Data Fundamentals: Scraping Data From PDFs

For this lab exercise, we are going to review different data formats you will usually come across in your work. We will start of with some of basic and common data formats and explore open-source tools we can leverage on to get our data into a format that is easy to work with.

In reference to the School of Data Pipeline, this lab exercise focuses on the third stage: **GET.**

Before we move forward, here is a refresher on data formats:

- Machine readable, structured: These are generated by a computer, and are organized in rows and columns. For example - CSV (comma-separated values), TSV (tab-separated values), Excel(.xls)

- **Unstructured:** These are sometimes generated by a computer, but are not organized as data tables by the computer. For example – some PDF, Word, and bitmap images (GIF, JPEG, PNG, BMP)

As part of your day-to-day work, you must have come across data in these file formats:

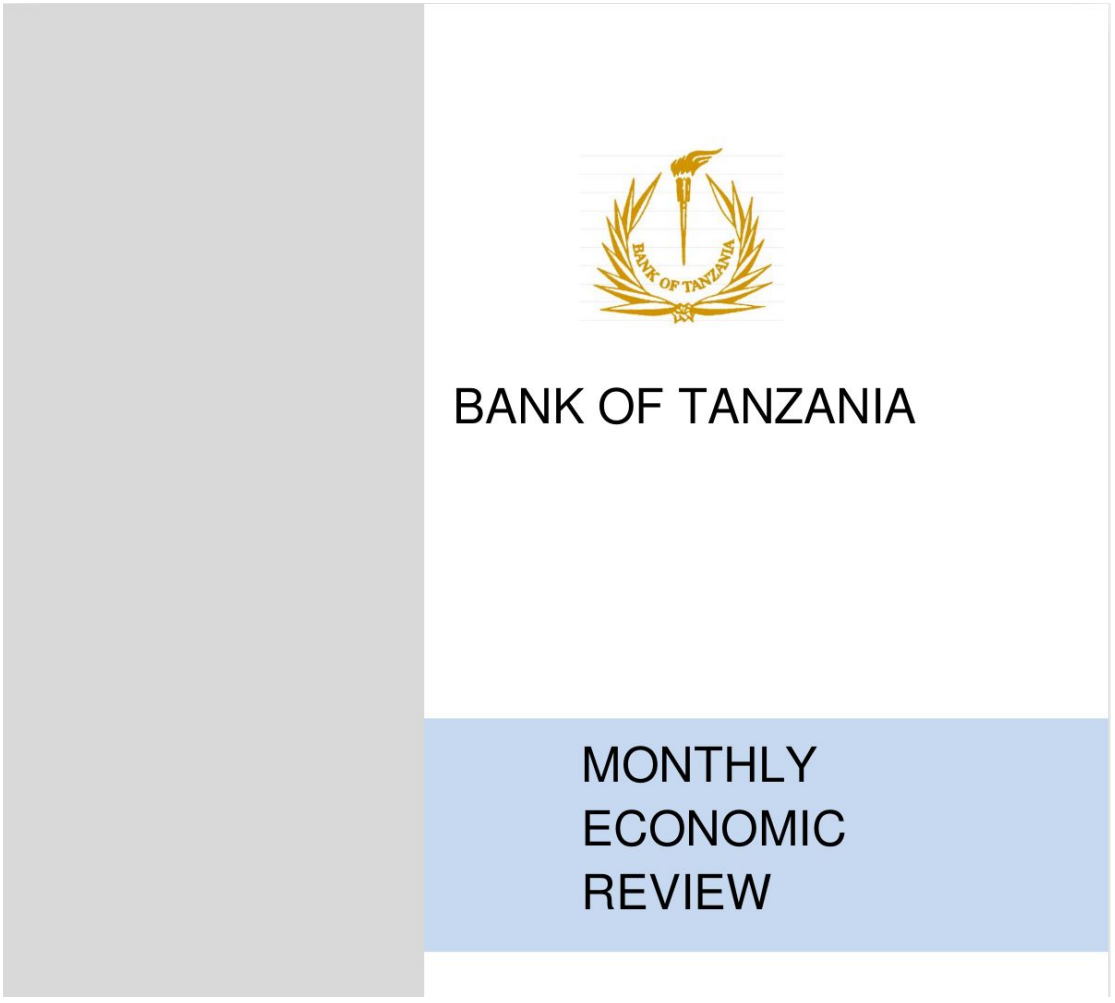
● **Portable Document Format (PDF):** These files may include charts that contain data, but the data is saved in a unified document with text.

● **Excel file (XLS):** These files save data as tables, which are readable by spreadsheet software such as Microsoft Excel, LibreOffice Calc or Google Sheets.

● **Comma separated values (CSV):** These are plain text files with each data point separated by a comma

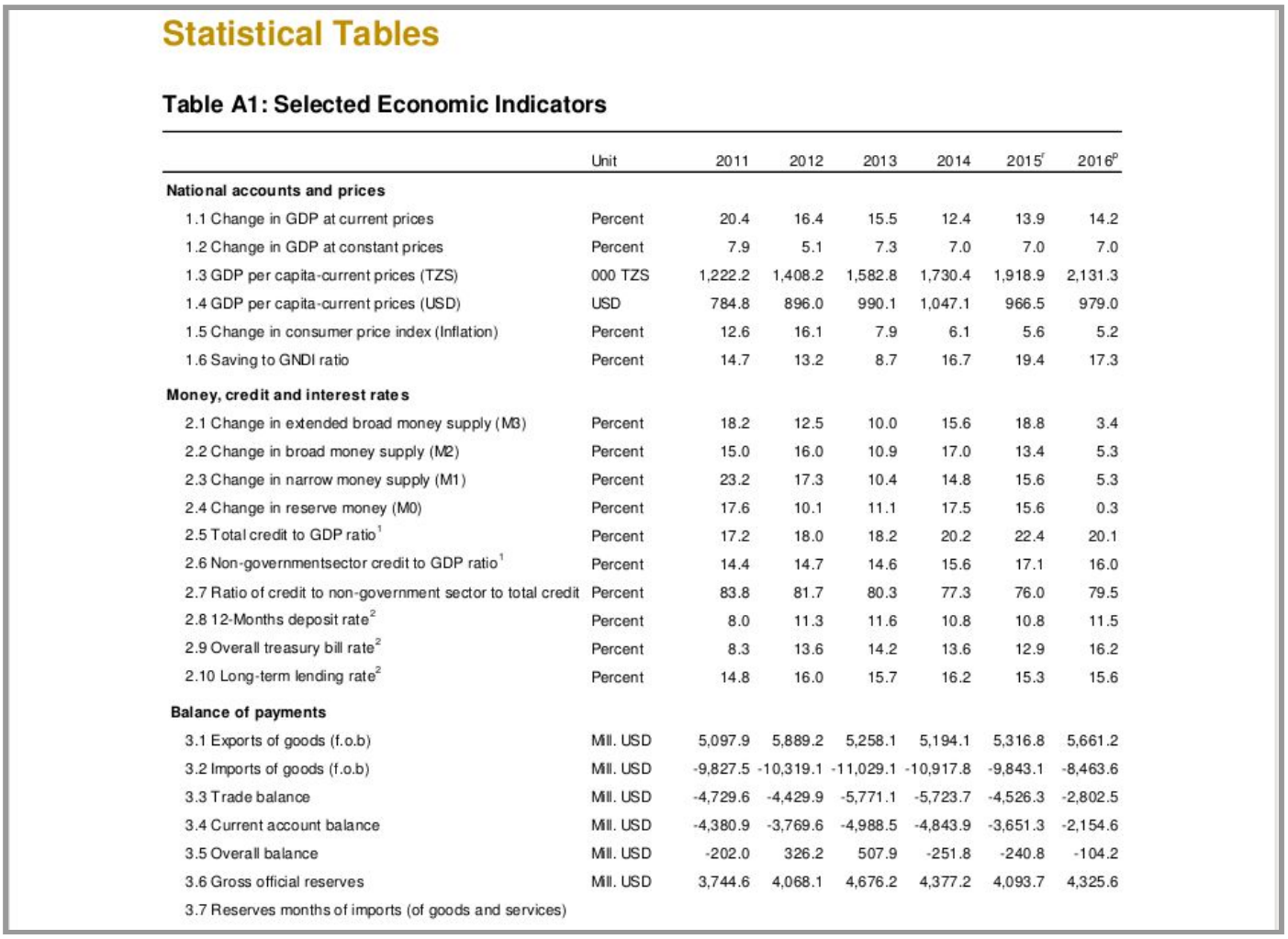
In order to analyse data that you find in a PDF, you will need to convert it into a format which is machine-readable and structured (for instance to XLS format)

Example: Bank of Tanzania Monthly Economic Review Data



The Bank of Tanzania releases a Monthly Economic Report with valuable information on finance such as economic indicators, interests rates and consumer price index. Despite this vast and useful source of data made available on the Bank of Tanzania website, they come in PDF format which makes it difficult for analysis.

Below is an example table with Selected Economic Indicators from 2011 to 2016.



Statistical Tables

Table A1: Selected Economic Indicators

	Unit	2011	2012	2013	2014	2015[r]	2016[p]
National accounts and prices							
1.1 Change in GDP at current prices	Percent	20.4	16.4	15.5	12.4	13.9	14.2
1.2 Change in GDP at constant prices	Percent	7.9	5.1	7.3	7.0	7.0	7.0
1.3 GDP per capita-current prices (TZS)	000 TZS	1,222.2	1,408.2	1,582.8	1,730.4	1,918.9	2,131.3
1.4 GDP per capita-current prices (USD)	USD	784.8	896.0	990.1	1,047.1	966.5	979.0
1.5 Change in consumer price index (Inflation)	Percent	12.6	16.1	7.9	6.1	5.6	5.2
1.6 Saving to GNDI ratio	Percent	14.7	13.2	8.7	16.7	19.4	17.3
Money, credit and interest rates							
2.1 Change in extended broad money supply (M3)	Percent	18.2	12.5	10.0	15.6	18.8	3.4
2.2 Change in broad money supply (M2)	Percent	15.0	16.0	10.9	17.0	13.4	5.3
2.3 Change in narrow money supply (M1)	Percent	23.2	17.3	10.4	14.8	15.6	5.3
2.4 Change in reserve money (M0)	Percent	17.6	10.1	11.1	17.5	15.6	0.3
2.5 Total credit to GDP ratio[1]	Percent	17.2	18.0	18.2	20.2	22.4	20.1
2.6 Non-governmentsector credit to GDP ratio[1]	Percent	14.4	14.7	14.6	15.6	17.1	16.0
2.7 Ratio of credit to non-government sector to total credit	Percent	83.8	81.7	80.3	77.3	76.0	79.5
2.8 12-Months deposit rate[2]	Percent	8.0	11.3	11.6	10.8	10.8	11.5
2.9 Overall treasury bill rate[2]	Percent	8.3	13.6	14.2	13.6	12.9	16.2
2.10 Long-term lending rate[2]	Percent	14.8	16.0	15.7	16.2	15.3	15.6
Balance of payments							
3.1 Exports of goods (f.o.b)	Mil. USD	5,097.9	5,889.2	5,258.1	5,194.1	5,316.8	5,661.2
3.2 Imports of goods (f.o.b)	Mil. USD	-9,827.5	-10,319.1	-11,029.1	-10,917.8	-9,843.1	-8,463.6
3.3 Trade balance	Mil. USD	-4,729.6	-4,429.9	-5,771.1	-5,723.7	-4,526.3	-2,802.5
3.4 Current account balance	Mil. USD	-4,380.9	-3,769.6	-4,988.5	-4,843.9	-3,651.3	-2,154.6
3.5 Overall balance	Mil. USD	-202.0	326.2	507.9	-251.8	-240.8	-104.2
3.6 Gross official reserves	Mil. USD	3,744.6	4,068.1	4,676.2	4,377.2	4,093.7	4,325.6
3.7 Reserves months of imports (of goods and services)							

Task 1: Converting PDF to Excel Using Online Tools

There are several tools that help you convert a data table within a PDF file to the Excel format. Each uses slightly different technology and it is worth it to save yourself time to try more than one and see which results in a cleaner data file. One such tool that's available online is ***Cometdocs***.

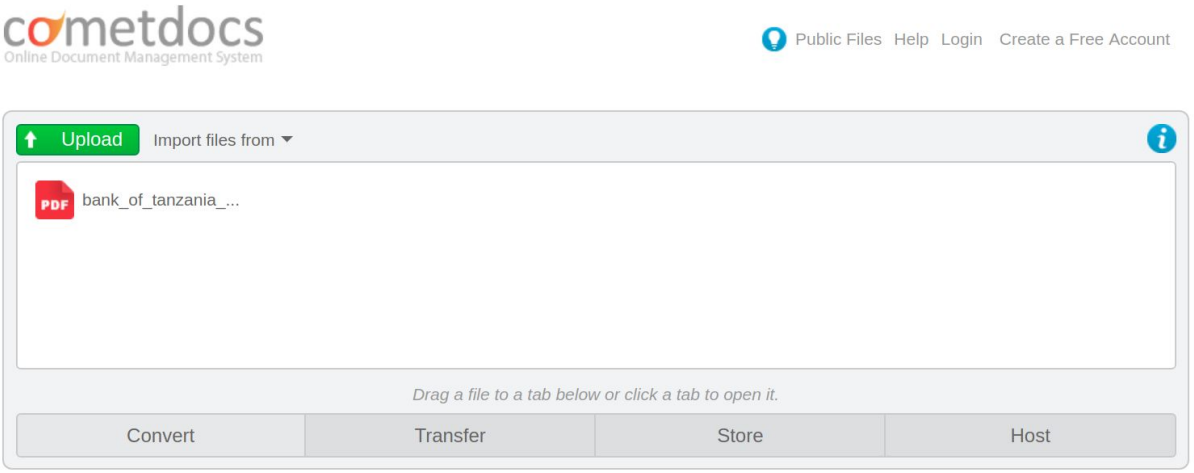
Cometdocs works especially well if your table has a lot of shading in different colors instead of just being a black and white table. You can access it here: www.cometdocs.com

Let's try working with Cometdocs to convert a PDF to Excel. Here are the steps:

1. Open www.cometdocs.com in your web browser

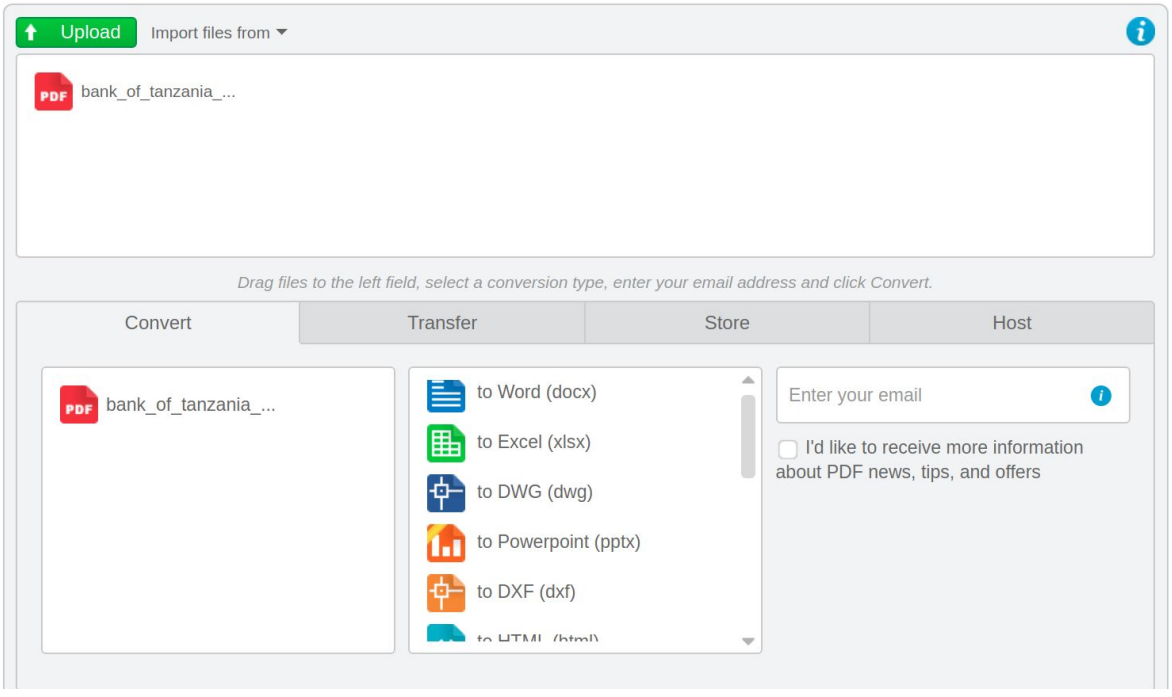
2. Click the **Go to the Web App** button. The screen updates.

3. Click the Upload button to upload the ***bank_of_tanzania_mer_approved_april_2018_pg14*** PDF to Cometdocs. Once uploaded the file displays in the window.



4. Click **Convert**. The screen updates with an empty box under **Convert**.

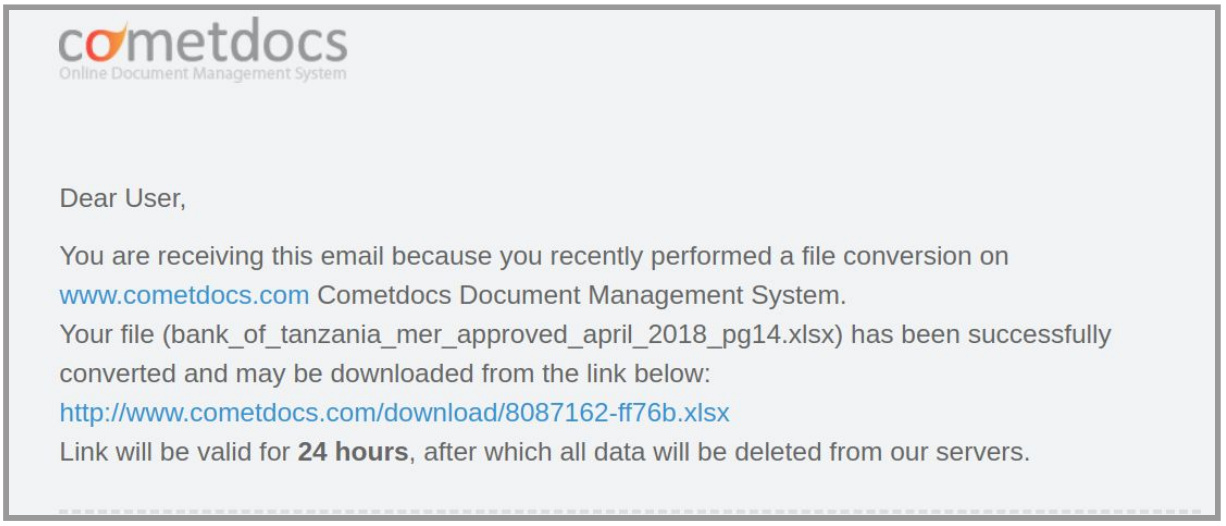
5. Now, click the PDF file icon in the box above, then drag-and-drop this file to the empty box under Convert.



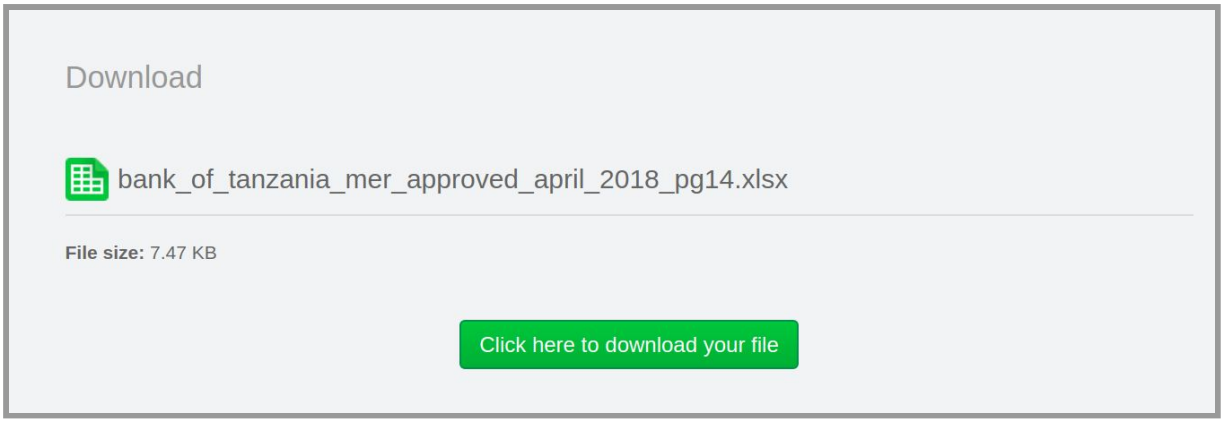
6. The screen updates. Now select the conversion type **to Excel (xlsx)**.

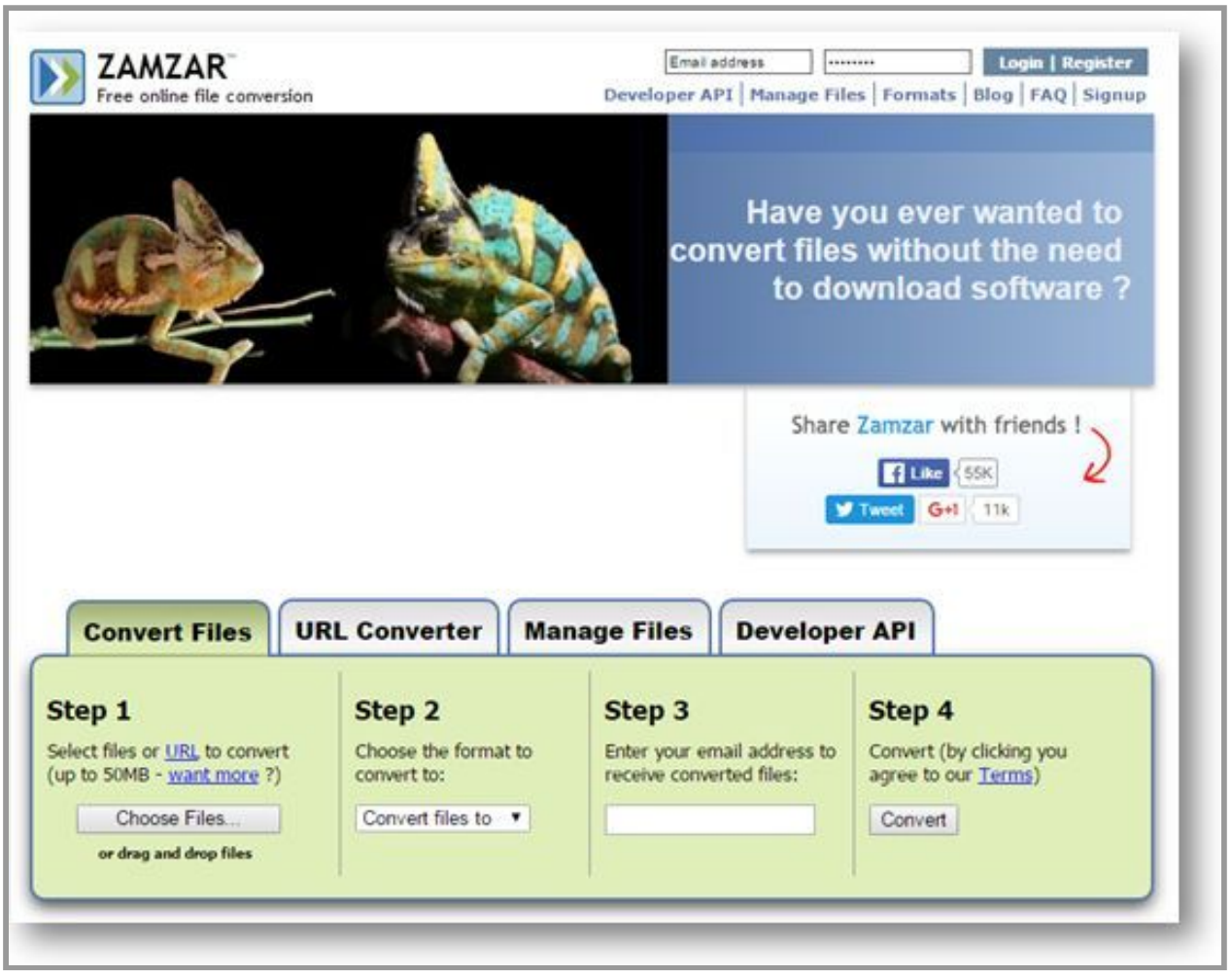
7. Next, enter your email address in the "Enter your email" field. Then, click **Convert**. Cometdocs will now send you a hyperlink to access the converted Excel file.

8. Open your email and click the link provided in the email from Cometdocs.



9. In the browser window that opens, click the link to download the converted Excel file.





Convert PDF to Excel Exercise

Open the following website, and try converting ***bank_of_tanzania_mer_approved_april_2018_pg14*** to Excel:

http://www.zamzar.com/

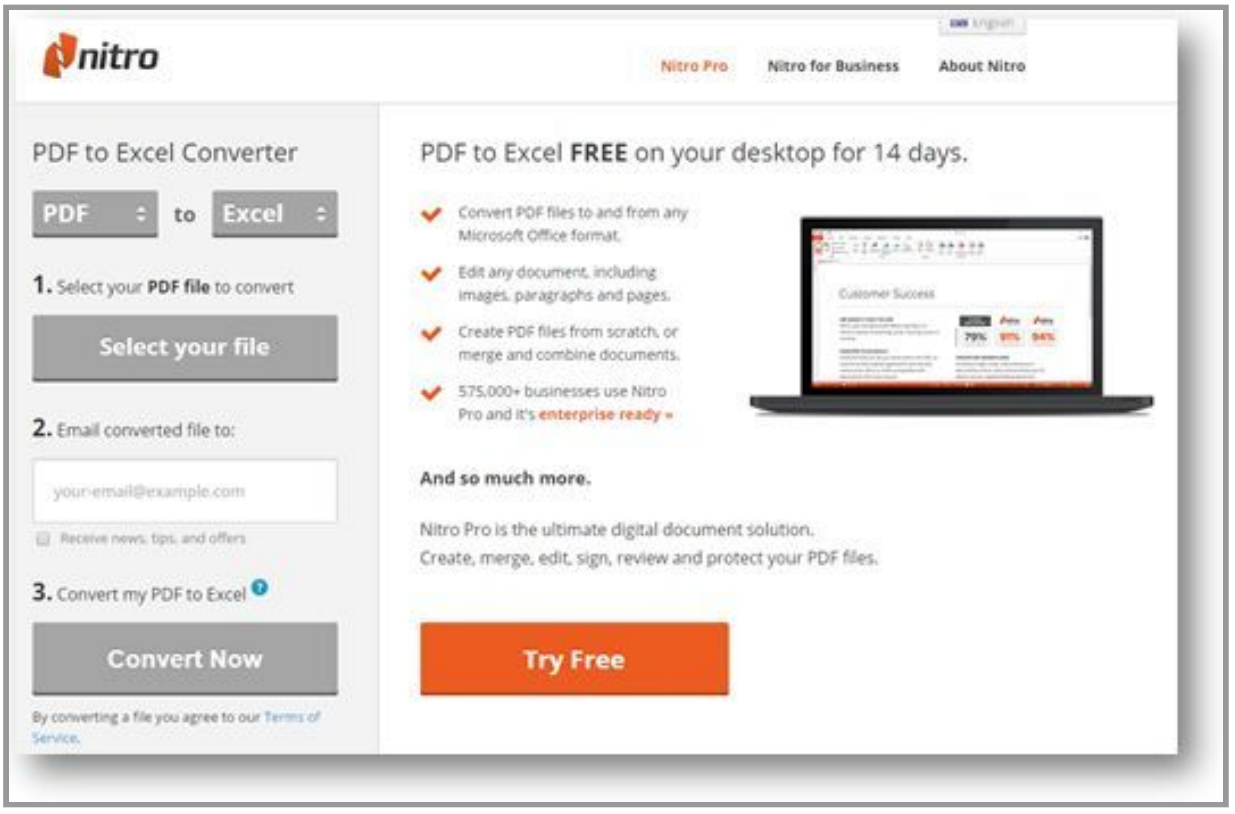
Note: Always check the converted document if the data scraped is as the same of the original document.

In this exercise the Excel file is not clean and will need to be corrected with the “replace” tool.

More Online Tools to Convert PDFs to Excel

Apart from the two online tools that you tried so far, here are two more websites that you can use to convert PDFs to Excel:

PDF to Excel Online



To access, go to: https://www.pdftoexcelonline.com/

PDF to Excel Org

To access, go to: http://www.pdftoexcel.org/

Task 2: Converting PDFs to Excel using Tabula

Tabula is a tool that you can install on your computer to extract data from PDF files. It works well for most PDFs with black and white data tables.

Installing Tabula

Here are the instructions to download and install Tabula:

1. Ensure Java is installed on your computer. You can download Java here: https://www.java.com/en/download/

2. Open the Tabula website: http://tabula.technology/

3. Download the version of Tabula for your operating system:
    - If you use Windows: https://github.com/tabulapdf/tabula/releases/download/v1.2.1/tabula-win-1.2.1.zip
    - If you use Mac: https://github.com/tabulapdf/tabula/releases/download/v1.2.1/tabula-mac-1.2.1.zip
    - If you use Linux / Other operating system: https://github.com/tabulapdf/tabula/releases/download/v1.2.1/tabula-jar-1.2.1.zip

4. Tabula downloads as a zip file on your computer. Extract the downloaded **zip** file – this creates a folder called "tabula" on your computer.
5. Go into the "tabula" folder. Run the **tabula.exe or tabula.dmg** program inside. A control window may open, allow this window to run.

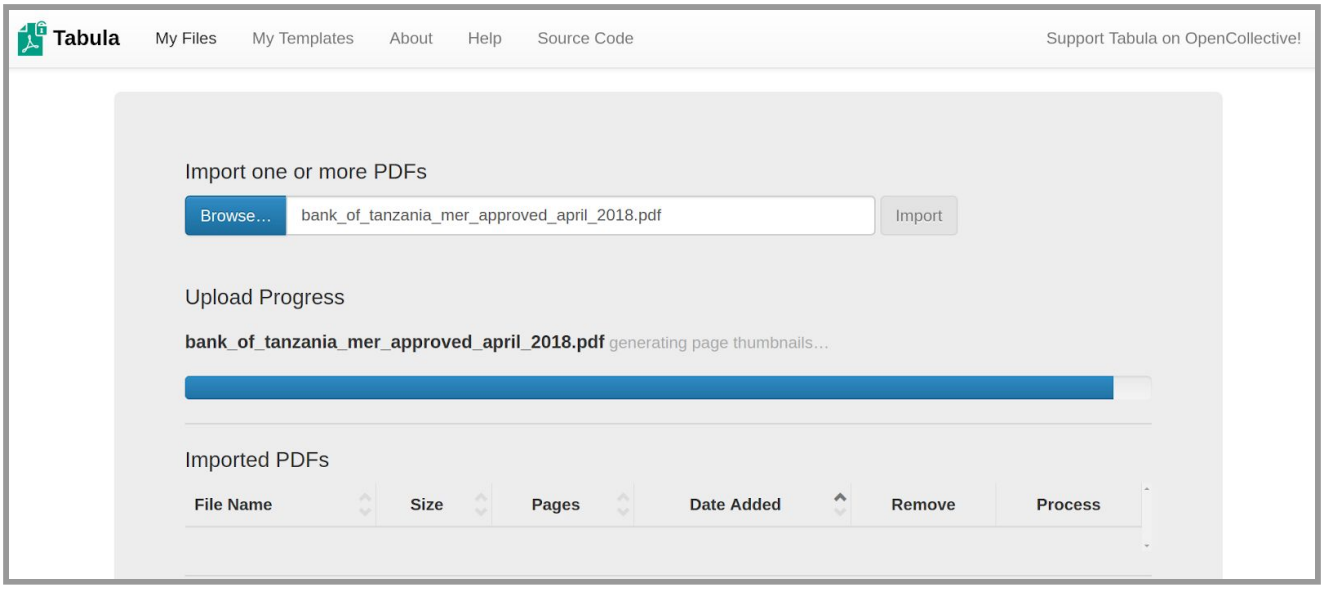
6. Next, a web browser will open – this is Tabula. If your web browser does not open, use your web browser to go to: http://localhost:34555

Using Tabula

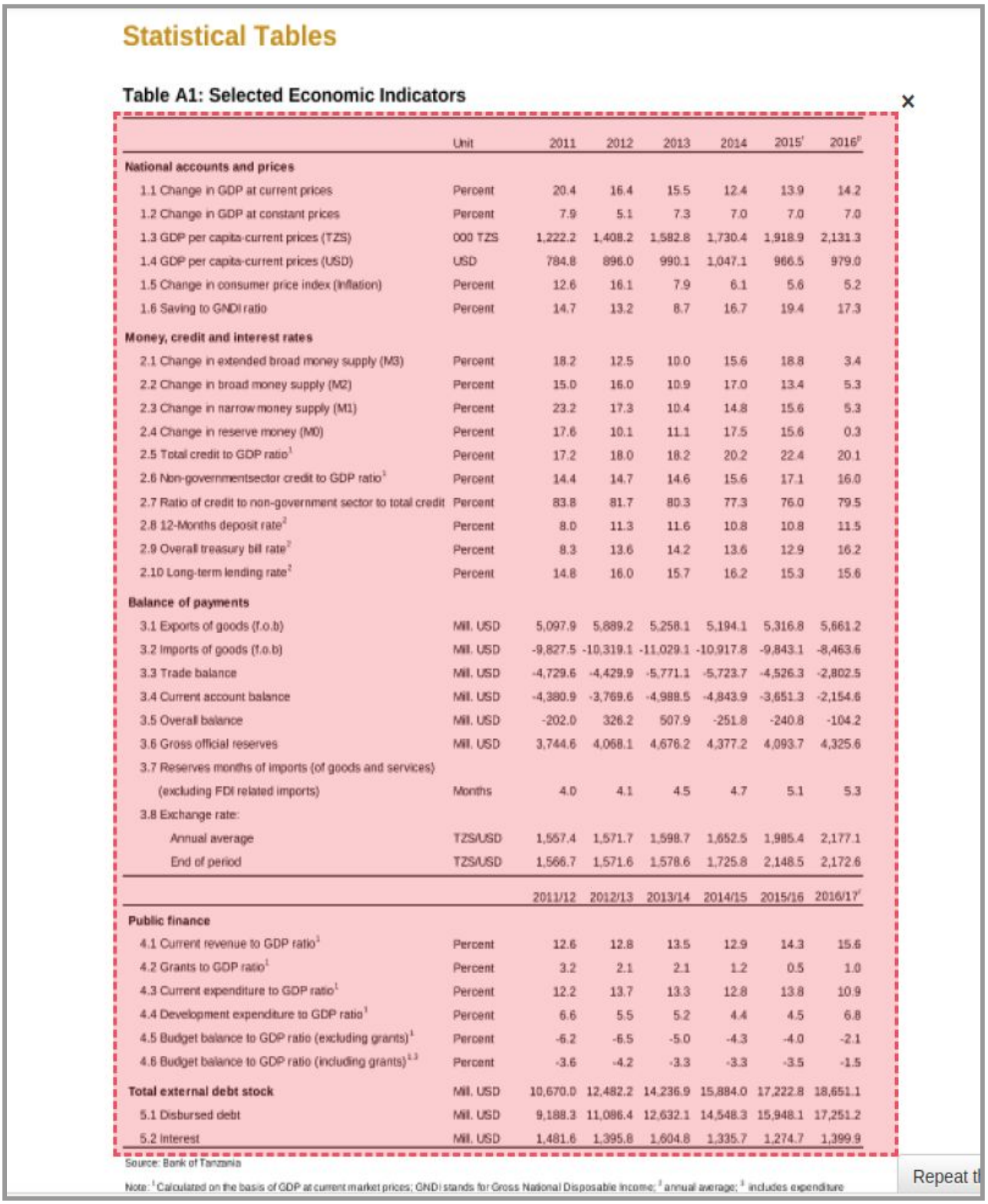
Here is an example, let's use Tabula to extract the data table that's included in a PDF file. The PDF file is called **bank_of_tanzania_mer_approved_april_2018.pdf**

7. Once Tabula is open in your browser window, click the **Browse** button to find and select **bank_of_tanzania_mer_approved_april_2018.pdf** . The file name is displayed in Tabula.

8. Now, click **Import**. Tabula processes the PDF file, and shows a preview of the data table included in the PDF **bank_of_tanzania_mer_approved_april_2018.pdf** within the Tabula window.



9. Go to page 18 and select Table A1: Selected Economic Indicators by clicking the top left corner of the table border and dragging the mouse pointer to the bottom right corner, until all of the data is included in the shaded selection area.



Statistical Tables

Table A1: Selected Economic Indicators

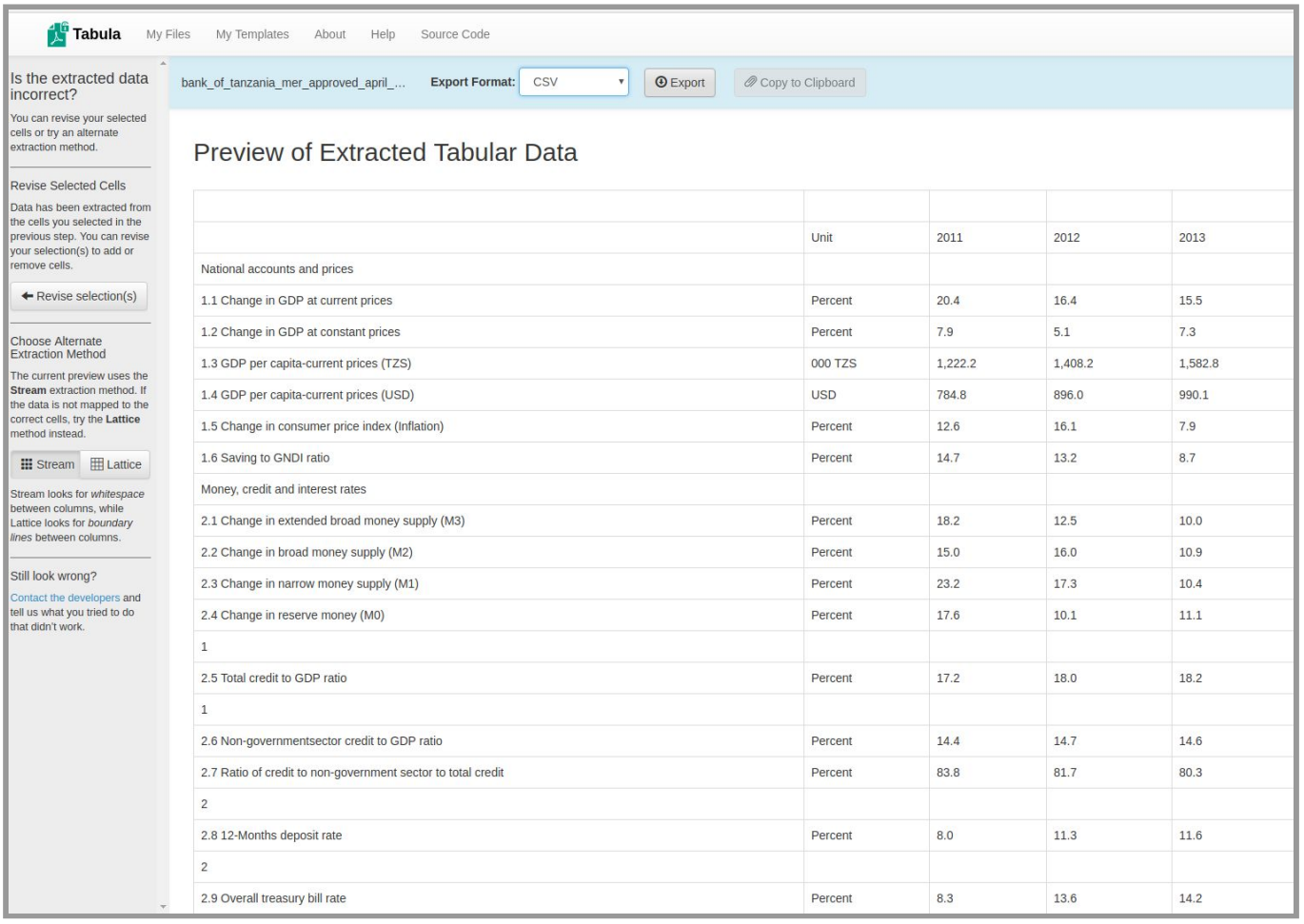
	Unit	2011	2012	2013	2014	2015[r]	2016[p]
National accounts and prices							
1.1 Change in GDP at current prices	Percent	20.4	16.4	15.5	12.4	13.9	14.2
1.2 Change in GDP at constant prices	Percent	7.9	5.1	7.3	7.0	7.0	7.0
1.3 GDP per capita-current prices (TZS)	000 TZS	1,222.2	1,408.2	1,582.8	1,730.4	1,918.9	2,131.3
1.4 GDP per capita-current prices (USD)	USD	784.8	896.0	990.1	1,047.1	966.5	979.0
1.5 Change in consumer price index (Inflation)	Percent	12.6	16.1	7.9	6.1	5.6	5.2
1.6 Saving to GNDI ratio	Percent	14.7	13.2	8.7	16.7	19.4	17.3
Money, credit and interest rates							
2.1 Change in extended broad money supply (M3)	Percent	18.2	12.5	10.0	15.6	18.8	3.4
2.2 Change in broad money supply (M2)	Percent	15.0	16.0	10.9	17.0	13.4	5.3
2.3 Change in narrow money supply (M1)	Percent	23.2	17.3	10.4	14.8	15.6	5.3
2.4 Change in reserve money (M0)	Percent	17.6	10.1	11.1	17.5	15.6	0.3
2.5 Total credit to GDP ratio[1]	Percent	17.2	18.0	18.2	20.2	22.4	20.1
2.6 Non-governmentsector credit to GDP ratio[1]	Percent	14.4	14.7	14.6	15.6	17.1	16.0
2.7 Ratio of credit to non-government sector to total credit	Percent	83.8	81.7	80.3	77.3	76.0	79.5
2.8 12-Months deposit rate[2]	Percent	8.0	11.3	11.6	10.8	10.8	11.5
2.9 Overall treasury bill rate[2]	Percent	8.3	13.6	14.2	13.6	12.9	16.2
2.10 Long-term lending rate[2]	Percent	14.8	16.0	15.7	16.2	15.3	15.6
Balance of payments							
3.1 Exports of goods (f.o.b)	Mil. USD	5,097.9	5,889.2	5,258.1	5,194.1	5,316.8	5,661.2
3.2 Imports of goods (f.o.b)	Mil. USD	-9,827.5	-10,319.1	-11,029.1	-10,917.8	-9,843.1	-8,463.6
3.3 Trade balance	Mil. USD	-4,729.6	-4,429.9	-5,771.1	-5,723.7	-4,526.3	-2,802.5
3.4 Current account balance	Mil. USD	-4,380.9	-3,769.6	-4,988.5	-4,843.9	-3,651.3	-2,154.6
3.5 Overall balance	Mil. USD	-202.0	326.2	507.9	-251.8	-240.8	-104.2
3.6 Gross official reserves	Mil. USD	3,744.6	4,068.1	4,676.2	4,377.2	4,093.7	4,325.6
3.7 Reserves months of imports (of goods and services) (excluding FDI related imports)	Months	4.0	4.1	4.5	4.7	5.1	5.3
3.8 Exchange rate:							
Annual average	TZS/USD	1,557.4	1,571.7	1,598.7	1,652.5	1,985.4	2,177.1
End of period	TZS/USD	1,566.7	1,571.6	1,578.6	1,725.8	2,148.5	2,172.6
		2011/12	2012/13	2013/14	2014/15	2015/16	2016/17[r]
Public finance							
4.1 Current revenue to GDP ratio[1]	Percent	12.6	12.8	13.5	12.9	14.3	15.6
4.2 Grants to GDP ratio[1]	Percent	3.2	2.1	2.1	1.2	0.5	1.0
4.3 Current expenditure to GDP ratio[1]	Percent	12.2	13.7	13.3	12.8	13.8	10.9
4.4 Development expenditure to GDP ratio[1]	Percent	6.6	5.5	5.2	4.4	4.5	6.8
4.5 Budget balance to GDP ratio (excluding grants)[1]	Percent	-6.2	-6.5	-5.0	-4.3	-4.0	-2.1
4.6 Budget balance to GDP ratio (including grants)[1,3]	Percent	-3.6	-4.2	-3.3	-3.3	-3.5	-1.5
Total external debt stock	Mil. USD	10,670.0	12,482.2	14,236.9	15,884.0	17,222.8	18,651.1
5.1 Disbursed debt	Mil. USD	9,188.3	11,086.4	12,632.1	14,548.3	15,948.1	17,251.2
5.2 Interest	Mil. USD	1,481.6	1,395.8	1,604.8	1,335.7	1,274.7	1,399.9

Source: Bank of Tanzania

Note: [1] Calculated on the basis of GDP at current market prices; GNDI stands for Gross National Disposable Income; [2] annual average; [3] includes expenditure

Repeat t

10. Click Preview & Export Extracted Data. A window appears that displays the preview of the extracted data in a structured, machine readable format. Inspect the data to make sure it looks correct. If any data is missing, you may have to slightly expand your selection. Sometimes, if headers are formatted strangely, you have to select the data tables without the headers and type in the column headers manually after.



	Unit	2011	2012	2013
National accounts and prices				
1.1 Change in GDP at current prices	Percent	20.4	16.4	15.5
1.2 Change in GDP at constant prices	Percent	7.9	5.1	7.3
1.3 GDP per capita-current prices (TZS)	000 TZS	1,222.2	1,408.2	1,582.8
1.4 GDP per capita-current prices (USD)	USD	784.8	896.0	990.1
1.5 Change in consumer price index (Inflation)	Percent	12.6	16.1	7.9
1.6 Saving to GNDI ratio	Percent	14.7	13.2	8.7
Money, credit and interest rates				
2.1 Change in extended broad money supply (M3)	Percent	18.2	12.5	10.0
2.2 Change in broad money supply (M2)	Percent	15.0	16.0	10.9
2.3 Change in narrow money supply (M1)	Percent	23.2	17.3	10.4
2.4 Change in reserve money (M0)	Percent	17.6	10.1	11.1
1				
2.5 Total credit to GDP ratio	Percent	17.2	18.0	18.2
1				
2.6 Non-governmentsector credit to GDP ratio	Percent	14.4	14.7	14.6
2.7 Ratio of credit to non-government sector to total credit	Percent	83.8	81.7	80.3
2				
2.8 12-Months deposit rate	Percent	8.0	11.3	11.6
2				
2.9 Overall treasury bill rate	Percent	8.3	13.6	14.2

11. From the Export Format drop-down, you can select a file format to download the extracted data in Choose a file format to work with – including the **CSV** format. Keep the selection as CSV, and click the **Export** button.

12. A CSV file called **tabula-bank_of_tanzania_mer_approved_april_2018.csv** downloads on your computer. Save this file at a suitable location.

13. Now you can work with your data using any spreadsheet software – including Excel, rather than a PDF. To open this file in Excel, first launch Microsoft Excel.

Exercise: Downloading and Extracting a Data Table from PDF

Try the following exercise to practice how to extract a specific data table from a PDF you find online. In this example, you will download the "APRIL 2018 MER APPROVED" report available online in the PDF format, and extract a data table (Table A5: Tanzania Balance of Payments).

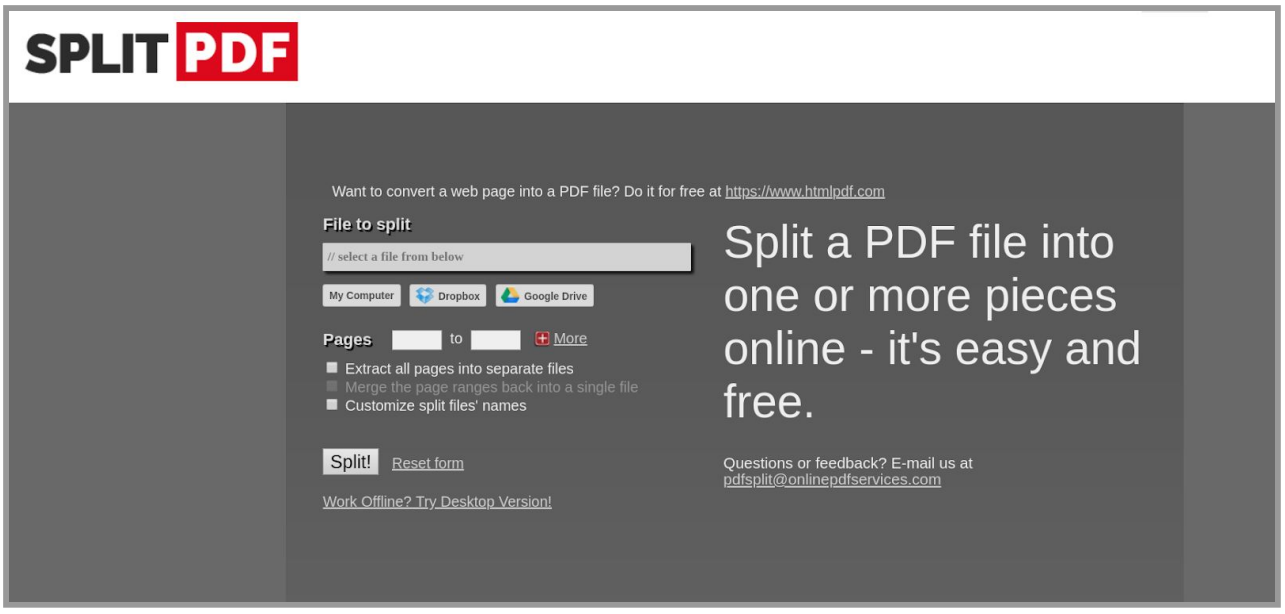
Here are the suggested steps for the process:

Download the PDF

Download the "APRIL 2018 MER APPROVED.pdf":
https://www.bot.go.tz/Publications/MonthlyEconomicReviews/APRIL%202018%20MER_APPROVED.pdf

Extract the required page

1. To extract Table A5 from this large file, go to http://www.splitpdf.com/



2. When the website opens:
    a. Upload the **APRIL 2018 MER APPROVED** to this website from your computer.

    b. Select page **22** to **22**.

    c. Click **Split**! The selected page downloads automatically. Save to your desktop.

Extract Data to Excel Format

3. Navigate to one of these online scraping services:
    a. www.cometdocs.com

    b. www.zamzar.com

4. Upload the PDF file that you saved after extracting page 22, and convert it to the Excel format.

5. Review the converted file to see how complete and clean it is.