail, is the one of structural stability in case of methodological changes. In other words, how future-proof is the DSD? How many and what kind of changes to the DSD are required when certain aspects of the underlying data change and the DSD needs to represent both, pre-change and post-change data, in different time series?

If certain future changes are already expected when the DSD is originally designed, dimensions (or attributes) covering these changes will be included in the DSD. If the DSD was designed without expecting changes, additional dimensions or attributes will have to be introduced at a later stage.

These guidelines provide recommendations on how to represent methodological changes for several use cases.

# Problem Definition

Examples of methodological change are changes in base years or changes in compilation practices resulting from new statistical manuals or guidelines. The most straightforward, but not always most practical, way of representing such changes is introducing one (or multiple) additional dimensions that distinguish the pre- and post-change data, e.g. the different base years or different compilation practices. However, there may also be corner cases, where some data series exist only for a very small subset of the observations (e.g. in international economic statistics data that are relevant for only one industry in only one country). Creating additional dimensions only for these corner cases is usually not justifiable in a global coding framework.

If a DSD framework exists already (e.g. a global DSD) and the methodological changes are not covered by the DSD, users may still want to exchange time series that are only differentiated by information that is not represented in the series key. Examples include changes in methodology (i.e. reference metadata), for which the key itself does not change, but both series need to be exchanged. In this case, the design of the DSD may need to be changed to enable the exchange of the relevant information.

The impact of the required design changes is a combination of the following points and depends on the structure of the existing DSD:

* Further details may need to be recorded in reference metadata.
* Attributes related to the data at different levels may need to be modified.
* The affected series may need to be recoded using a different code for a given dimension (i.e. a different series key).
* A different DSD may be required.

For example, in National Accounts the UNSD records a new time series in five cases:

1. Change of statistical framework
2. Change in currency unit
3. Change in fiscal year
4. Change in base year
5. Significant change in compilation and methodological practices [[3]](#footnote-3)

If the UNSD applied the above five cases to coding data using the global DSDs for National Accounts[[4]](#footnote-4), this would result in the following:

1. Change of statistical framework:

This includes major changes in classifications used. For the example of a switch from the 1993 System of National Accounts (SNA93) to the 2008 System of National Accounts (SNA2008), a switch of DSD from NA\_MAIN93 to NA\_MAIN solves the problem. SNA93 data are coded using the DSD NA\_MAIN93; SNA2008 data are coded using the DSD NA\_MAIN.

1. Change in currency unit:

Currency is already a dimension in the given DSD. The new time series key is created by using the code of the new currency. The two series have different series keys as they use two different currency codes.

1. Change in fiscal year:

This includes changing from fiscal to calendar year reporting or different fiscal year definitions. The given DSD already contains the attributes REFYEARSTART and REFYEAREND to record the definition of the fiscal year. The values of these attributes need to be changed to represent the new fiscal year definition.

1. Change in base year:

The given DSD already contains the attribute REF\_YEAR\_PRICE. Analogously to 3. “Change in fiscal year”, the value of this attribute needs to be changed to the new base year.

1. Significant change in compilation and methodological practices:

This includes changes that affect the comparability between the new data observations and the previously reported data observations without changing the statistical framework. This can happen for instance when you get new sources, when you use different compilation methods, etc. These changes cannot be coded within the existing National Accounts DSD. The change will be recorded as reference metadata (usually structured using an MSD and expressed as a metadata message).

Cases 3, 4 and 5 have the issue that they would result in observations with duplicate series keys in the data message with the existing DSD. Cases 3 and 4 only differ through series level attributes. Case 5 only differs in reference metadata. The old and new series would therefore have the same series key. Information linked to the time series (e.g. series level attributes) might lead to ambiguous values. Options to resolve these issues are described below.

# Implementation Options for Methodological Changes

The suggested implementation options described below cover these possible scenarios:

1. Include known potential changes in the dimensionality during DSD design
2. Include unknown potential changes in the dimensionality during DSD design
3. Include a series break attribute (not part of the DSD dimensionality/series key)

It is recommended to follow the options in the order above, depending on design flexibility or the mandate for design changes. Any changes that are known at design time should be represented accordingly (option 1). In case changes that are unknown at design time should be supported, option 2 can also be implemented. Option 3 may be used if the dimensionality of existing DSDs cannot be changed, for example to avoid a structural change.

Two examples (following cases 4 and 5 above) are used to illustrate the implementation options.

## Example UNSD Case 4:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Country X** |  |  |  |  |  | **Change**  |  |  |
| Time period | **1995** | **1996** | **1997** | **1998** | **1999** | **2000** | **2001** | **2002** |
| Values before the change | 10 | 20 | 30 | 40 | 50 | 60 | - | - |
| Reference year before | 1995 | 1995 | 1995 | 1995 | 1995 | 1995 | - | - |
| Values after the change | - | - | - | - | - | 70 | 80 | 90 |
| Reference year after | - | - | - | - | - | 2000 | 2000 | 2000 |

In the year 2000, country X changed the reference year, causing a break in the reporting series. At the time of the change, the country observes both, the new and the old value of the time series.

## Example UNSD Case 5:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Country X** |  |  |  |  |  | **Change**  |  |  |
| Time period | **1995** | **1996** | **1997** | **1998** | **1999** | **2000** | **2001** | **2002** |
| Values compiled from old sources | 10 | 20 | 30 | 40 | 50 | 60 | - | - |
| Values compiled from new sources | - | - | - | - | - | 70 | 80 | 90 |

In the year 2000, country X changed their data sources, expecting better quality data. Since data cannot anymore be compared directly, this also causes a break in the reporting series. As before, at the time of the change, the country observes both the new and the old value of the time series.

## 1. Represent known changes in the dimensionality

This implementation option assumes that the changes that the DSD needs to represent are known. It proposes the specification of (possibly multiple) conceptually clean and clear dimensions in the DSD that can represent possible changes in future data exchanges, for example for reference year or data source.

The advantages of this option are that it is conceptually clean and specific and therefore better supports automated processing of the represented information. However, these benefits come at the cost of a potentially higher number of dimensions and less flexibility in case of future changes that were not known at design time. The higher number of dimensions may not be justified in corner cases such as the one described above. The structural specificity (and, thus, rigidity) of this approach requires a structural change and with that a new major version of the DSD every time an unforeseen change occurs.

This approach is recommended for new DSD design projects and in cases when DSDs are redesigned and structural changes acceptable, also in the future. The key requirement is that the nature of changes to be covered is largely known.

**Implementation Details and Examples**

For [example UNSD case 4](#_Example_UNSD_Case), this option would mean to derive a DSD with three dimensions: reference area, time period and reference year. Each data element can be uniquely identified through those three dimensions.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Reporting of initial time series before the change (up to year 2000)** |  |  |  |  |
| Time period | **1995** | **1996** | **1997** | **1998** | **1999** | **2000** |  |  |
| REF\_YEAR\_PRICE*dimension* | **1995** | **1995** | **1995** | **1995** | **1995** | **1995** |  |  |
| OBS\_VALUE | 10 | 20 | 30 | 40 | 50 | 60 |  |  |
|  |  |  |  |  |  |  |  |  |
| **Reporting of the new time series after the change (year 2000 and after)** |
| Time period |  |  |  |  |  | **2000** | **2001** | **2002** |
| REF\_YEAR\_PRICE*dimension* |  |  |  |  |  | **2000** | **2000** | **2000** |
| OBS\_VALUE |  |  |  |  |  | 70 | 80 | 90 |

For [example UNSD case 5](#_Example_UNSD_Case_1), an additional dimension for "source" would need to be added. This may not be practical, especially if "change of sources" is only relevant for a few observations and therefore represents a corner case in this domain. Additionally, redundancy may be created, because data sources are often recorded in reference metadata. This redundancy can be resolved by removing the redundant metadata attribute from the DSD or MSD.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Reporting of initial time series before the change (up to year 2000)** |  |  |  |  |
| Time period | **1995** | **1996** | **1997** | **1998** | **1999** | **2000** |  |  |
| Data Source *dimension* | **Source A** | **Source A** | **Source A** | **Source A** | **Source A** | **Source A** |  |  |
| OBS\_VALUE | 10 | 20 | 30 | 40 | 50 | 60 |  |  |
|  |  |  |  |  |  |  |  |  |
| **Reporting of the new time series after the change (year 2000 and after)** |
| Time period |  |  |  |  |  | **2000** | **2001** | **2002** |
| Data Source *dimension* |  |  |  |  |  | **Source B** | **Source B** | **Source B** |
| OBS\_VALUE |  |  |  |  |  | 70 | 80 | 90 |

Adding a dimension for each possible aspect that could change over time (and create multiple series with duplicate keys) may turn out to be very difficult to maintain at the DSD level. In this case, option 2 below can be implemented.

## 2. Represent unknown changes in the dimensionality

The [first implementation](#_1._Represent_known) option is only feasible in cases when types of changes are known and can be related to concrete concepts to be represented as dimensions in a DSD. In some cases, methodological changes such as compilation methods do not relate to a clear and concise statistical concept. In other cases, changes may occur that were unforeseeable or unexpected, or the fact that changes would happen was expected but the nature of changes was unknown.

An approach to deal with such changes is to add a completely flexible, loosely defined dimension that can incorporate any unforeseen change in the structure of the data exchange. This makes the DSD future-proof, as it will not require a structural change and major new version when previously unknown changes occur. While this flexibility is a clear advantage of this approach, the downside is the introduced ambiguity, especially when the “integer data type” option (see below) is used. Therefore, the “code list” option is recommended if unknown changes need to be represented in a DSD.

Overall, this approach is recommended if the key requirement of option 1 described above is not met, i.e. the nature of changes to be covered is largely unknown, or it is mostly rare cases that need to be covered.

**Implementation Details and Examples**

During the time when the DSD is constructed, two new concepts are included:

* **“Time series differentiator”** dimension (integer or code list, TS\_DIFF): TS\_DIFF does not represent a statistical concept. It is added as an artificial concept to differentiate time series. The context is described either by the **time series differentiator comment** attribute or reference metadata.
* **“Time series differentiator comment”** attribute (string, COMMENT\_TS\_DIFF, optional): a textual attribute that is attached at the time series level and describes information related to the type of the structural change.

The **Time series differentiator** dimension may be represented by an integer data type or a code list:

* **integer data type**: the same number for this dimension is likely to mean different things. For example, TS\_DIFF:2 could mean “Change of reference year from 1995 to 2000 in year 2000” or “Change from fiscal year reporting to calendar year reporting in year 2010” or anything else. The **time series differentiator comment** attribute or reference metadata would be required to provide the context.
* **code list**: the context is maintained by the maintenance agency through this code list (e.g. code 1 for "Change of reference year from 1995 to 2000 in year 2000", code 2 for "Change from fiscal year reporting to calendar year reporting in year 2010"). It would mean a minor change of the DSD for adding an event with the advantage of being able to compare events over different series. Likewise, the part *in year xxxx* may be modelled as a separate concept to keep the code list shorter and stay more flexible. The **time series differentiator comment** attribute is not needed, if the “code list” option is used.

This implementation approach generates three time series:

* Before the change occurs (TS\_DIFF=1)
* After the change occurs (TS\_DIFF=2)
* Concatenated time series of 1 and 2 (TS\_DIFF=0): As a coding convention, 0 is reserved for concatenated series. Also for cases when no change occurred yet, the time series are coded with TS\_DIFF 0.

For [example UNSD case 4](#_Example_UNSD_Case), reporting will look as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Reporting of initial time series before the change (up to year 2000)** |  |  |  |  |
| TS\_DIFF | 1 |
| COMMENT\_TS\_DIFF *time series level* | **-** |
| REF\_YEAR\_PRICE*time series level* | 1995 |  |  |
| Time period | **1995** | **1996** | **1997** | **1998** | **1999** | **2000** |  |  |
| OBS\_VALUE | 10 | 20 | 30 | 40 | 50 | 60 |  |  |
| OBS\_STATUS | A | A | A | A | A | A |  |  |
| PRE\_BREAK\_VALUE  | - | - | - | - | - | - |  |  |
|  |  |  |  |  |  |  |  |  |
| **Reporting of the new time series after the change (year 2000 and after)** |
| TS\_DIFF | 2 |
| COMMENT\_TS\_DIFF *time series level* | Change of reference year from 1995 to 2000 in year 2000 |
| REF\_YEAR\_PRICE *time series level* |  |  |  |  |  | 2000 |
| Time period |  |  |  |  |  | **2000** | **2001** | **2002** |
| OBS\_VALUE |  |  |  |  |  | 70 | 80 | 90 |
| OBS\_STATUS |  |  |  |  |  | A | A | A |
| PRE\_BREAK\_VALUE  |  |  |  |  |  | - | - | - |
|  |  |  |  |  |  |  |  |  |
| **Reporting of the concatenated series covering the period before and after the change**  |
| TS\_DIFF | 0 |
| COMMENT\_TS\_DIFF *time series level* | Change of reference year from 1995 to 2000 in year 2000 |
| REF\_YEAR\_PRICE [[5]](#footnote-5)*time series level* | 2000 | 2000 |
| Time period | **1995** | **1996** | **1997** | **1998** | **1999** | **2000** | **2001** | **2002** |
| OBS\_VALUE | 10 | 20 | 30 | 40 | 50 | 70 | 80 | 90 |
| OBS\_STATUS | A | A | A | A | A | B | A | A |
| PRE\_BREAK\_VALUE  | - | - | - | - | - | 60 | - | - |

For [example UNSD case 5](#_Example_UNSD_Case_1), the reporting will look very similar. Series 1 and 2 will look basically the same as above. Series 0 will only have a different comment:

|  |
| --- |
| **Reporting of the concatenated series covering the period before and after the change**  |
| TS\_DIFF | 0 |
| COMMENT\_TS\_DIFF *time series level* | Change of data sources in year 2000 |
| Time period | **1995** | **1996** | **1997** | **1998** | **1999** | **2000** | **2001** | **2002** |
| OBS\_VALUE | 10 | 20 | 30 | 40 | 50 | 70 | 80 | 90 |
| OBS\_STATUS | A | A | A | A | A | B | A | A |
| PRE\_BREAK\_VALUE  | - | - | - | - | - | 60 | - | - |

In this case, the change is reflected as a structured attribute and no information is lost. The detailed description of the old and new sources can be done through a reference metadata message, targeting the relevant partial key (e.g. reference area and TS\_DIFF dimensions).

The advantage of this approach is that the structure of the DSD stays the same, even in cases of unforeseen changes. There is no need to implement a new major version of the DSD every time a different type of change occurs. It also allows keeping series level attributes available in a structured way. Reference metadata can target TS\_DIFF to provide additional context that cannot be described through the DSD.

For time series attributes (such as REF\_YEAR\_PRICE in that example case 4), some information cannot be retained in the series 0. When generating the TS\_DIFF 0 series from the 1 and 2 series, the comment as well as the break flag can be generated automatically. However, for detailed data analysis, it would be recommended to use the original series (TS\_DIFF <> 0) in order to maintain all attributes in a structured way. The concatenated series (TS\_DIFF = 0) may be used for visualisation or for other use cases not relying on having all attributes available in a fully structured way. This is also relevant in cases where multiple changes are chained (e.g. change of reference year in year 2000 and change of sources in year 2010).

## 3. Working with a series break attribute

In cases when it is not possible to create additional dimensions (either following option 1 or 2), changes can also be represented using a "break in series" attribute along with textual explanations. It is possible to use two observation-level attributes to transmit an observation value in effect prior to the series break and to indicate a break in a series. In this respect and in the context of the above example the observation-level PRE\_BREAK\_VALUE and OBS\_STATUS attributes must exist in the DSD.

The reporting is done in a single series, basically reflecting the concatenated series of option 2 (TS\_DIFF=0), but without the differentiator dimension. Other dimensions are the same as before. The OBS\_STATUS attribute is set to "B" (break) and the PRE\_BREAK\_VALUE attribute indicates the value for the specific observation before the change. In addition, a comment can be added to a textual attribute (e.g. BREAK\_REASON) to give additional information about the change.

For the examples used above, the data messages would look as follows:

**Example Case 4:**

|  |  |
| --- | --- |
| BREAK\_REASON*time series level* | Change of reference year from 1995 to 2000 in year 2000 |
| REF\_YEAR\_PRICE [[6]](#footnote-6)*time series level* | 2000 | 2000 |
| Time period | **1995** | **1996** | **1997** | **1998** | **1999** | **2000** | **2001** | **2002** |
| OBS\_VALUE | 1 | 2 | 3 | 4 | 5 | 1006 | 1007 | 1008 |
| OBS\_STATUS | A | A | A | A | A | B | A | A |
| PRE\_BREAK\_VALUE  | - | - | - | - | - | 6 | - | - |

**Example Case 5:**

|  |  |
| --- | --- |
| BREAK\_REASON*time series level* | Change of data sources in year 2000 |
| Time period | **1995** | **1996** | **1997** | **1998** | **1999** | **2000** | **2001** | **2002** |
| OBS\_VALUE | 1 | 2 | 3 | 4 | 5 | 1006 | 1007 | 1008 |
| OBS\_STATUS | A | A | A | A | A | B | A | A |
| PRE\_BREAK\_VALUE  | - | - | - | - | - | 6 | - | - |

The advantage of this approach is that the structure of the DSD in terms of dimensions stays the same. There is no need to create a new major version of the DSD. In case some of the attributes used above are not yet defined in the DSD, adding them is a minor change.[[7]](#footnote-7) The same series key can continue to be exchanged and backward compatibility is ensured.

Since the "original series" (see option 2) are not available in that approach, series level attributes are not anymore fully maintained in a structured way. Note in the example, that the fact of the reference year being 1995 before time period 2000 is only available in the free text attribute (COMMENT\_TS) and the REF\_YEAR\_PRICE attribute is only correct for the year 2000 onwards.

In case multiple such changes are chained, then multiple such comments are required. Additionally, if data is compared between countries for partial time series and processed automatically, this may lead to errors in data processing. E.g. if data is shown for country X and Y for the year 1999 only:

|  |  |  |
| --- | --- | --- |
| Time period = **1999** | **Country X** | **Country Y** |
| OBS\_VALUE | 5 | 10 |
| OBS\_STATUS | A | A |
| REF\_YEAR\_PRICE | 2000 | 2000 |
| PRE\_BREAK\_VALUE  |  |  |
| COMMENT\_TS *time series level* | Change of reference year from 1995 to 2000 in year 2000 |  |

Without parsing the free-text comment, it will not be clear that country X used base year 1995 in year 2000. Since the dataset is partial, also the break flag is not visible. This may become a major issue in case data is linked and automatically processed between different datasets for different purposes.

1. A parsimonious DSD does not contain any redundant dimensions that are not needed to uniquely identify a data point. [↑](#footnote-ref-1)
2. See Modelling Guidelines <https://sdmx.org/wp-content/uploads/Modelling-statistical-domains-in-SDMX-v2-201806.docx> [↑](#footnote-ref-2)
3. This is often triggered by a change of the statistical framework. [↑](#footnote-ref-3)
4. National Accounts page on SDMX.org: <https://sdmx.org/?page_id=1498> [↑](#footnote-ref-4)
5. This is a time series level attribute and therefore has the value 2000 only once. It is displayed here twice to illustrate that the wrong value would be assigned to the pre-change part of the time series. [↑](#footnote-ref-5)
6. This is a time series level attribute and therefore has the value 2000 only once. It is displayed here twice to illustrated that the wrong value would be assigned to the pre-change part of the time series. [↑](#footnote-ref-6)
7. SDMX Versioning guidelines: <https://sdmx.org/wp-content/uploads/Guidelines_on_versioning_v1_0.pdf> [↑](#footnote-ref-7)