Download whole datasets facility

(Usually called Bulk Download)

To access to the bulk downloads facility: Registered users

Commission

Useful information for the bulk downloads facility:

– 'Table of contents':   gives the list of tables and datasets available in the navigation tree and in the 'Data' directory. The table of contents (TOC) is available in two formats: xml and txt.

– 'Data' directory: includes all zipped datasets and tables.

– 'Dic' directory: includes the 'dictionaries' of all the coding systems used in the datasets, but also provides:

a 'Dimlst' file which includes a description of all dictionaries;

a 'Table' file' which gives the list of all tables and datasets.

– Excel hints: provides information on how to make better use of Excel with 'dft' and 'tsv' files.

Data update: the information stored here is updated twice a day, at 11:00 and 23:00.

1. TABLE OF CONTENTS

The 'table of contents' (TOC) file provides information on the datasets and tables available on the Eurostat website (see below). These datasets and tables are stored in 'tsv', 'dft' and 'sdmx' formats in the 'Data' directory.

Title

Titles in the TOC are available in English only.

Code

Each item (dataset, table and folder) of the TOC has a unique code which allows it to be identified both in the navigation tree and in the directory data of the bulk download. However, some of them can be classified at two different places in the navigation tree.

Examples:

- Classification of a **dataset** under two different themes:

Regional transport data are available in both 'Theme 1 / Regional statistics' and 'Theme 7 / Transport'.

- Some **tables** are classified both by Themes (1 to 9) and by EU policy indicators.

Type

Dataset: i.e. an 'open dataset' available in the 'Data' directory.

Table: i.e. a 'pre-defined table' available in the 'Data' directory.

Folder: i.e. a 'section' or 'chapter'.

Last update

Indicates the last time the dataset/table was modified.

Data start

Date of the oldest value included in the dataset (if available).

Data end

Date of the most recent value included in the dataset (if available).

Values

Number of actual values included in the dataset.

2. DATA DIRECTORY

This directory includes approximately 4000 datasets available in three formats: 'tsv', 'dft' and 'sdmx–ml'. Each of these formats is compressed in zip files ('gz'). To access and download the data, first make your choice between the three available formats. Then click on the letter corresponding to the code(s) you are looking for and make your choice.

You can sort the files by name, date of update or size by clicking on the header of the corresponding column.
All 'gz' files can be uncompressed with the help of WinZip.

2.1. The 'tsv' format (tab separated values)

Definition

'Tsv' files are flat files that include a 'tab delimited' sequence of values in each line instead of one value per line/record.

– In most files the sequences of values are time series.
– For datasets without the dimension (or attribute) time (e.g. 'area of the regions'), or that cover only one period of time, the sequences of values are not time series but another dimension, e.g. geographical series.
– The columns (or fields or cells) of the records are 'tab delimited'.
– Time series are sorted in descending order (for explanation see the chapter 'Hints for Excel users').

Example of dataset with time series (with made-up values)

unit,s_adj,partner,flow,indic,geo\time	2004 m05	2004 m04	2004 m03	2004 m02
mio-eur,nsa,ext_eurozone,net,bp-100,eurozone	11148	10660	13398	9437
mio-eur,nsa,ext_eurozone,net,bp-200,eurozone	3386 e	539	-185	-432
mio-eur,nsa,ext_eurozone,net,bp-300,eurozone	-5626 e	-6696 i	1902	919
mio-eur,nsa,ext_eurozone,net,bp-379,eurozone	-5758 e	-4165	-3970	-4703
mio-eur,nsa,ext_eurozone,net,bp-993,eurozone	3151.5 e	338.7 i	11146.1	5221.0
mio-eur,nsa,ext_eurozone,net,bp-994,eurozone	2314	669	543	2113
mio-eur,nsa,ext_eurozone,net,bp-010,eurozone	5465.1	1006.5	11689.0	7334.3

– First line: header.
– Other lines: records with the sequence of values.
– First column — first line: sequence of codes separated by a comma followed by a code separated by a back slash '\'
    The codes separated by a comma ',' are the 'names' of the dimensions used for identifying each (time) series.
    For each of these codes there is a file (with the same name plus the extension 'dic') in the directory dic.
    The code separated by a back slash '\' is the 'name' of the dimension of the sequence of values, e.g. 'time' (if this is a time series) or 'geo' (in the case of a geographical series).

– First column except the first line: sequence of codes separated by a comma ',' that represent the 'names' of the items (or instances or positions) of the dimensions. The label/title of these codes can be found in the 'dic' file that has the same name of the corresponding dimension.
– Other columns of the first line: sequence of codes corresponding to the items of the dimension.
– All other columns but the first line: sequence of values.

  Where available, flags are attached to values. The separator used between values and flags is a blank. If there are no flags, the value is followed by a blank.
– The decimal symbol used in the files is the dot '.'.

Note for Excel users: these files can be straightforwardly opened in Excel (see chapter Hints for Excel users).

2.2. The 'dft' format

DFT files are intended to efficiently store data organised as multi-dimensional tables. It is a compact way of storing figures and it uses codes for the positions in each dimension and the dimensions. Other files (dictionaries) are used to provide additional textual information.

Description of DFT sections	DFT file
General information	INFO Created: Tue 18 Mar 2008
Last update	LASTUP Tue 18 Mar 2008, 08:30:17
Type of DFT file	TYPE V
Characters used as delimiters	DELIMS ( ),@~
List of dimension codes	DIMLST (soft,theme,domain,table,indic,geo,time)
List of dimension categories	DIMUSE (R,R,N,N,V,V,V)

Table continuing….

Description of DFT sections	DFT file
Lists of position codes (one per dimension)	POSLST (newcronos) (theme1) (euro) (na_q) (na010,na010q1,na010q4,na020,na020q1,na020q4,na030,na030q1,na030q4,na040,na040q1,na040q4,na050,na050pc,na060,na060q1,na073,na074,na075,na076) (eu25,eu15,be,cz,dk,de,ee,gr,es,fr,ie,it,cy,lv,lt,lu,hu,mt,nl,at,pl,pt,si,sk,fi,se,uk,no) (2005q04,2005q03,2005q02,2005q01,2004q04,2004q03,2004q02,2004q00, 2003q04,2003q03,2003q02,2003q01,2002q04,2002q03,2002q02,2002q01, 2001q04,2001q03,2001q02,2001q01,2000q04,2000q03,2000q02,2000q01, 1999q04,1999q03,1999q02,1999q01,1998q04,1998q03,1998q02,1998q01)
type of DFT format (FORMAT0 or FORMATR)	FORMAT0 FORMATR
Character used for missing values	NOTAV :
List of values	VALLST (0) (2017484.1,2001955.6,1982021.2,1966906.6,1951177.7,1939337.9,1927135.0,1905575.9,1886010.4,1872216.7,1848105.6,1840077.6,1834075.6,1822790.7,1802773.8,1788589.1, [...] 94.0,92.2,91.1,89.2,86.0,80.6,81.5,84.3,84.2,81.5,76.4,76.7,74.9,68.7,65.4,62.0,61.4,57.8,54.4,53.2,54.0,52.8,53.3,51.9,51.3,51.4,:@296) (-1)

'Dft' files can be divided into two parts:

- A header with information about the table and a description of the dimensions used.
- A data set with values that can be displayed in the data presentation area of the CBX browsing and editing programs.

TYPE

The 'dft' files are always of type 'V', which stands for values. From this section onwards, space or tabulators must not be used in the text.

DELIMS

Delims, i.e. delimiters, tell you the beginning of a list (1st symbol), the end of a list (2nd symbol) and the separator between two common elements of a list (3rd symbol); the 4th symbol used in the 'rlc' format is a multiplier (@) and the 5th is the delimiter for flags. None of these characters may be used in any other way. To avoid ambiguity, these characters must not be present in other elements (such as position, dimension, value, etc.).

DIMLST

Dimlst is a list of all the dimensions used. This list contains service information, identification of the dataset and 'true' dimensions. The **DIMUSE** block specifies their use as:

– Reserved dimension (= R): service information including source database if applicable;
– Name dimension (= N): definition of the data in the source database;
– Variable dimension (= V): the true dimensions, definitions of the cells in the 'dft'.

The DIMUSE block refers to the block DIMLST.

– The first item on the DIMLST list is used as specified by the first item on the DIMUSE list.
– The second item on the DIMLST list is used as specified by the second item on the DIMUSE list.
– The last item on the DIMLST list is used as specified by the last item on the DIMUSE list.

The **POSLST** paragraph shows, for each dimension, the list of positions belonging to this dimension. (For R or N dimensions, this list must contain exactly one element). Empty lists are not allowed.

FORMAT

Format is used to specify the coding of the values in the 'dft' file. Different formats use different conventions to define the cells. The old original format (FORMAT0) used to store cells one by one, i.e. all the values were mentioned. The new format we have adopted (FORMATR) is an 'rlc' format, which uses a reduced VALLST: each sequence of identical values is written as '*one value followed by a multiplier*'.

Example:

the sequence:	**:,:,:,:,:,:,:**
is written as:	**:@7**
to be read as:	***7 times ':'***

NOTAV

Notav specifies the string used in a cell to represent an empty cell ('not available value'). The value declared in this field must correspond to the value used in the VALLST list of values.

2.3.   The 'sdmx–ml' format

The 'sdmx–ml' is the XML-based version of the Statistical Data and Metadata Exchange (SDMX) standards. In the bulk download facility 'Data' directory, each compressed 'gz' file includes both data and DSD (data structure definitions) files.

Complete information about the 'sdmx-ml' format can be found at: http://www.sdmx.org

3. DICTIONARIES ('DIC') FILES

3.1. In general

The 'dic' files give the definition and meaning of the codes used in the first columns of the 'tsv' files. They are available in German, English and French. They are stored in the 'de', 'en', and 'fr' sub-directories respectively.

There is a 'dic' file for each dimension used in the 'tsv' files (i.e. for each code that appears in the first column of first line of the files, excluding 'time' — see below).

Each 'dic' file includes the items of the corresponding dimension, followed by the corresponding 'labels' (separated by a 'tab').

3.2. For the 'TIME' dimension

There is no 'dic' file for the time dimension. The syntax of the codes used for this dimension is the following:

– Yearly data: **YYYYyMM** where YYYY is the year and MM is the number of the month (01 to 12) or '00' ('zero zero') if the value refers to the 'whole year' (e.g. is the annual total or average).

– Semi-annual data: **YYYYsSS** where YYYY is the year and SS is the number of the semester (01 or 02).

– Quarterly data: **YYYYqQQ** where YYYY is the year and QQ is the number of the quarter (01 to 04).

– Monthly data: **YYYYmMM** where YYYY is the year and MM is the number of the month (01 to 12).

– Daily data: **YYYYmMMdDD** where YYYY is the year, MM is the number of the month (01 to 12) and DD is the number of the day (01 to 28 or 29 or 30 or 31).

4. HINTS FOR EXCEL USERS

When you open a 'tsv' file, Excel shows you the Text Import Wizard. For a 'basic' import, you can just click on the Finish button and get the table in Excel (but don't forget that the decimal symbol used in the files is the dot '.', not the comma ',').

4.1. Data import and delimiters

For a more 'sophisticated' import, you can click on Next and, in the second step of the wizard, specify the character(s) used for delimiting the columns .Therefore if you choose:

Tab as delimiter (= the choice by default)

You will get as many columns as specified in the table_of_contents file, i.e. all the 'comma delimited' codes will be in the first column and the flags will be in the same cells as the values:

In the *Data preview* panel you can see what the table will look like.
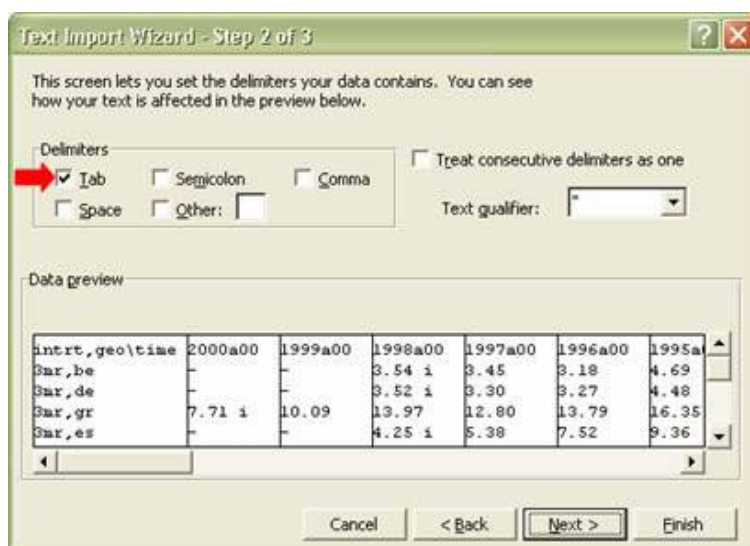
Tab and Space as delimiters

If in addition you uncheck the option 'Treat consecutive delimiters as one', you will get values and flags in separated columns. The total number of columns will be double (minus one) the number of columns given in the table_of_contents file:

In the Data preview panel you can see that values and flags are in different columns.
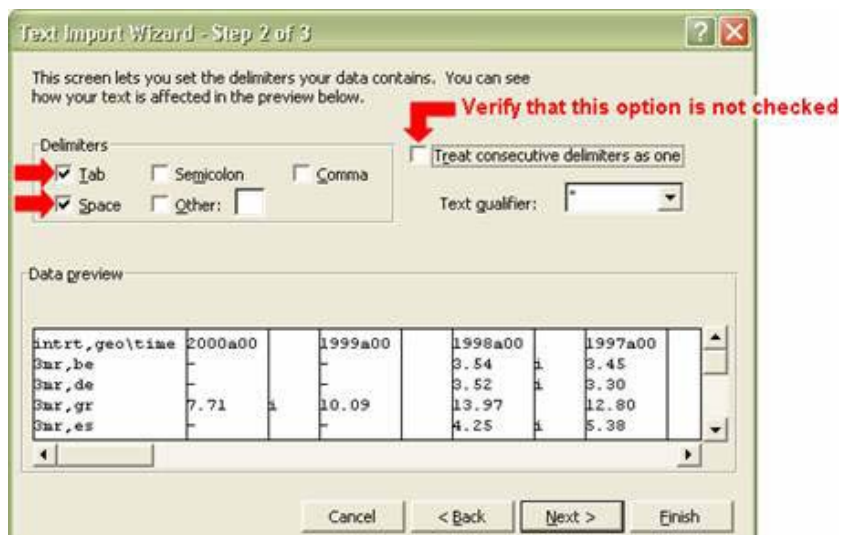
Tab and the Comma as delimiters

You will get the dimensions and their items in distinct columns (with the exception of the code that is separated by a back slash '\' in the first line);

In the *Data preview* panel you can see that the codes at the beginning of each line are in separate columns.

Tab and the Space and the Comma as delimiters

If you uncheck the option 'Treat consecutive delimiters as one', you will get:

In the *Data preview* panel you can see that codes, values and flags are in different columns.

4.2. Column data format

In the third and last step of the Text Import Wizard (see the picture below) you will be able to:

– Chose the column data format. The 'General' default option works well for importing 'tsv' files, unless you have decided to split the first column (using the comma as delimiter). In that case, we advise you to choose the 'Text' format for the columns that contain the codes of the dimensions and their items, otherwise Excel may interpret them incorrectly (e.g. the code for the region DEC1 may become 1/12/2000!);

– Chose to skip columns;

– Define the decimal separator, by clicking on the Advanced button (the decimal symbol used in the 'tsv' files is the dot '.').
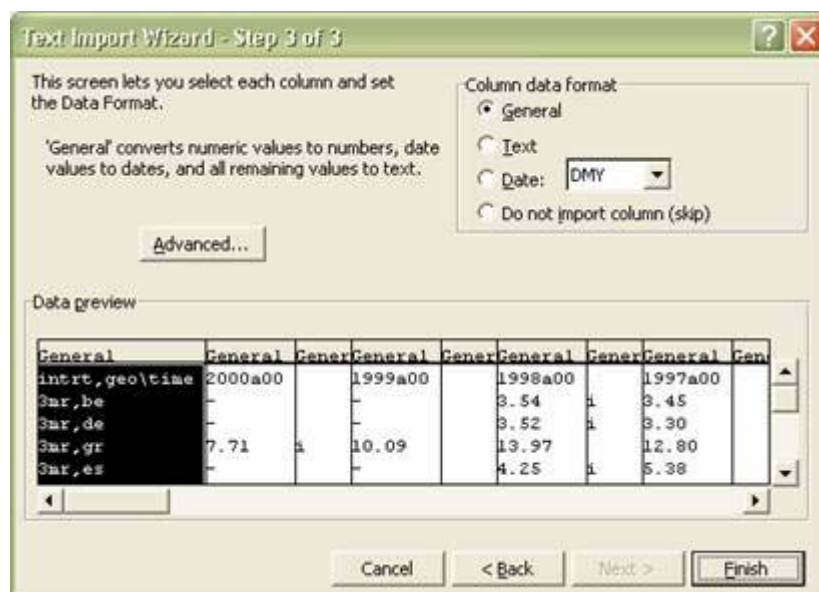
4.3. Data importation size limit

The 'notional' limits of an Excel sheet are 256 columns and 65536 lines (the actual limits depend on the memory available on your PC). If you try to open a file that goes beyond these limits, Excel will 'truncate' the excess columns and/or lines and give a message of this kind:

– In the case of time series with more than 255 periods, you will keep the most recent ones because, in 'tsv' files, the time series are sorted in descending order;

– If there are too many lines, you can open the file with Word and split it;

– Huge files can only be used/exploited with more powerful software such as Access, Dbase, Oracle, SAS, SyBase, etc.

4.4. Importation of values and flags

If you import in Excel the values and flags in different columns, the total number of columns will be double (minus one) the number of columns given in the table_of_contents file. If the file is then 'truncated' by Excel, the last imported column of values may lose its associated column of flags. In that case last column should be deleted.

If you import in Excel values and flags in the same columns, check for special combinations of values and flags such as '10 p', because Excel may automatically convert them into '10:00 PM'. To solve this problem you can format the cell(s) as 'text'

4.5. Data sorting with Excel:

How to sort the time series (i.e. the columns) in ascending order:

If you have imported in Excel values and flags in the same columns

– Select all the columns that contain values (typically all the columns but the first one) including the first row

– Open the menu Data and choose the item 'Sort'.

– Check that the option 'Header row' is selected (as shown in the picture).

– Click on 'Options' and select the option 'Sort left to right'

– Click on 'OK' and you will go back to the Sort menu.
– Click on 'OK' again.

If you want values and flags in different columns, then you can sort the columns in ascending order either by writing an ad hoc macro or by using a double import:

–   Import the file in Excel with flags and values in the same columns

–   Sort the columns as explained in the previous paragraph (a)

–   Save the file as Text (Tab delimited)

–   Import the file in Excel with flags and values in separate columns (as explained in previous pages).