**Notes for input of new tax schedule data and checking**

**SOURCES**

Hierarchy of sources:

1. EY reports: when available, use the **Worldwide Estate and Inheritance Tax Guide (“EYb”)**, if not, use **Worldwide Personal Tax and Immigration Guide (“EYa”)**

Archive 2006-2023 EYa, 2012-2023 EYb: <https://www.ey.com/en_gl/tax-guides/tax-guide-library-archive>

Last source available 2023-2024: <https://www.ey.com/en_gl/tax-guides>

1. PWC World Tax Summaries: <https://taxsummaries.pwc.com/>, covering information for the CURRENT year only (no archive). Useful to check countries neither in EYa nor in EYb.
2. YaleInheritance data (Excel), up to 2009

<https://isps.yale.edu/research/data/d025>

1. TID data for the year of introduction of the tax

<https://tid.seelkopf.eu/tidData.php>

1. When available, tax law for the country

There are several other sources that we used for the previous version of the dataset, but for the moment we start from these ones. The absolute priority will be EY reports.

1. **STRUCTURE OF THE MAIN FOLDER taxsched\_input**

The MAIN FOLDER **…\gcwealth\handmade\_tables\taxsched\_input** contains some general files which are independent of the source, and a subfolder for each type of **source.** Up to now, we focused on the two folders EY EIG Guide and EY Personal Tax Guide. The main aim is now to populate the other two folders TIDData and Yale\_Inheritance\_Data.

A screenshot of a computer

Description automatically generated

General files:

* These **Operational notes**
* A Worfflow\_record.xslx Excel file to keep track of all the progress made, to be filled every day of work
* The Excel file Input\_format to take as reference for the data input in each source-country
* The **Python code split\_pdf.py** to split the EY PDFs into single country PDFs
* The **Stata codes** for verification of **validity** (data have valid entries) and **consistency** (data are compatible between columns) and for **automatic filling** of missing columns when possible. The code is made of two parts:
  1. **Validation\_code.do**, to be opened and modified for the specific source once the Excel is ready
  2. **Eigt\_verify.ado** not to be modified (it is automatically called by **Validation\_code.do**, you do not need to interact with it unless you are curious).

1. **STRUCTURE OF THE SUB FOLDERS** EY EIG Guide AND EY Personal Tax Guide

Each EY subfolder (either …\gcwealth\handmade\_tables\taxsched\_input\EY EIG Guide or …\gcwealth\handmade\_tables\taxsched\_input\EY Personal Tax Guide) contains:

* as many subfolders as the countries covered by the source
* the original PDFs (one per year) of the source downloaded from the web
* Two Word files EIG taxes Question Schemeand Wealth Taxes Question Schemeguiding the question flows to ask to ChatGPT

Immagine che contiene testo, schermata, schermo, software

Descrizione generata automaticamente

For the moment, you see there four folders for Australia, Belgium, Italy, Philippines, for which we prepared some examples using the 2022-2023 EY reports.

1. **STRUCTURE OF THE SUB FOLDERS** FOR EACH COUNTRY

Each subfolder named as a country contains:

* The Excel file EYa\_country.xlsx or EYb\_country.xlsx containing the data taken or inferred from the source in the sheet **Data**, and several sheets of legend and documentation. The reference file for this excel is the Input\_format that you find in the mail folder.
* The parts of the PDFs related to the country, generated through the Python code to query ChatGPT
* The Excel file GPT\_questions.xlsx containing the questions and answers of the conversation with ChatGPT for each country included in the source. The excel contains one sheet per year covered by the source.
* The text fileverification\_logfile.txt reporting the output of the data verification performed through Stata. It is automatically created by the Stata do file in the main folder
* The data output in Stata format data\_longformat.dta generated by the Stata do file in the main folder

1. STRUCTURE OF THE SUB FOLDERS TIDDATA AND YALE\_INHERITANCE\_DATA

Both folders contain the corresponding original source in an excel file named TIDData.xlsx and Inheritance\_taxation\_Masterfile.xlsx, respectively. For each folder, it is necessary to create as many subfolders as the countries covered by the source. For example, if Yale covers 10 countries, the folder Yale\_Inheritance\_Data will contain 10 subfolders with the countries covered by the source.

1. **STRUCTURE OF THE SUB FOLDERS** FOR EACH COUNTRY

Each subfolder named as a country contains:

* The Excel file Yale\_country.xlsx (or TIDD\_country.xlsx) containing the data collected following the format and instructions of the Excel file Input\_format.

1. **INPUT STEPS**
   1. Download the PDF of the source and save it in the source folder.
   2. Get from the whole PDF the part related to a single country using the Python code split\_pdf.py and save it into the subfolder of the country (create the folder if it is not already there). Proceed in alphabetical order.

These are the only two lines that require manual input: pdf\_path specifies the current location of the PDF to be split, and the result is needed to define the range of pages delimited by two cities within the country; insert your path and run:

Immagine che contiene schermata, testo, Carattere

Descrizione generata automaticamente

* 1. Copy the Input\_format file into the folder for the country and change the name to EYa\_countryname or EYb\_countryname depending on the source.
  2. Query ChatGPT using the instructions in Question Scheme.doc. Start with EIG taxes which have priority. Document all the questions and answers in the Excel file GPT\_questions creating a different sheet for each year.
  3. Report the output in the Excel EYa\_country.xlsx or EYb\_country.xlsx in the **sheet “Data”**, following the instructions in the “Legend” sheet.
  4. Open in Stata the validation do file Validation\_code.do. The very first time on your computer, you need to save your username in the empty space as you see below:

**Immagine che contiene testo, schermata, software, Software multimediale

Descrizione generata automaticamente**

Your username is displayed in Stata when you run lines 15-17: copy it in the blank space in line 25 (Martino) or 30 (Ghulam). Once you save it the first time, you will be able to run the entire do file without manual changes.

Before running the do file, set the source and the country names in lines 38 and 39 (they must correspond to the names in the folders!). Then run the do file.

If errors or warnings pop up, correct them in Excel and repeat the check until no message is generated.

* 1. Check in the data\_longformat.dta that the information is coherent with the input you made. Bro with the condition bracket == 0 if you want to check bracket-independent information.