**SDMX STATISTICAL GUIDELINES**

**SDMX Cross-Domain Code Lists**

## CL\_CONF\_STATUS

**Name**: Code list for Confidentiality Status (CONF\_STATUS).

**Description**: This code list provides coded information about the sensitivity and confidentiality status of the data.

**Established international standard used as input for the code list**: None.

**Version**: 1.2

# DOCUMENT HISTORY

|  |  |  |
| --- | --- | --- |
| **Version** | **Date** | **Comment** |
| 1.0 | 2009 | Initial version. |
| 1.1. | 26/6/2014 | Clarification changes to descriptions; addition of primary confidentiality codes; additional remarks for specific issues. |
| 1.2 | 19/12/2017 | Added this VERSION HISTORY section; removed the previous changes section marked as tracked changesAdded CONF\_STATUS:E; included CONF\_REDIST descriptions for codes N and EAdded REMARKS section |

**Date**: 19 December 2017.

|  |  |  |
| --- | --- | --- |
| **Recommended code value** | **Recommended code description** | **Annotation** |
| F | Free (free for publication) | Used for observations without any special sensitivity considerations and which can thus be freely shared. Usually, source organisations provide information and guidance on general requirements for re-dissemination  (like mentioning the source) either on their websites or in their paper publications. In some institutional environments the term "unclassified" is used in a sense that still denotes implied restrictions in the circulation of information. If this is the case, the organisations concerned may probably consider that "free" (value F) is not the appropriate tag for this kind of "unclassified" category and that "Not for publication, restricted for internal use only" (value N) may be more appropriate.  |
| N | Not for publication, restricted for internal use only | Used to denote observations that are restricted for internal use only within organisations. This code may be accompanied with an additional observation-level attribute: CONF\_REDIST which defines the secondary recipient(s) to whom the sender allows the primary recipient to forward confidential data. |
| C | Confidential statistical information | Confidential statistical information (primary confidentiality) due to identifiable respondents. Measures also should be taken to prevent not only direct access, but also indirect deduction or calculation by other users and parties, probably by considering and treating additional observations as "confidential" (secondary confidentiality management). |
| D | Secondary confidentiality set by the sender, not for publication  | Used by the sender of the data to flag (beyond the confidential statistical information) additional observations in the dataset so that the receiver knows that he/she should suppress these observations in subsequent stages of processing (especially dissemination) in order to prevent third parties to indirectly deduct the observations that are genuinely flagged with "C".  |
| S | Secondary confidentiality set and managed by the receiver, not for publication  | If senders do not manage the secondary confidentiality in their data and/or there are also other countries' data involved (with the intention to eventually compile a regional-wide aggregate that is going to be published), the value "S" is used by the receiver to flag additional suppressed observations (within sender’s data and/or within the datasets of other senders) in subsequent stages of processing (especially, dissemination) in order to prevent third parties to indirectly deduct the observations that were genuinely flagged with "C" by the sender. |
| A | Primary confidentiality due to small counts | A cell is flagged as confidential if less than m units ("too few units") contribute to the total of that cell. The limits of what constitutes "small counts" can vary across statistical domains, countries, etc.  |
| O | Primary confidentiality due to dominance by one unit | Used when one unit accounts for more than x % of the total of a cell. The value of x can vary across statistical domains or countries, be influenced by legislation, etc.  |
| T | Primary confidentiality due to dominance by two units | Used when two units account for more than x % of the total of a cell. The value of x can vary across statistical domains or countries, be influenced by legislation, etc. |
| G | Primary confidentiality due to dominance by one or two units | Used when one or two units account(s) for more than x % of the total of a cell. The value of x can vary across statistical domains or countries, be influenced by legislation, etc. |
| M | Primary confidentiality due to data declared confidential based on other measures of concentration | Cells declared confidential using mathematical definitions of sensitive cells, e.g. p-percent, p/q or (n,k) rules.  |
| E | Not for publication, restricted for internal use only (equivalent to the code N) until the embargo time elapses; Free for publication (equivalent to the code F) after the embargo time elapses. | Used for embargoed data. The embargo time has to be specified in the EMBARGO\_TIME attribute.This code may be accompanied with an additional observation-level attribute: CONF\_REDIST which defines the secondary recipient(s) to whom the sender allows the primary recipient to forward confidential data. |

# REMARKS

For guidance on how to implement CONF\_STATUS, see the SDMX content-oriented Guidelines for Confidentiality and Embargo in SDMX.

A clarification on the difference between codes N and C is that N is restricted for an unknown reason, and C is confidential for a known reason (due to identifiable respondents).

# Specific issues

The values A, O, T, G and M can only appear with an observation value, not with a missing value. If suppressed, the CL\_OBS\_STATUS code would be Q (and the CL\_OBS\_CONF code F).

This code list combines two distinct concepts: "Confidentiality status" (Confidential versus not confidential data) and "Reasons for confidentiality" (e.g. dominance rule).

The reasons for keeping the two concepts together are the following:

* the first version of the code list already combined confidentiality status and reasons for confidentiality, and it was felt inappropriate to change this;
* the mixed approach seems also to be more practical in collection and international data sharing environments as it lowers the complexity of the DSDs by having less concepts (in line with the simplicity principle mentioned in the "[Guidelines for the Design of Data Structure Definitions](http://sdmx.org/wp-content/uploads/2014/02/SDMX_Guidelines_for_DSDs_1.0.doc)");

The values of this code list are usually attached at the observation level; if the use of this concept is defined as "mandatory" in a DSD, a value from this code list should be assigned to all observations.

Code values "**D**" and "**S**" are useful when organisations manage secondary statistical confidentiality. Primary confidentiality (code "C") applies to data that directly or very easily reveal information about individuals (persons or companies) and which is not publicly available. Obviously, this kind of information needs to be strictly protected. Moreover, if a figure is composed, for example, of information provided by only two agents, then one of these parties or a third party could easily deduct the suppressed information of the individual person or firm. Also, accounting identities could possibly be used or, using longer time series, other algorithms and techniques (e.g. regression) to deduct a suppressed "C" observation out of a large dataset. Thus, there is possibly a need to suppress additional information which is not confidential statistical information in the first place or, in general, to take measures (secondary confidentiality management) against indirect deduction of the suppressed "C" information from third parties who might be using the rest of the information that is available.