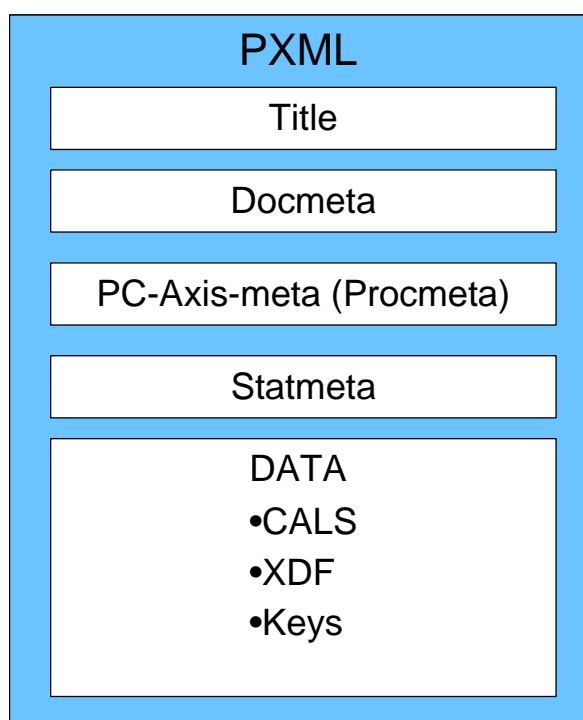


The CoSSI compatible PXML-format version 1.0.

This document describes how the metadata information in a PC-Axis file is matched with the metadata definitions of Statistics Finland's CoSSI-model. In the table below "Name" is the name of the PC-Axis keyword, "Description" is the description copied from the PC-Axis documentation and "DTD equivalent" is the equivalent element in the DTD (statmeta.dtd, docmeta.dtd, pxmeta.dtd, matrix.dtd, table.dtd and statkeys.dtd). These modular DTDs are described in the DTD documentation.

If a PC-Axis keyword is statistical or document metadata, it is included in the corresponding element in statmeta.dtd, docmeta.dtd, table.dtd, matrix.dtd and statkeys.dtd. Other keywords are placed in the pxmeta.dtd.



A short introduction to the CoSSI model

The CoSSI defines the structures of statistical data (matrices and tables), metadata (document and statistical metadata, and quality declarations), and publications. XML DTDs have been selected as the technical means for implementing these structures. The CoSSI model is comprised of several DTDs that can be modularly combined for different types of documents. The basic document types are a statistical table (CALS), a statistical matrix (XDF) and a publication. These documents are XML documents that are compatible with the CoSSI model and also contain the metadata and the language versions necessary for describing a set of statistics.

PC-Axis is also using CoSSI model as the XML format for the data and metadata.

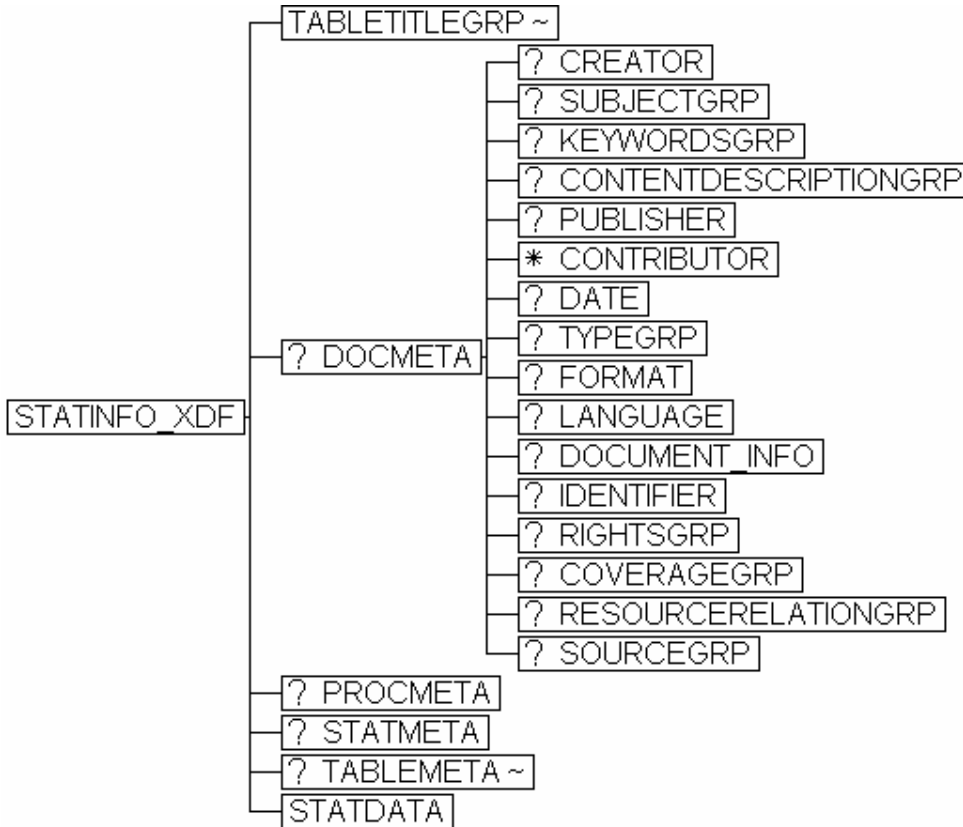
Metadata

In the CoSSI model, metadata are divided into four categories:

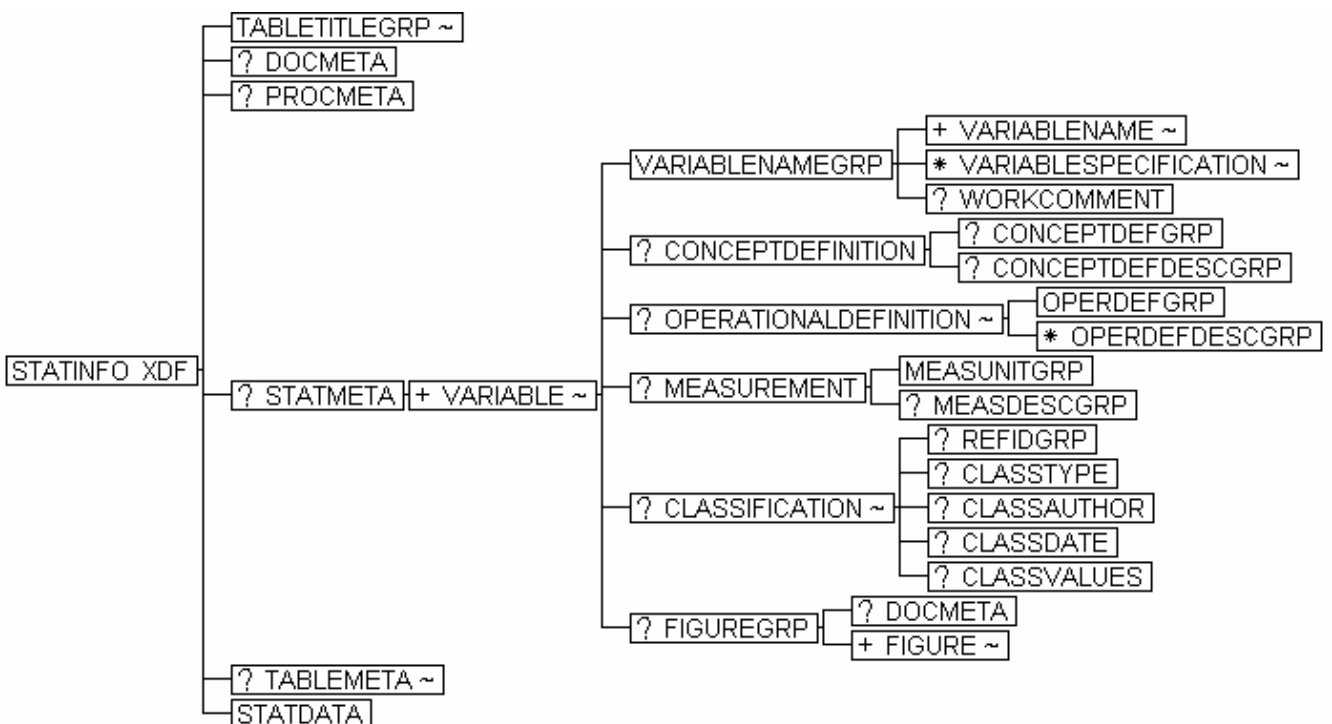
1. Document metadata
2. Statistical metadata
3. Quality declarations
4. Processing metadata

The division is based on the content and character of metadata. Document metadata describe the content of a document, its creator, date, keywords, statistical topic and identifiers connected with the document. Statistical metadata contain descriptions of the variables that are present in statistical data and tables, calculation rules and any classifications that may apply to a variable. The quality declaration is a standard format description of the data collection, the data and the applied statistical methods. Processing metadata are metadata for statistical software applications. For the PC-Axis there is a specific processing metadata module called pxmeta.dtd. The keywords in a PC-Axis file are converted to these metadata according to the information they carry.

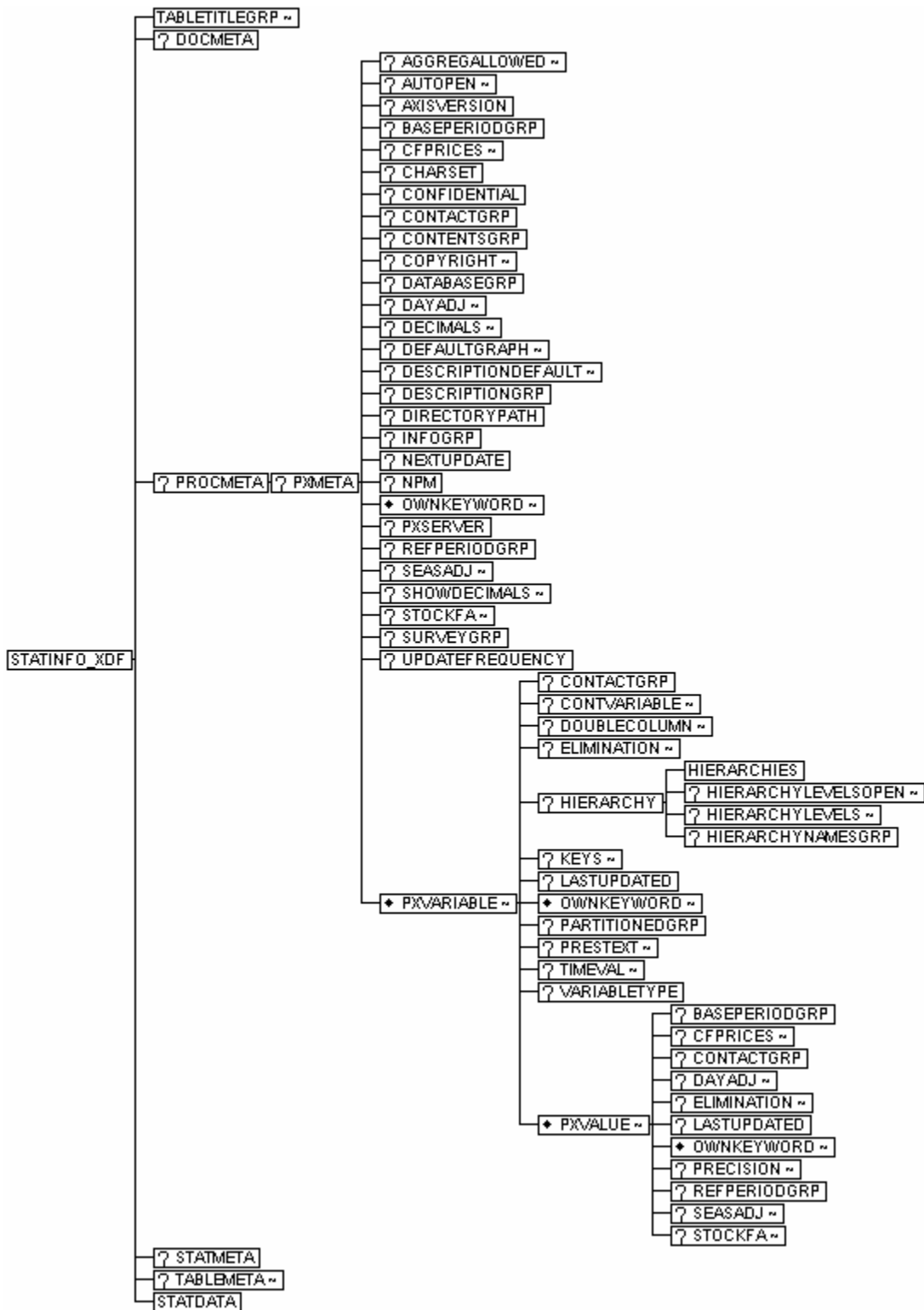
Document metadata



Statistical metadata



Processing metadata

*Statistical data in CoSSI*

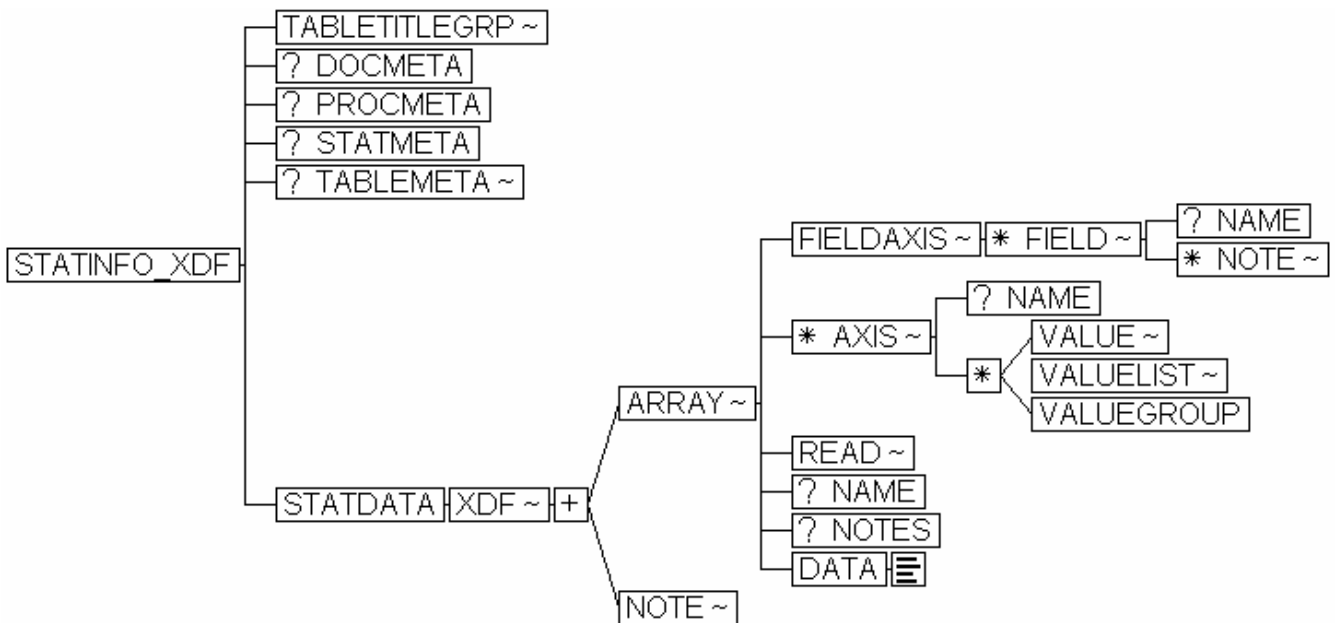
The CoSSI model contains three ways for describing statistical data – matrices, tables and sparse matrices. The difference between the two matrices formats and tables is a technical one, for they all contain the same statistical information. The data in a matrix is organised in the XDF matrix format, whereas a table is based on the CALS model. The sparse matrix format is similar to the PC-Axis Keys format. However, they all contain the same metadata modules (document, statistical and processing metadata), so the matrices can be automatically converted into tables, and vice versa.

Matrices

The matrix format (XDF) is primarily intended for the saving of statistical data and with PC-Axis, this may be the most relevant format. In the matrix format, the variables included in the matrix, and their order, are described first, followed by the portion containing all the actual data separated by character spacing. This eliminates unnecessary repetition and even large volumes of statistical data can be packed into a small space. So this is very similar to the traditional PC-Axis file format.

In the CoSSI model, the metadata concerning a record or document, as well as statistical and processing metadata are attached around the XDF matrix. The contents of the matrix, its creator and the included variables can be exhaustively described with the help of these metadata.

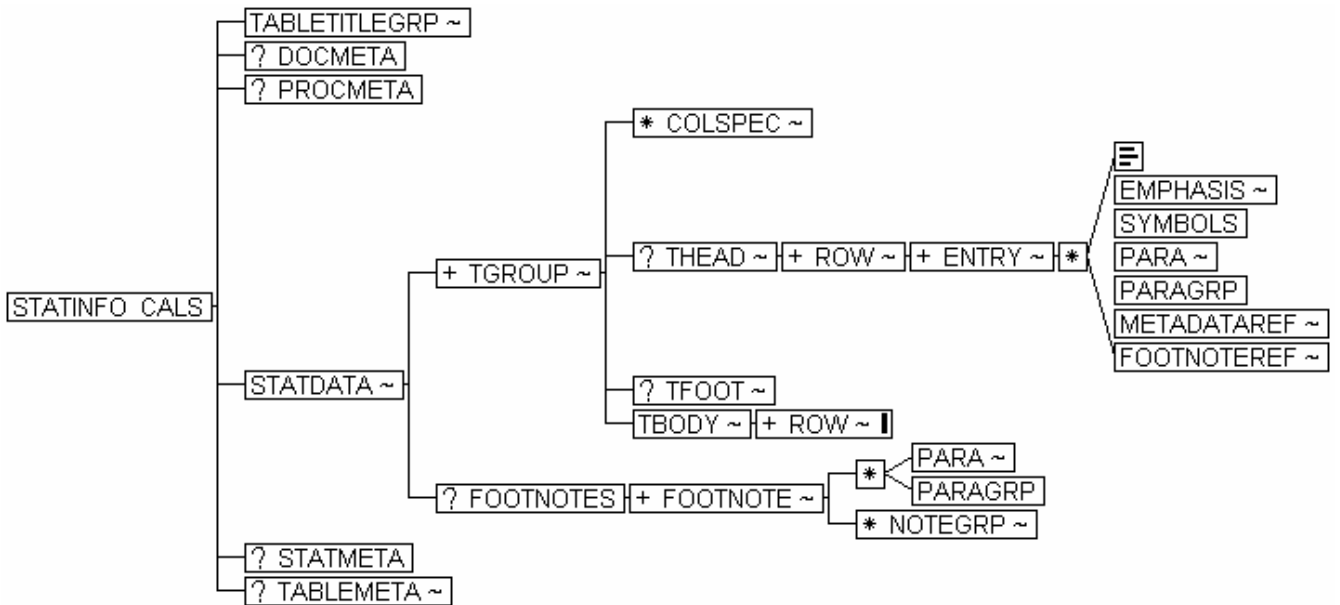
As the matrix model is so effective with large data volumes, the matrix format is excellent for the archiving of statistical data and as an original for tables that are to be disseminated from a database. For example, statistical data in the matrix format can be converted automatically into traditional PC-Axis format for dissemination via PX-Web.



Tables

The table format is based on the CALS model, enlarged for the presentation of statistical tables (Statistical CALS). It is well suited for the publishing of tables with a predefined structure and presentation format. The CoSSI model's definition for a publication structure specifies a table conforming with the table model, whereby a table can be attached direct to a publication. A statistical table is described very precisely with XML elements in the CALS model, so that its conversion into, e.g. an HTML table requires no special processing as with the matrix format, and the conversion can be done easily with standard XML techniques.

The same metadata modules as in the matrix format are added around the table model in the CoSSI model. The contents of the table and the variables can be described precisely with these metadata and the data can be used all the way down to dissemination.



Sparse matrix

This sparse matrix format is very similar to the XDF matrix format. The only difference is that zero values are excluded from the data part. So this format is useful for the data sets with lot of zeros.

Mapping from the PC-Axis keywords to the CoSSI XML elements

| Name | Description | DTD equivalent: DM = Document Metadata (docmeta.dtd) SM = Statistical Metadata (statmeta.dtd) PM = PC-Axis Metadata (pxmeta.dtd) |
|----------------------|---|--|
| AGGREGALLOWED | If the contents of the table cannot be added, contains for instance index or average, the keyword AGGREGALLOWED=NO; is used to stop the user from making a sum. If the keyword is missing aggregations are allowed. | PM: <aggregallowed valid="no" /> |
| AUTOPEN | If the file is published on the Internet and the user selects a number of variables and values it is possible to remove the windows "Select variables and values" in PC-Axis and instead show the complete table in PC-Axis when the file is downloaded. AUTOPEN=YES;. | PM: <autopen valid="yes" /> |
| AXIS-VERSION | Version number for PC-AXIS (max 80 chars). Is read and saved but otherwise not used in version 1.5 and later. | PM: <axisversion> |
| BASEPERIOD | Base period for for instance index series. Is shown with the footnote. If there is a contents variable the keyword is repeated for each value of the contents variable. | PM: <baseperiodgrp><baseperiod xml:lang="...">... |
| CELLNOTE | Footnote for a single cell or a group of cells. Which cell it refers to is given by values and variables. If a value is given as * the note refers to all values for that variable. Only one value can be given for each variable. The values are given in the variable order indicated by STUB and HEADING, starting | XDF and KEYS: <notes><location>i j k</location><notegrp forcednote="no"><note xml:lang="en">cellnote</note></notegrp> CALs: <footnote> |

| | | |
|----------------------|---|---|
| | with STUB. | |
| CELLNOTEX | As CELLNOTE but shown mandatory as for NOTEX.. | XDF and KEYS: <notes><location>i j k</location><notegrp forcedNote="yes"><note xml:lang="en">cellnote</note></notegrp> CALs: <footnote> |
| CFPRICES | Indicates if data is in current or fixed prices. C is used for Current and F for Fixed prices. Quotation marks must be used. CFPRICES="C" or CFPRICES("value")="C". | PM: <cfprices type="c" /> |
| CHARSET | CHARSET="ANSI"; indicates that the texts in the file are written in Windows format. If the keyword is missing or if it has CHARSET="OEM" it means that the texts in the file are in DOS format. They will be translated by PC-Axis to Windows. This keyword must appear in the beginning of the file before any texts that can include characters outside A-Z, 0-9. | PM: <charset>ANSI</charset> |
| CODEPAGE | To be used for conversion to get correct characters. Default iso-8859-1. Max 20 chars. | Encoding for the xml-files is: encoding="iso-8859-1" If Chinese encoding "big5" |
| CODES | The key word CODES is used if a variable exists both in code and plain text. The codes are written in the same way as VALUES. Not more than 256 characters. | SM: classvaluecode |
| CONFIDENTIAL | Possibility to do some manipulation with the data in the data part of the file. Is only suitable if the user cannot download the total file since the data can be read in any editor. Max 20 chars. | PM: <confidential> |
| CONTACT | States the person who can give information about the statistics. Is written in the form <i>name, organization, telephone, fax, e-mail</i> . Several persons can be stated in the same textstring and are then divided by the #-sign. Is shown with the footnote. If there is a contents variable the keyword is repeated for each value CONTACT("value")="xx". | PM: (no contents variable) <contactgrp>. PM: (content variable) <pxvariable><contactgrp><contact xml:lang="...">... |
| CONTENTS | Information about the contents, which makes up the first part of a table heading created when retrieving tables from PC-AXIS. The text must not exceed 256 characters. | PM: <contentsgrp><contents xml:lang="...">... |
| CONTVARIABLE | This is used to indicate that the table has two or more different contents. For instance the contents Import and the contents Export. The variable name must also be found either as STUB or HEADING. When a contvariable exists a number of keywords will be indexed: DAYADJ, SEASADJ, STOCKFA, UNITS, CONTACT, LASTUPDATED, REFFPERIOD, BASEPERIOD, CFPRICES. The keyword CONTVARIABLE must precede the first keyword that will be indexed. | PM: <contvariable valid="yes" /> |
| COPYRIGHT | Copyright is given as YES or NO. If COPYRIGHT=YES the copyright refers to the organization given in SOURCE. Is shown together with footnotes. | PM: <copyright valid="yes" /> |
| CREATION-DATE | Date when file was created. Written in format CCYYMMDD hh:mm, e.g. "19960612 | DM: <date><published> <day> |

| | | |
|---------------------------|--|---|
| | 14:20". Is shown together with footnotes. | <time> |
| DATA | The key word DATA must be placed at the end of the file, followed by all the data cells or if Keys are used the variable values and all datacells that differ from 0. How the cells are to be written is described later in this paper. | This information is in the table (soextblx.dtd), matrix (statxdf.dtd) or keys (statkeys.dtd) part of the XML document. Therefore it is not needed in the metadata part of the document. How this is equivalent with table and matrix information see the documentation of table.dtd and matrix.dtd. |
| DATABASE | The name of the database from where the statistics is retrieved. Is shown with the footnote. | PM: <database> |
| DAYADJ | DAYADJ=YES means that data is adjusted e.g. to take into account the number of working days. Default is DAYADJ=NO or DAYADJ("value")=NO. | PM: <dayadj valid="no" /> |
| DECIMALS | The number of decimals in the table cells. 0 - 15. (0-6 if SHOWDECIMALS not included). Indicates how many decimals will be saved in the PC-Axis file. Written without quotation marks. Compare SHOWDECIMALS. | PM: <decimals number="1" /> Value for the attribute <i>number</i> must be a number from 0 to 15. |
| DEFAULT-GRAPH | This keyword is read and saved in the PX-file but not shown in PC-axis. | PM: <defaultgraph graphtype="1"> Value for the attribute <i>graphtype</i> must be a number from 1 to 10. |
| DESCRIPTION | If a file contains DESCRIPTION, when fetching from a disk, this text is used to show the contents of the px file. C.f. TITLE. If the user wants to save a file in PC-AXIS and writes a text that describes the file, this text will be saved as DESCRIPTION. The text will be used to show the contents of tables in the folder. The text is not presented as a note nor in any other way when the table is presented on the screen unless the keyword Descriptiondefault is used. In this case the description is shown instead of the title. The Description can have lines of max 98 characters. | PM: <descriptiongrp><description xml:lang="..."> |
| DESCRIPTIONDEFAULT | For some languages it is difficult to build a table title dynamically. The keyword DESCRIPTIONDEFAULT=YES; means that the text after keyword Description will be used as title for the table. | PM: <descriptiondefault valid="yes"> |
| DIRECTORY-PATH | This keyword is read and saved in the PX-file but not shown in PC-axis. | PM: <directorypath> |
| DOMAIN | Can occur once for each variable. Is used to determine which value sets are of interest, and thus which aggregation lists can be used. The text must not exceed 80 characters. | SM: <variable><classification><refidgrp><refid> |
| DOUBLECOLUMN | This keyword is used to get code and text in separate columns for the specified variable DOUBLECOLUMN("region")=YES;. It only has effect if the user selects presentation in matrix format. | PM: <doublecolumn valid="yes" /> |
| ELIMINATION | States if and how a variable may be eliminated in a table. If the key word is written as ELIMINATION("variable name")="value name" this value will be used as an elimination value if the user does not select the variable to the table. If the key word is written ELIMINATION("variable name")=YES this means that the variable will be eliminated by the summing up of all the | PM: <pxvariable><pxvalue><elimination valid="yes"> or <pxvariable><elimination valid="yes"> |

| | | |
|--------------------------------|--|---|
| | values for that variable in the file. | |
| HIERARCHIES | HIERARCHIES("Country")="E25","E25": E15","E15":E12","E12":AT","E12":BE", "E12":FI","E12":FR","E12":DE", "E12":GR","E12":IR","E12":IT","E12": LU","E12":NL"; | PM: <pxvariable><hierarchy><hierarchies>"E25","E 25":E15","E15":E12","E12":AT","E12":B E","E12":FI","E12":FR","E12":DE", "E12":GR","E12":IR","E12":IT","E12":LU ","E12":NL"</hierarchies> |
| HIERARCHY-LEVELS | To indicate the number of levels existing, only if all branches are the same length. | PM: <hierarchylevels existinglevels="1" /> |
| HIERARCHY-LEVELSOPEN | To allow the database manager to decide at which default hierarchical level the list of values will be presented to the end-users when selecting values for the table. | PM: <hierarchylevelsopen openlevels="1" /> |
| HIERARCHY-NAMES | Possibility to name the different levels | PM: <hierarchynamesgrp> <hierarchynames xml:lang="en"> "NameOfLevel1", "NameOfLevel2", .. |
| INFO INFO[EN] | This keyword is read and saved in the PX-file but not shown in PC-axis. | PM: <infogrp><info xml:lang="en"> text</info></infogrp> |
| INFOFILE | Name of a file containing more information on the statistics. If the keyword exists in the file a new button is shown in the toolbar and the user can click for more information. Depending on the file type and to which program the type is associated the corresponding program is started and the information shown. (Word for .DOC, Notepad for .TXT etc). | DM: <resourcerelation> |
| KEYS | If the keyword is missing it is equivalent to KEYS=NO;. If it is used it must occur as many times as there are variables in the stub. It contains the name of the variable and whether the key is taken from VALUES or CODES. Example: KEYS("age")=VALUES; KEYS("region")=CODES;. When it is used all data rows start with the Value/code of the variables in the stub. See Description of Data. | PM: <keys valuesCodes="values" /> |
| LANGUAGE | The language used in the PC-Axis file (2 chars), sv for Swedish, en for English etc. Compare language codes for text files. | DM: <main_language> |
| LANGUAGES | If more than one language in the PC-Axis file then all languages are mentioned here (2 chars each), Example: LANGUAGES="en","sv"; | DM: "The main language" is in the <main_language> element and the other languages are in the <other_language> element. Example: <language> <main_language>en</main_language> <other_language>sv</other_language> </language> |
| LAST-UPDATED | Date and time for latest update format CCYYMMDD hh:mm. Example "19960528 11:35". Is also used in Aremos file format. The date is not updated in PC-AXIS when changes are made to the table. If there is a contents variable it is written as LAST-UPDATED("value")= 19990318 18:12 | DM: <date><modified> <day> <time> or PM: <pxmeta><pxvariable><pxvalue><lastupdated> |
| LINK LINK[EN] | This keyword is read and saved in the PX-file but not shown in PC-axis. | DM: <resourcerelationgrp><resourcerelation> <link target="http://www.stat.fi"> |
| MAP | Used for a geographic variable for which maps can be made (max 80 chars). Example:MAP("region")="Sweden_municipality"; | SM: <variable><figuregrp><figure>. |
| MATRIX | The name of the matrix. Is suggested as file | DM: <documentnumber> |

| | | |
|--------------------|--|---|
| | name when the file is saved. Max 8 characters for older PC-Axis versions, Max 20 for PC-Axis 2006. | |
| NEXT-UPDATE | This keyword is read and saved in the PX-file but not shown in PC-axis. | PM: <nextupdate> |
| NOTE | Contains a footnote which is showed in the statistical data bases if the user demands it. In PC-AXIS it is shown if the user presses F7. The footnote may either refer to the entire table or to a table variable. In the latter case the key word must be followed by the variable name in parentheses. | TM: (table note) <tabletitlegrp tableIdRef="meta"> -> <tablemeta tableId="meta"> <notegrp forcedNote="no"><note xml:lang="en"> Text </note></notegrp> SM: (variable note) <variablename>variable</variablename> <conceptdefdesc xml:lang="en" forcednote="no"> Text</conceptdefdesc> |
| NOTEX | Contains a note which is always shown in the statistical data bases. In PC-AXIS the note is shown automatically, before the table is presented on the screen. The note may either refer to the entire table or to a table variable. In the latter case the key word should be followed by the variable name in parentheses. | TM: (table note) XDF and KEYS: <tabletitlegrp tableIdRef="meta"> -> <tablemeta tableId="meta"> <notegrp forcedNote="yes"> <note xml:lang="en"> Text </note></notegrp></matrixmeta> SM: (variable note) XDF and KEYS: <axis axisIdRef="var1"> CALs <metadatarref linken="var1"> SM: <variablename>variable</variablename> <conceptdefdesc xml:lang="en" forcedNote="yes"> Text</conceptdefdesc> |
| OWN KEYWORD | Ignored in PC-axis. | PM: <ownkeyword name="name_of_the_keyword"> value_of_the_keyword</ownkeyword> |
| PARTITIONED | This is used to partition a variable into levels. for instance PARTITIONED("region")="municipality",1,4; PARTITIONED("region")="subarea",5; states that the first four positions for the regional values contain the municipality code, and that the subarea code starts in position 5. Thus the values for the variable region after the key word VALUES must be written in code, not plain text. Max 3 levels can be used, each gives start position and length except for the last level where length is implied as rest of the code. . | PM: (ex 3 levels) <partitioned position="1" length="2">County <partitioned position="3" length="2"> Municipality <partitioned position="5">Subarea |
| PRECISION | Can occur for single values. Determines that the value shall be presented with a number of decimals that differs from the keyword SHOWDECIMALS. Is to be written as PRECISION("variable name","value name")=n where n is a figure between 0 and 6. | PM: <precision number="1" /> Value for the attribute <i>number</i> must be a number from 0 to 6. |
| PRETEXT | States if texts or codes are shown for the keyword VALUES. Normally a file is created | PM: <pretext ="1" /> |

| | | |
|--------------------------|--|---|
| | so that texts are found after the keyword VALUES and codes after the keyword CODES. This is equivalent to PRETEXT("variable name")=1; which is the default. If a user changes presentation from texts to codes and saves the file the value codes will be saved after the keyword VALUES and value texts after the keyword CODES. In this case the keyword PRETEXT is written as PRETEXT("variable name")=0. This enables PC-Axis to know it is necessary to switch to codes if aggregation is selected. The user can also decide to show both codes and texts for a value and in this case the keyword is saved as PRETEXT("variable name")=2 or PRETEXT("variable name")=3. PRETEXT becomes 2 if it originally was 1 and 3 if it originally was 0. | |
| PX-SERVER | This keyword is read and saved in the PX-file but not shown in PC-axis. | PM: <pxserver> |
| REFPERIOD | Text with information on the exact period for the statistics. Is shown with the footnote. If there is a contents variable the keyword is repeated for each value of the contents variable. | PM: <refperiodgrp><refperiod xml:lang="...">... |
| SEASADJ | SEASADJ=YES means that data is seasonally adjusted. Default is SEASADJ=NO or SEASADJ("value")=NO. | PM: <seasadj valid="no" /> |
| SHOWDECIMALS | The number of decimals to be shown in the table, 0-6. Must be the same or smaller than the number stored as indicated by the keyword DECIMALS. If SHOWDECIMALS is not stated in the file the number stated by DECIMALS will be used. | PM: <showdecimals number="1" /> Value for the attribute <i>number</i> must be a number from 0 to 6. |
| SOURCE | States the organization which is responsible for the statistics. Is shown with the footnote. | DM: <sourcegrp><source> |
| STOCKFA | Indicates if data is stock, flow or average. The used characters S (stock), F (flow) and A (average) must be within quotation marks. STOCKFA="S" or STOCKFA("value")="S". | PM: <stockfa type="s" /> |
| STUB and HEADING | At least one of the keywords STUB or HEADING must be included. Usually both are included, as you choose one or several variables for the stub and the heading, respectively. The keywords are followed by a list with the chosen variables. The variables are within quotation marks and separated by commas. Each variable name must not exceed 80 characters. If the list with the variables has to be divided up into several lines, this should be done after a comma and not within the variable name. | The names of the variables name are in <statmeta>. The names of the variables in Heading are also in <statdata>, array, fieldAxis, field, name |
| SUBJECT-AREA | The name of the subject area in plain text. The text must not exceed 100 characters. | DM: <subject> |
| SUBJECT-CODE | Subject area code. It is used to categorize the contents. The text must not exceed 5 characters for older PC-axis versions. Can be max 20 for PC-Axis 2006.. | DM: <categories> |
| SURVEY SURVEY[EN] | This keyword is read and saved in the PX-file but not shown in PC-axis. | PM: <surveygrp><survey xml:lang="en"> |

| | | |
|-------------------------|---|--|
| TIMEVAL | Keyword to enable the use of time series. After the keyword is the name of the time variable given, e.g. TIMEVAL("time"). TLIST gives information on timescale and timeperiods. The time periods must be consecutive | PM: <timeval><frequency> |
| TITLE | The title of the table, reflecting its contents and variables. How the heading of a table will look depends on which variables the user chooses; the text created will then be saved as TITLE. If DESCRIPTIONDEFAULT exists the table is not built dynamically but the text from DESCRIPTION, is used. (Compare also Descriptiondefault) | <tabletitletitlegrp> <tabletitletitle xml:lang="..."> <tablemaintitle> |
| UNITS | Unit text, e.g. ton, index. The text must not exceed 80 characters. Compare UNITS for CONTVARIABLE. | <i>Unit for a table:</i> TABLEMETA: <tableunitsgrp><tableunits xml:lang="..."> <i>Unit for a value:</i> SM: <measunit> |
| UNITS | When there is a CONTVARIABLE the keyword UNITS takes an index and is repeated for every value for the contents variable. UNITS("value")="tons". | <i>Unit for a table:</i> Tablemeta: <tableunitsgrp><tableunits xml:lang="..."> <i>Unit for a value:</i> SM: <measurement> |
| UPDATE-FREQUENCY | This keyword is read and saved in the PX-file but not shown in PC-axis. | PM: <updatefrequency> |
| VALUENOTE | A footnote for separate variable values. Should be written with the variable name and the value names in parentheses. | XDF and KEYS: <axis axisIdRef="var1"> CALs <metadatarref linken="var1"> SM: <variablename>variable</variablename> <classvaluename>value</classvaluename> <classvaluedesc xml:lang="en" > <i>Text</i> </classvaluedesc> |
| VALUENOTEX | Mandatory footnote for single values for a variable. Is written with the variable name and the value name in parentheses. Is shown the same way as NOTEX. | XDF and KEYS: <axis axisIdRef="var1"> CALs <metadatarref linken="var1"> SM: <variablename>variable</variablename> <classvaluename>value</classvaluename> <classvaluedesc xml:lang="en" forcedNote="yes"> <i>Text</i> </classvaluedesc> |
| VALUES | The key word VALUES occurs once for each variable in the table, and is followed by the variable name in parentheses, within quotation marks. The values will be in the same order as in the stub and heading, respectively. They are within quotation marks and separated by commas. Each value name must not exceed 256 characters. If the values has to be divided up into several lines, this should be done after a comma and not within the value name. See also the key word TIMEVAL below. | The names of the values are in <statdata> array, fieldAxis, axis, value, name. |
| VARIABLE-TYPE | This keyword is read and saved in the PX-file but not shown in PC-axis. | PM: <variabletype>text</variabletype> |