Bulk download in ILOSTAT:

Instructions and Guidelines

ILO Department of Statistics

2019-05-15

Contents

Acknowledgements

This document is the result of an extensive collaboration among various members of the Data Production and Analysis Unit of the ILO's Department of Statistics. The first draft was prepared by David Bescond and Rosina Gammarano, and it benefited from valuable comments by Steven Kapsos, Yves Perardel and Marie-Claire Sodergren.

Introduction

The Bulk Download facility featured in ILOSTAT is a key functionality allowing advanced users to easily download the data required. Through the Bulk Download, users can access entire individual datasets or even the complete ILOSTAT database. The files downloaded through these means are presented in csv format (comma separated values), which can later on be imported using the user's preferred tool. ILOSTAT's Bulk Download is the basis for ILOSTAT's R package ("Rilostat'), which was designed to give data users the ability to access the ILOSTAT database, search for data, rearrange the information as needed, download it in the desired format, and make various data visualizations, all in a programmatic and replicable manner, with the possibility of quickly re-running the queries as required. This document provides instructions and guidelines for using the Bulk Download facility, as well as detailed information on how to automate the downloading of datasets.

Contents of the bulk download repository

The Bulk Download facility contains in its repository various types of files containing data, metadata and documentation. These include datasets in csv format, "dictionaries" for all the codes used in the datasets, and this document on guidelines and instructions to use the Bulk Download. The directories containing the datasets by indicator (for instance the unemployment rate by sex and age) or by ref_area (abbreviation for

Table 1: Contents of the Bulk Download repository

Directory	Contents
[**indicator**]	All ILOSTAT tables presented by indicator and frequency
[**ref_area**]	All ILOSTAT tables presented by ref_area and frequency
[**dic**]	Dictionaries of all the codes used (code lists)
BulkDownload_Guidelines.pdf	Documentation, including guidelines and instructions

reference area which is the relevant geographical unit such as a country) present, in addition to all the data tables available, a table of contents detailing the list of tables available by indicator or reference area and the time period covered by the corresponding data. The following table summarizes the contents of the bulk download repository and provides a brief description of each item included in it.

All of the information stored in the Bulk Download repository is updated once a week, every Sunday at 10:00 pm (Europe/Paris time zone). The updating procedure only involves datasets for which there is new data or that have undergone a modification or a structural change. All other datasets are left untouched.

The following sections go through each item of the Bulk Download repository, providing in-depth descriptions of their structure and uses.

Data directories: tables by indicator or by ref_area

The Bulk Download gives access to ILOSTAT datasets through two different directories, based on two different ways of presenting the corresponding tables: organizing them by 'indicator' (and frequency) or by 'ref_area' (and frequency). The indicator refers to the title of each specific table, including the represented variable and the eventual disaggregations used for it (for instance, 'labour force by sex and age', 'employment by sex and economic activity' and 'unemployment rate by sex, age and rural / urban areas' are ILOSTAT indicators). The ref_area (from reference area) refers to the geographic areas for which data are available. Since ILOSTAT includes both country-level data and regional and global estimates, the reference area can either refer to countries, to regions (geographic regions such as Africa, Americas or Arab States, income groups such as low income countries, or other groups such as the BRICS or the G20) or to the world as a whole. However, it is important to note that global and regional estimates are only available for some indicators, and so most datasets would only include country-level data. The frequency refers to whether the various data points are annual, quarterly or monthly. Data directories, whether by indicator or by ref_area, are presented in csv format as compressed zip files ('gz'). All 'gz' files can be uncompressed using WinZip or 7zip. For further information on the csv files, see the following section.

After selecting one of the two approaches proposed (tables by indicator or by ref_area) by clicking on the name of the directory, you can access and download the desired data by clicking on the code name(s) of the table(s) you are looking for.

The [dic] directory (dictionaries) provides dictionaries of all the code names needed to identify the indicator or reference area that you are looking for. For reference, please note that code names all follow the same structure. The indicator code names include, in this order, the code of the topic, the represented variable, the disaggregations included ('NOC' for 'no classification' if there is no disaggregation), the unit ('NB' for absolute values or numbers and 'RT' for percentages or rates) and the frequency ('A' for annual data, 'Q' for quarterly data and 'M' for monthly data). Similarly, the code names of the files by reference area refer to the country (ISO Alpha-3 country code) or the region (codes starting in X) and the frequency ('A' for annual data, 'Q' for quarterly data and 'M' for monthly data).

The table below show the contents of the [indicator] directories.

The table below show the contents of the [ref_area] directories.

\begin{table}[t]

Table 2: Contents of the indicator directory, approximately 500 datasets

Files	Contents
table_of_contents_en	Table of contents in EN
table_of_contents_fr	Table of contents in FR
table_of_contents_sp	Table of contents in SP
EAP_TEAP_SEX_AGE_NB_A.csv.gz	Dataset with annual labour force by sex and age
EMP_DWAP_NOC_RT_A.csv.gz	Dataset with annual employment-to-population ratio
...	...

\caption{Contents of the ref_area directory, approximately 700 datasets}

Files	Contents
table_of_contents_en	Table of contents in EN
table_of_contents_fr	Table of contents in FR
table_of_contents_sp	Table of contents in SP
ABW_A.csv.gz	Dataset containing all annual data available for Aruba
ABW_M.csv.gz	Dataset containing all monthly data available for Aruba
...	...

\end{table}

Downloaded 'csv' (comma separated values) files format

Files in 'csv' format are files storing tabular information (whether numbers or text) in the form of plain text, as comma separated values. That is, the columns (or fields) from the original table are separated by commas, allowing for each row or line of the file to correspond to one data record (the data record may thus consist of one or more fields, separated by commas). These files can easily and straightforwardly be opened in Excel.

In ILOSTAT 'csv' files, the first row contains the headers (of the fields or columns). The subsequent rows present the data records, consisting of the key of the record (the 'names' of the dimensions used to identify each record, including the data collection, the reference area, the source of the data, the classifications used, etc., referring to all fields from 'collection' to 'time'), the observation value ('obs_value') and any other metadata available (such as the geographical coverage of the source or the specific definitions used for some concepts, referring to all fields from 'obs_status' to 'note_source').

All of the labels corresponding to the code names used as field headers in the csv files available for download are presented in the code lists dictionary ([dic] files, see following section for further information). The only code name not explained in the [dic] files is 'obs_value', which corresponds to the observation value. It is noteworthy that there is no dictionary (or no 'dic' file) for the time dimension.

The syntax of the codes used for this dimension is the following:

- Yearly data: YYYY where YYYY is the year.
- Quarterly data: YYYYqQ where YYYY is the year and Q is the quarter (the number corresponding to the quarter from 1 to 4).
- Monthly data: YYYYmMM where YYYY is the year and MM is the month (the number corresponding to the month from 01 to 12).

The number format applied in ILOSTAT files uses a dot as the decimal symbol ('.').

Code lists directories : dictionary files ('dic')

Code lists are predefined sets of terms from which statistical concepts (statistical characteristics of data) that have been coded take their values. All of the code lists presented in ILOSTAT are available in three languages ('en' for English, 'fr' for French and 'sp' for Spanish). All ILOSTAT code list files have the same structure, consisting of three columns: the variable name or code ('var_name'), the variable label or description of the code ('var_label') and a number used to sort the information in the file ('var_sort'). The following table provides an example of ILOSTAT code list.

\begin{table}[t]

\caption{Extract of code list file 'indicator_en.csv'}

Indicator	Indicator.label
SDG_0852_SEX_AGE_RT	[8.5.2] Unemployment rate (%)
POP_XWAP_SEX_AGE_NB	Working-age population by sex and age (thousands)
...	...

\end{table}

The various code lists available in English, French and Spanish in the [dic] directory correspond to the fields used in the downloaded csv files described in the previous section (except for the 'obs_value' field used for the observation value and not requiring a dictionary with labels). The following table enumerates the code lists included in the [dic] directory.

\begin{table}[t]

\caption{Extract of code list file 'indicator_en.csv'}

Variable name (used also as code list name)	Brief description
collection	Data collection or compilation from which the data was derived, from all the various data compilations carried out by the ILO and disseminated in ILOSTAT
ref_area	Reference area this can refer to countries, geographic regions, groups of countries (by income level or others) or the world
source	The specific source of the data, including information on the country or region for which it is used and the main type of source (population census, labour force survey, administrative records, etc.) as well as the precise name of the source.
indicator	The indicator, including information on the represented variables, the classifications used (if any) and the unit.
sex	The disaggregation by sex and the items of this disaggregation.
classif1	All classifications used as the first disaggregation in the various indicators available (excluding the disaggregation by sex, which is treated separately) and the corresponding classification categories or items.
classif2	All classifications used as the second disaggregation in the various indicators available (excluding the disaggregation by sex, which is treated separately) and the corresponding classification categories or items.
obs_status	The value status or flags on the values, such as breaks in series or provisional values.
note_classif	Metadata and/or footnotes related to the classifications used and the specific classification categories.
note_indicator	Metadata and/or footnotes related to the indicator.
note_source	Metadata and/or footnotes related to the data source.

\end{table}

It should be noted that these code lists present only the label corresponding to each code. For further methodological information, including definitions of the main statistical terms used in ILOSTAT, detailed indicator descriptions and statistical standards, refer to the metadata section of ILOSTAT www.ilo.org/ilostat/faces/ilostat-home/metadata.

Tables of contents

The two data directories included in ILOSTAT's Bulk Download facility ([indicator] directory and [ref_area] directory) present a table of contents, available in csv format and in three languages ('en' for English, 'fr' for French and 'sp' for Spanish). These tables of contents list all of the data files available for download in the corresponding directory, and provide summary information on each data file. The table of contents of the [indicator] directory lists all the indicators available, with the label of the indicator and the frequency of the data. The table of contents of the [ref_area] directory lists all the reference areas available (countries, regions, groups of countries), with the label of the reference area and the frequency of the data. Both tables of contents also indicate the size of each data file, the time period covered by the data in the file and the date when the data file was last updated. Since ILOSTAT's datasets include projections of the main labour market indicators, the time period covered by some data files can go as far as 2050. The codes or identifiers used in the tables of contents for the indicators and reference areas in the first field or column ('id') are unique and allow for the unequivocal identification of the corresponding item. The two tables presented next show extracts of the tables of contents of the [indicator] and the [ref_area] directories.