Bulk download guidelines

including FAQ

Last modification:	08-NOV-2019
Source:	Eurostat
Contact:	User support

Table of Contents

1. CONTENT OF THE BULK DOWNLOAD REPOSITORY

Directory	Content
[data]	All Eurostat tables as available in the data navigation tree
[dic]	Code lists (dictionaries). They include the files: 'Dimlst.dic' which provides a description of all code lists 'Table_dic.dic' corresponding to the list of all tables
[metadata]	Reference metadata including 'Euro-SDMX Metadata Structure (ESMS)'
metabase.txt.gz	Structure definition of datasets and tables including all codes
table_of_contents…	Table of content in three languages (EN, DE and FR) and three formats (xml, txt and pdf)
BulkDownload_Guidelines.pdf	documentation including frequent asked questions
[comext]	International trade database (COMEXT)

Update: the information stored in the bulk download repository is updated twice a day, at 11:00 and 23:00.

2. DETAILED INFORMATION

2.1. DATA DIRECTORY

This directory includes approximately 6300 datasets and tables available in two formats, 'tsv' and 'sdmx–ml', compressed in zip files ('gz'). All 'gz' files can be uncompressed with the help of WinZip.

To access and download data, choose your language first, select the format and click on the code(s) of the table you are looking for.

A filtering feature is available by clicking on the code (name), size, date and type.

2.1.1. 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 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').

– ATTENTION: cells for which there is "no data available" are NOT stored in the tsv files on the bulk download, doing so would explode the size of the tsv files. When using the on-line extraction tools like Tables, Graphs and Maps or Data Explorer, such cells contain the symbol ":"

EXAMPLE

Dataset with time series (with made-up values)

unit,s_adj,partner,flow,indic,geo\time	2004m05	2004m04	2004m03	2004m02
mio-eur,nsa,ext_eurozone,net,bp-100,eurozone	11148	10660	13398	9437
mio-eur,nsa,ext_eurozone,net,bp-200,eurozone	3386e	539	-185	-432
mio-eur,nsa,ext_eurozone,net,bp-300,eurozone	-5626e	-6696i	1902	919
mio-eur,nsa,ext_eurozone,net,bp-379,eurozone	-5758e	-4165	-3970	-4703
mio-eur,nsa,ext_eurozone,net,bp-993,eurozone	3151.5e	338.7i	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).

2.1.2. 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 repository '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

2.2. CODE LISTS FILES ('DIC')

2.2.1. Definitions

Code list: predefined set of terms from which some statistical coded concepts take their values.

Statistical concept: a statistical characteristic of data.

Further definitions: see Eurostat CODED

2.2.2. In the bulk download

The complete list of 'code lists' can be found in the file 'Dimlst.dic'.

Code lists are defined by a code and a label. Ex:

Code:	MARSTA
Label:	Marital status

In the bulk download, the code lists are classified by code only (the labels are not available).

Code lists have a two columns structure of code and labels that are used in the datasets definitions. The separator is a 'tab'. Ex: MARSTA includes:

TOTAL	Total
SIN	Single persons (never in legal union)
LUN	Persons in legal union (married or in registered partnership)
MAR	Married persons
MARO	Persons in an opposite-sex married couple
Etc.	

Code lists are available in German, English and French and stored in the 'de', 'en', and 'fr' sub-directories respectively.

There is a 'dic' file for each code list used in the 'tsv' or 'sdmx-ml' files. In the tsv, they appear at the top left (first row and first column); in the sdmx-ml format they are defined by the 'CodeList' id. Ex for the dataset 'lfso_05reduc1':

- Tsv format: sex,age,ynlfs,worktime,marsta,time\geo etc…
- Sdmx-ml format:
  - <structure:CodeList id="CL_MARSTA" isFinal="true" agencyID="EUROSTAT">

2.2.3. '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).

2.2.4. Mapping tables

Code lists do have a life cycle. Major revisions of code list are recorded in mapping files that give the correspondence between old and new code/label.

These tables are available in the 'dic' repository.

Example: 'coicop_mapping.xlsx'.

2.3. METADATA

The directory includes the following reference metadata available in the data navigation tree:

- 'Euro-SDMX Metadata Structure (ESMS)': set of international standards for exchange of statistical information between organisations.

- ESS Standard for Quality Reports Structure (ESQRS): standard structure for the collection of quality report.
  The ESQRS is also used for the collection of national quality reports from National Statistical Institutes.

See also further information on Eurostat web pages:
. Metadata: http://ec.europa.eu/eurostat/data/metadata/metadata-structure
. Statistics explained:
https://webgate.ec.europa.eu/fpfis/mwikis/sdmx/index.php/ReferenceMetadata

These metadata are accessible in the data navigation tree or the the data browsers via the icon  . They might correspond to specific tables or, which is most of the case, to groups of datasets or tables.

2.4. METABASE

The 'Metabase' file provides the structure definition for datasets and tables. The first column corresponds to the code of the table, the second one to the dimensions and the third one to the codes used for each dimension.

The example below gives a preview of the file for the dataset 'aact_ali01'.

aact_ali01	geo	EU28
aact_ali01	geo	EU27
aact_ali01	geo	EU25
aact_ali01	geo	EU15
aact_ali01	geo	EA19
aact_ali01	geo	EA16
aact_ali01	geo	EA12
aact_ali01	geo	EA11
aact_ali01	geo	BE
aact_ali01	geo	BG
aact_ali01	geo	CZ
aact_ali01	geo	DK
aact_ali01	geo	DE
aact_ali01	geo	EE

2.5. TABLE OF CONTENTS

2.5.1. Table of content formats and content

The 'table of contents' (TOC) file provides information on datasets and tables available on the Eurostat website. The TOC is available in three formats: txt, xml and pdf.

Table 1: TOC by format and type of information

	Type of information	pdf	txt	xml
.	code	x	x	x
.	title - english	x	x	x
.	title - french		x	x
.	title - german		x	x
.	sub-title - english			x
.	sub-title - french			x
.	sub-title - german			x
.	short description - english			x
.	short description - french			x
.	short description - german			x
.	type (dataset, table, folder)		x	x
.	last update of data		x	x
.	last table structure change		x	x
.	data start		x	x
.	data end		x	x
.	number of values			x
.	reference metadata (ESMS) access via url			x
.	reference metadata (ESMS) download file (zip)			x
.	data download in format tsv	x		x
.	data download in format sdmx			x
.	direct access to data via data browser	x		
.	hierarchy of the data navigation tree	x	x	x

2.5.2. REMARKS

CODE

Each item (dataset, table and folder)[1] 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.

[1] **'Tables'** correspond to all data available in 'Tables by themes'. They offers a selection of the most important Eurostat data in a user-friendly way. All data are presented in simple two or three dimensional tables.

'Database' correspond the all data available in 'Database by themes'. They contain the full range of data publically available at Eurostat. They are presented in multi-dimensional tables with various selection features and export formats.

Both branches are organised according to nine statistical themes.

However, some of them can be classified at two different places in the data navigation tree.

Examples:

- Classification of a dataset under two different themes. For instance Regional transport data are available in both:
  'Theme 1 / Regional statistics'
  '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 OF DATA

Indicates the last time the dataset/table was updated.

LAST TABLE STRUCTURE CHANGE

Indicates the last time the dataset/table structure 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.

2.6. 'ESTAT_<date>' FILES

There are 3 files of this type:

- ESTAT_<date>.zip: This zip file contains the description in rdf format of all datasets and tables available in the data navigation tree.

- ESTAT_<date>_updates.zip: This zip file contains the description in rdf format of all datasets and tables available in the data navigation tree which were updated since the previous update cycle.

- ESTAT_<date>_delete.zip: This zip file contains the description in rdf format of all datasets and tables available in the data navigation tree which were deleted since the previous update cycle.

3. HINTS FOR EXCEL USERS

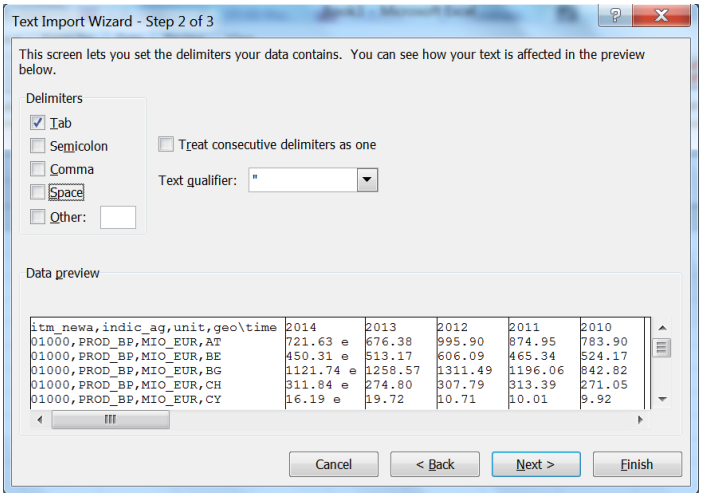
When you open a 'tsv' file, Excel shows you the Text Import Wizard. For a 'basic' import, you can just click on the Finish button and get the table in Excel (but don't forget that the decimal symbol used in the files is the dot '.', not the comma ',').

3.1. 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:

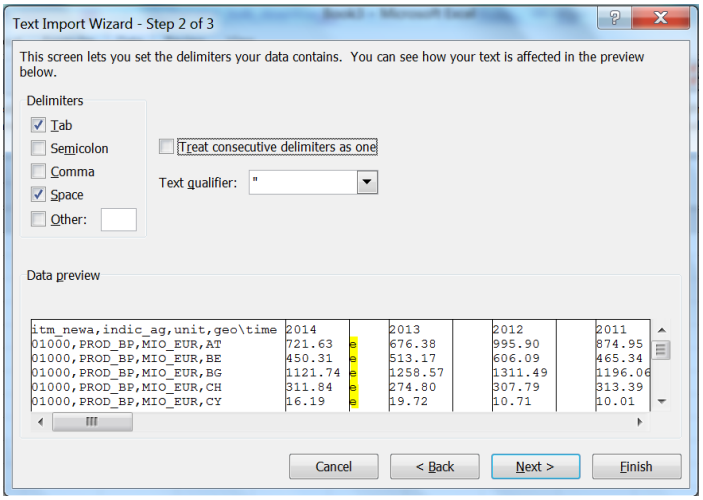
TAB AS DELIMITER (= THE CHOICE BY DEFAULT)

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.



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.



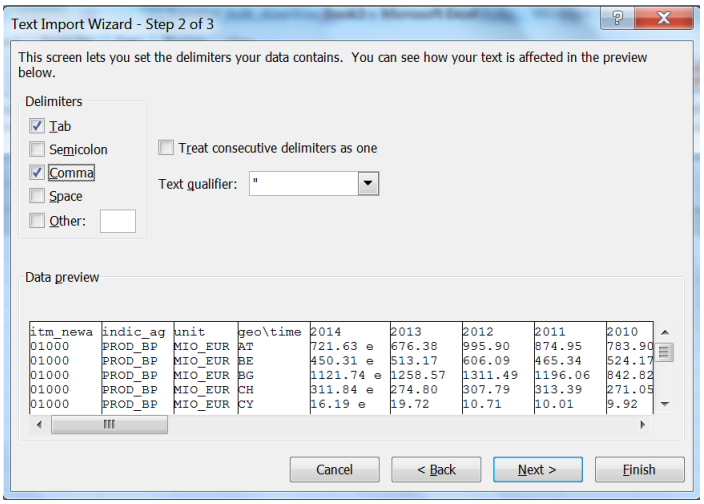
In the Data preview panel you can see that values and flags are in different columns.

Verify that the 'Treat consecutive delimiters as one' is not checked.

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).

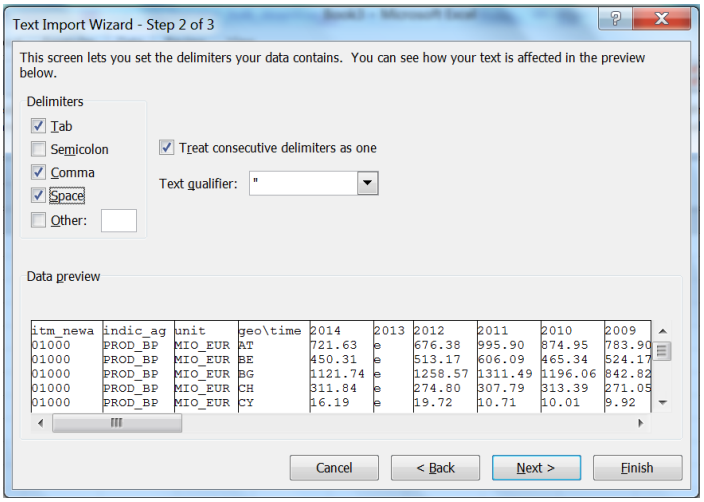
In the Data preview panel you can see that the codes at the beginning of each line are in separate columns.



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.



3.2. 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 '.').

3.3. Data importation size limit

The 'notional' limits of an Excel sheet are 4445 columns and 1048576 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 an error message.

– If there are too many lines, you can open the file with a text editor (txt) and split it;

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

3.4. 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. FREQUENT ASKED QUESTIONS (FAQ)

4.1. In what format are datasets supplied?

Data files in the bulk download are supplied in the following formats:

- SDMX-ML
- tsv (tab separated value)

Any of these formats will enable you to easily import the data in a tool of your choice.

The "official" format for data exchange adopted by Eurostat is SDMX-ML. You can find more information on SDMX on the following address: http://www.sdmx.org

4.2. Where can I download data files?

All data can be downloaded from the bulk download repository available at the following location: http://ec.europa.eu/eurostat/estat-navtree-portlet-prod/BulkDownloadListing

4.3. Where can I find a list of table codes?

There are several places where to find tables and their codes:

- The data navigation tree. The codes are indicated in brackets behind the titles.
- Table of contents in the bulk download. They are available in three languages (English, French and German) and formats (xml, pdf and txt).
- In 'table_dic.dic' in the bulk download 'dic' repository (English only).

4.4. How do I know if a table has been updated?

Consult the table of contents file (xml format), which is available from the bulk download repository.

Then, simply search for the data codes you're interested in and check the *lastUpdate* date. Below is an extract of the *table_of_contents.xml* file.

It includes also a *lastModified* date which corresponds to the last table structure change.

Example: Table of content in xml format

```
<nt:code>t_demo_gen</nt:code>
- <nt:children>
  - <nt:leaf type="table">
      <nt:title language="en">Total population</nt:title>
      <nt:title language="fr">Population totale</nt:title>
      <nt:title language="de">Gesamtbevölkerung</nt:title>
      <nt:code>tps00001</nt:code>
      <nt:lastUpdate>21.09.2011</nt:lastUpdate>
      <nt:lastModified>21.09.2011</nt:lastModified>
      <nt:dataStart>2000</nt:dataStart>
      <nt:dataEnd>2011</nt:dataEnd>
      <nt:values>455</nt:values>
      <nt:unit language="en">At 1 January</nt:unit>
      <nt:unit language="fr">Au 1<sup>er</sup> janvier</nt:unit>
      <nt:unit language="de">Am 1. Januar</nt:unit>
      <nt:shortDescription language="en">The inhabitants of a given area on 1 January of the year in question (or, in some cases, on 31 December of the previous year). The population is based on data from the most recent census adjusted by the components of population change produced since the last census, or based on population registers.</nt:shortDescription>
      <nt:shortDescription language="fr">Ensemble des habitants d'une zone donnée au 1<sup>er</sup> janvier d'une année donnée (ou, dans certains cas, au 31 décembre de l'année précédente). Cette population est soit calculée à partir des données obtenues du dernier recensement, ajustées par les données sur les composantes de l'accroissement de population depuis ce dernier recensement, soit à partir des registres de population.</nt:shortDescription>
      <nt:shortDescription language="de">Die Einwohner eines bestimmten Gebietes am 1. Januar des betreffenden Jahres (oder in einigen Fällen am 31. Dezember des vorangegangen Jahres). Die Einwohnerzahl basiert auf den Daten der jüngsten Volkszählung, bereinigt unter Berücksichtigung der Komponenten des Bevölkerungswachstums seit der letzten Volkszählung, oder auf den Daten der Bevölkerungsregister.</nt:shortDescription>
      <nt:metadata>http://epp.eurostat.ec.europa.eu/cache/ITY_SDDS/en/demo_pop_esms.htm</nt:metadata>
      <nt:downloadLink format="tsv">http://epp.eurostat.ec.europa.eu/NavTree_prod/everybody/BulkDownloadListing?file=data/tps00001.tsv.gz</nt:downloadLink>
      <nt:downloadLink format="dft">http://epp.eurostat.ec.europa.eu/NavTree_prod/everybody/BulkDownloadListing?file=data/tps00001.dft.gz</nt:downloadLink>
      <nt:downloadLink
        format="sdmx">http://epp.eurostat.ec.europa.eu/NavTree_prod/everybody/BulkDownloadListing?file=data/tps00001.sdmx.zip</nt:downloadLink>
    </nt:leaf>
  - <nt:leaf type="table">
      <nt:title language="en">Total population, Candidate countries and potential candidates</nt:title>
      <nt:title language="fr">Population totale, Pays candidats et candidats potentiels</nt:title>
      <nt:title language="de">Gesamtbevölkerung, Kandidaten- und potentielle Kandidatenländer</nt:title>
      <nt:code>tgs00027</nt:code>
```

4.5. When are datasets updated?

Datasets are updated twice a day, at 11:00 and at 23:00. The updating procedure only updates the datasets which have new data or have a structural change. Other datasets are left untouched.

4.6. Where can I find more information on Flags and special values (:) used in Eurostat's online database?

Flags are codes added to the data and defining a specific characteristic. A special value is a code that replaces the real data. More information is found here:
http://ec.europa.eu/eurostat/data/database/information

4.7. How many datasets are available?

In January 2019 about 6300 datasets were accessible from the bulk download area.

4.8. What can be found in the metabase file?

The metabase file includes the structure definition of all datasets available in 'Database by themes' of the data navigation tree i.e; the dimensions used and the related positions;

Example for the table '*Agricultural Labour Input Statistics: absolute figures (1 000 annual work units)*' which is identified with the code 'aact_ali01'.

CODE	DIMENSION	POSITION
aact_ali01	geo	EU28
aact_ali01	geo	EU27
aact_ali01	geo	EU25
…	…	…
aact_ali01	geo	UK
aact_ali01	geo	IS
aact_ali01	geo	NO
aact_ali01	geo	CH
aact_ali01	geo	MK
aact_ali01	geo	TR
aact_ali01	itm_newa	40000
aact_ali01	itm_newa	41000
aact_ali01	itm_newa	42000

4.9. Does Eurostat offer other possibilities to automatically download data?

Yes. Eurostat also offers to download data by means of web services based on SDMX, JSON and Unicode.

Here you can find more information on this service:

- SDMX: http://ec.europa.eu/eurostat/web/sdmx-web-services/about-this-service
- JSON and Unicode: http://ec.europa.eu/eurostat/web/json-and-unicode-web-services/about-this-service