Data for Media - Spring 2024

https://schoolofdata.org/extracting-data-from-pdfs/

Lesson 04:
Data Scraping

Extracting Data from PDFs

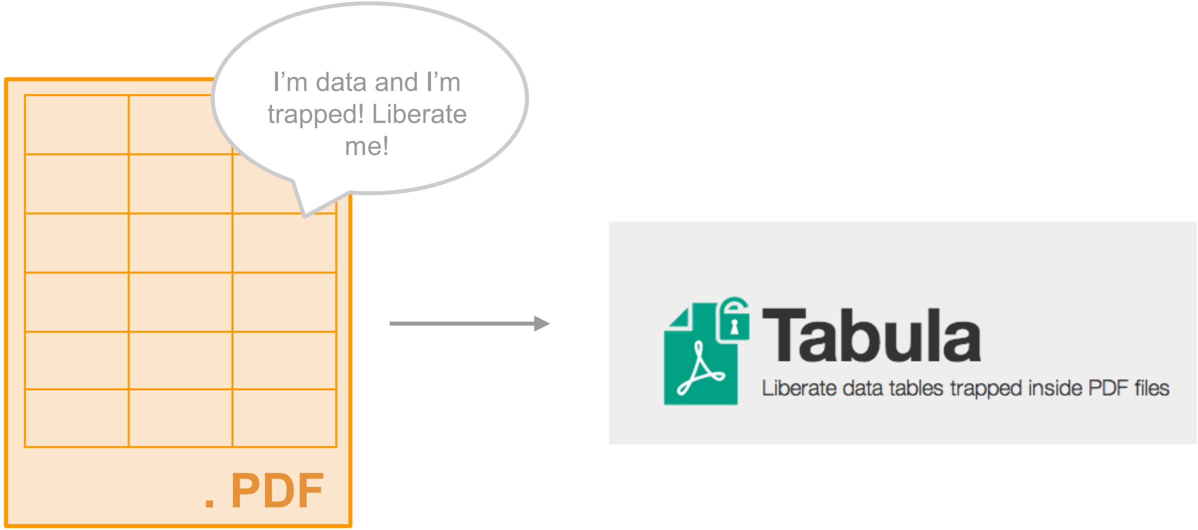
Professor:
Amr Eleraqi

aeleraqi@aucegypt.edu

Unlock it!

Your problem
PDFs are great for humans to read. But not for machines. When researchers want to analyze the data trapped in a PDF, what tools can we recommend?

Your solution
Tabula


I'm data and I'm trapped! Liberate me!
. PDF
Tabula
Liberate data tables trapped inside PDF files

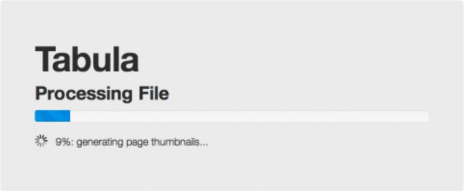
PDF (It's a trap!)

https://tabula.technology/

Unlock it!

Tabula magic:

1. upload file



Tabula

Processing File

9%: generating page thumbnails...

2. select stuff

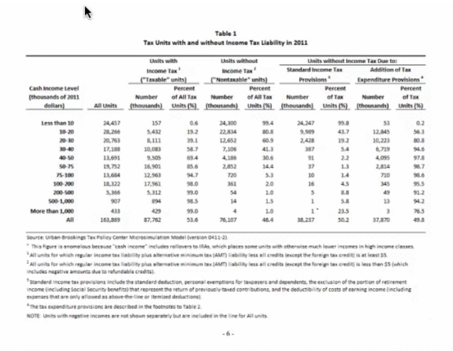


Table 1
Tax Units with and without Income Tax Liability in 2011

		Units with Income Tax [1] ("Taxable" units)		Units without Income Tax [2] ("Nontaxable" units)		Units without Income Tax Due to:			
						Standard Income Tax Provisions [3]		Addition of Tax Expenditure Provisions [4]	
Cash Income Level (thousands of 2011 dollars)	All Units	Number (thousands)	Percent of All Tax Units (%)	Number (thousands)	Percent of All Tax Units (%)	Number (thousands)	Percent of Tax Units (%)	Number (thousands)	Percent of Tax Units (%)
Less than 10	24,457	157	0.6	24,300	99.4	24,247	99.8	53	0.2
10-20	28,266	5,432	19.2	22,834	80.8	9,989	43.7	12,845	56.3
20-30	20,763	8,111	39.1	12,652	60.9	2,428	19.2	10,223	80.8
30-40	17,188	10,083	58.7	7,106	41.3	387	5.4	6,719	94.6
40-50	13,691	9,505	69.4	4,186	30.6	91	2.2	4,095	97.8
50-75	19,752	16,901	85.6	2,852	14.4	37	1.3	2,814	98.7
75-100	13,684	12,963	94.7	720	5.3	10	1.4	710	98.6
100-200	18,322	17,961	98.0	361	2.0	16	4.5	345	95.5
200-500	5,366	5,312	99.0	54	1.0	5	8.8	49	91.2
500-1,000	907	894	98.5	14	1.5	1	5.8	13	94.2
More than 1,000	433	429	99.0	4	1.0	1 *	23.5	3	76.5
All	163,869	87,762	53.6	76,107	46.4	38,237	50.2	37,870	49.8

Source: Urban-Brookings Tax Policy Center Microsimulation Model (version 0411-2).

* This figure is anomalous because "cash income" includes rollovers to IRAs, which places some units with otherwise much lower incomes in high income classes.

[1] All units for which regular income tax liability plus alternative minimum tax (AMT) liability less all credits (except the foreign tax credit) is at least $5.

[2] All units for which regular income tax liability plus alternative minimum tax (AMT) liability less all credits (except the foreign tax credit) is less than $5 (which includes negative amounts due to refundable credits).

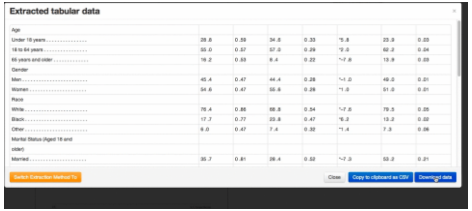
[3] Standard income tax provisions include the standard deduction, personal exemptions for taxpayers and dependents, the exclusion of the portion of retirement income (including Social Security benefits) that represent the return of previously-taxed contributions, and the deductibility of costs of earning income (including expenses that are only allowed as above-the-line or itemized deductions).

[4] The tax expenditure provisions are described in the footnotes to Table 2.

NOTE: Units with negative incomes are not shown separately but are included in the line for All units.

- 6 -

3. download data!



Extracted tabular data

Age							
Under 18 years	28 .6	0 .59	34 .6	0 .33	*5 .8	23 .9	0 .03
18 to 64 years	55 .0	0 .57	57 .0	0 .29	*2 .0	62 .2	0 .04
65 years and older	16 .2	0 .53	8 .4	0 .22	*–7 .8	13 .9	0 .03
Gender							
Men	45 .4	0 .47	44 .4	0 .28	*–1 .0	49 .0	0 .01
Women	54 .6	0 .47	55 .6	0 .28	*1 .0	51 .0	0 .01
Race							
White	76 .4	0 .86	68 .8	0 .54	*–7 .6	79 .5	0 .05
Black	17 .7	0 .77	23 .8	0 .47	*6 .2	13 .2	0 .02
Other	6 .0	0 .47	7 .4	0 .32	*1 .4	7 .3	0 .06
Marital Status (Aged 18 and							
older)							
Married	35 .7	0 .81	28 .4	0 .52	*–7 .3	53 .2	0 .21

About Tabula

- The program is available in different versions that are compatible with various operating systems.
- It requires an internet connection.
- No one can view the files and documents that are loaded inside the program, nor does the program keep a copy of them.
- The program works well with both structured and unstructured data.
- The program handles data effectively in Arabic language.

Pros

- Free, open source.
- Easy to install and use.
- Pretty accurate, especially for PDFs with simple table layout.
- Also has options for Python or R.

Cons

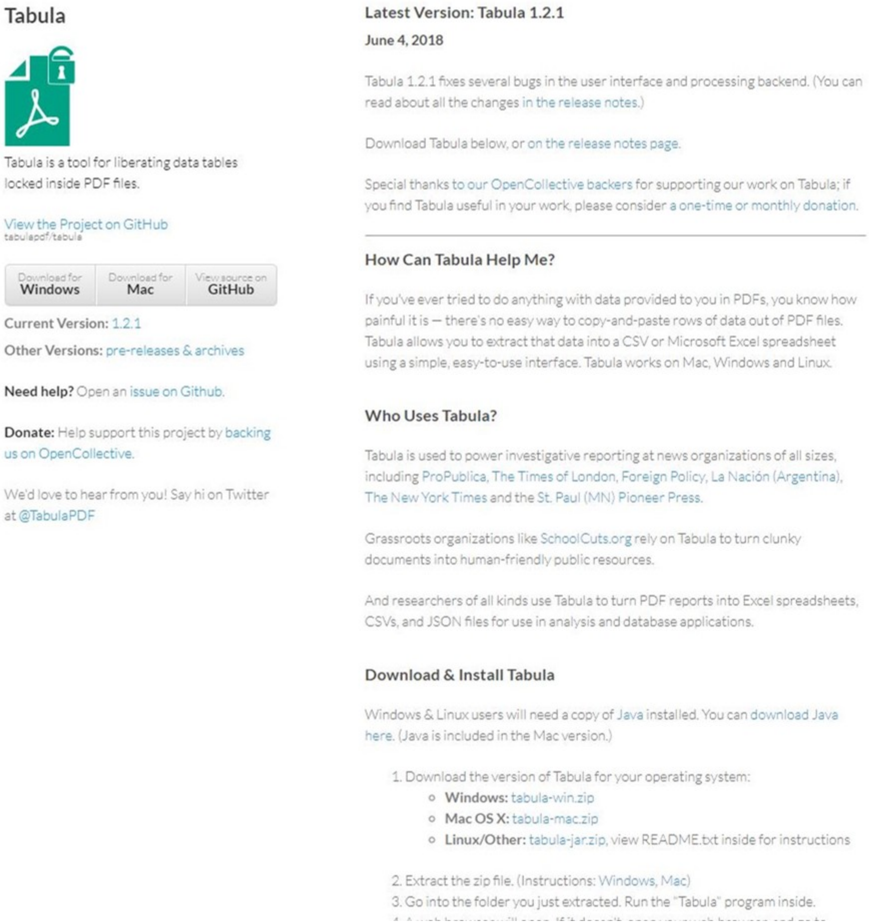
- No OCR, so that must be done separately.
- Had some issues with complex table layout.
- Minimal documentation.
- Not in active development.

Installing Tabula

Here are the instructions to download and install Tabula:

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

- Open the Tabula website: http://tabula.technology/
- Download the version of Tabula for your operating system



Tabula

Tabula is a tool for liberating data tables locked inside PDF files.

View the Project on GitHub
tabulapdf/tabula

Current Version: 1.2.1

Other Versions: pre-releases & archives

Need help? Open an issue on Github.

Donate: Help support this project by backing us on OpenCollective.

We'd love to hear from you! Say hi on Twitter at @TabulaPDF

Latest Version: Tabula 1.2.1

June 4, 2018

Tabula 1.2.1 fixes several bugs in the user interface and processing backend. (You can read about all the changes in the release notes.)

Download Tabula below, or on the release notes page.

Special thanks to our OpenCollective backers for supporting our work on Tabula; if you find Tabula useful in your work, please consider a one-time or monthly donation.

How Can Tabula Help Me?

If you've ever tried to do anything with data provided to you in PDFs, you know how painful it is — there's no easy way to copy-and-paste rows of data out of PDF files. Tabula allows you to extract that data into a CSV or Microsoft Excel spreadsheet using a simple, easy-to-use interface. Tabula works on Mac, Windows and Linux.

Who Uses Tabula?

Tabula is used to power investigative reporting at news organizations of all sizes, including ProPublica, The Times of London, Foreign Policy, La Nación (Argentina), The New York Times and the St. Paul (MN) Pioneer Press.

Grassroots organizations like SchoolCuts.org rely on Tabula to turn clunky documents into human-friendly public resources.

And researchers of all kinds use Tabula to turn PDF reports into Excel spreadsheets, CSVs, and JSON files for use in analysis and database applications.

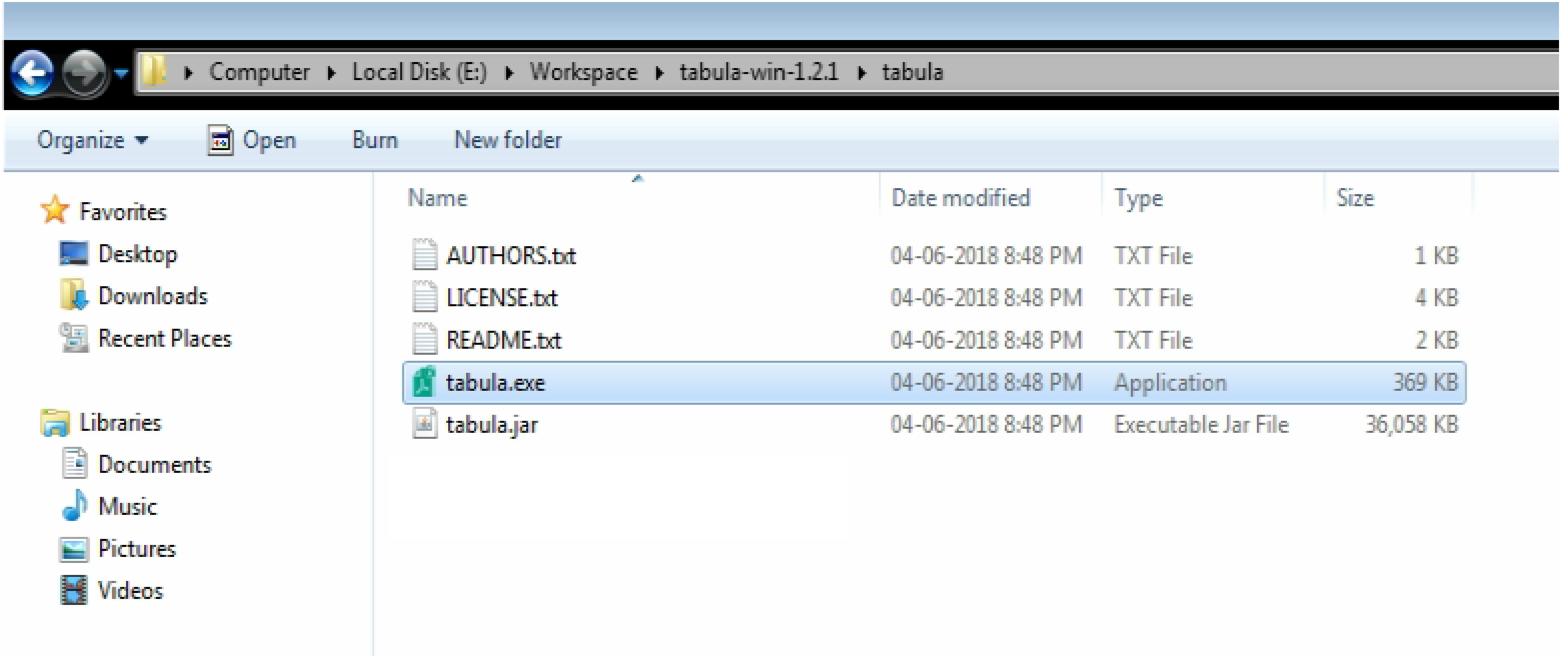
Download & Install Tabula

Windows & Linux users will need a copy of Java installed. You can download Java here. (Java is included in the Mac version.)

1. Download the version of Tabula for your operating system:
   - **Windows:** tabula-win.zip
   - **Mac OS X:** tabula-mac.zip
   - **Linux/Other:** tabula-jar.zip, view README.txt inside for instructions
2. Extract the zip file. (Instructions: Windows, Mac)
3. Go into the folder you just extracted. Run the "Tabula" program inside.

Installing Tabula

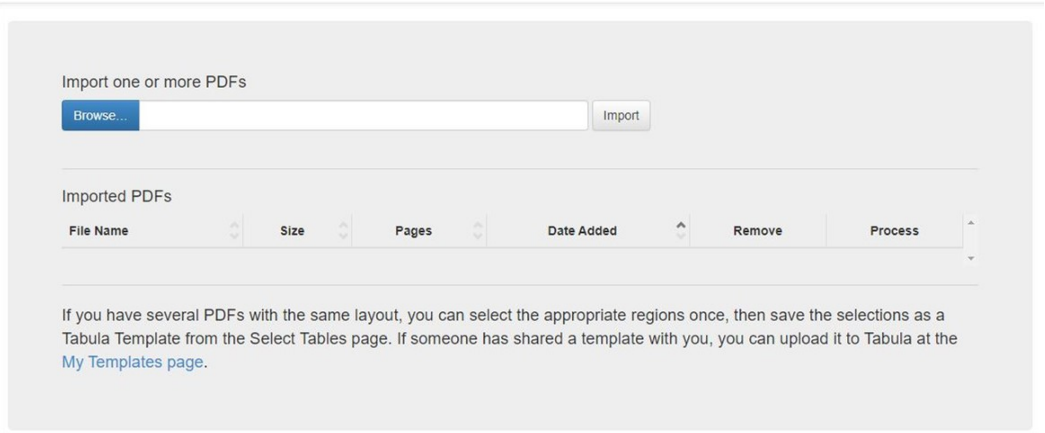
- Tabula downloads as a zip file on your computer. Extract the downloaded zip file – this creates a folder called “tabula” on your computer.
- Go into the “tabula” folder. Run the tabula.exe program inside. A control window may open; allow this window to run.
- Next, a web browser will open – this is Tabula. If your web browser does not open, use your web browser to go to: http://localhost:34555



Name	Date modified	Type	Size
AUTHORS.txt	04-06-2018 8:48 PM	TXT File	1 KB
LICENSE.txt	04-06-2018 8:48 PM	TXT File	4 KB
README.txt	04-06-2018 8:48 PM	TXT File	2 KB
tabula.exe	04-06-2018 8:48 PM	Application	369 KB
tabula.jar	04-06-2018 8:48 PM	Executable Jar File	36,058 KB

Installing Tabula

- After running the program, 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
- There’s Tabula!



Import one or more PDFs

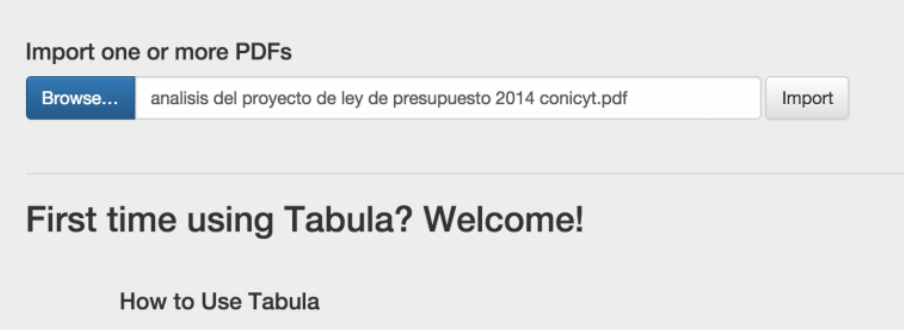
Imported PDFs

File Name	Size	Pages	Date Added	Remove	Process

If you have several PDFs with the same layout, you can select the appropriate regions once, then save the selections as a Tabula Template from the Select Tables page. If someone has shared a template with you, you can upload it to Tabula at the My Templates page.

Tabula in Action

1. Upload PDF File: In home screen you will find file selection option where you need to browse and upload PDF file from which you want to extract data. After selecting the file, click on the Import button. After submission, you will be shown uploaded PDF file as shown in the screenshots below:

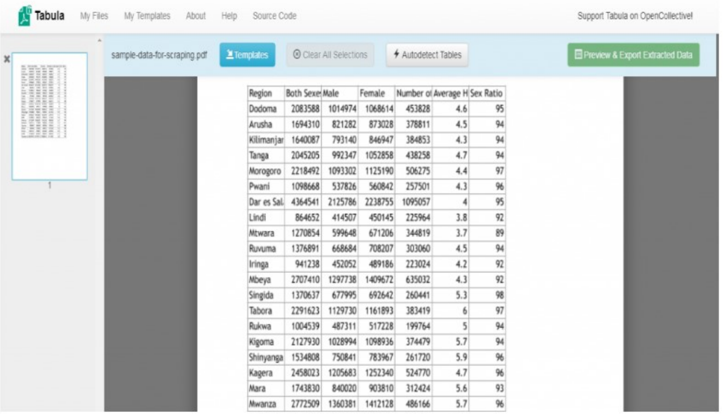


Import one or more PDFs

analisis del proyecto de ley de presupuesto 2014 conicyt.pdf

First time using Tabula? Welcome!

How to Use Tabula



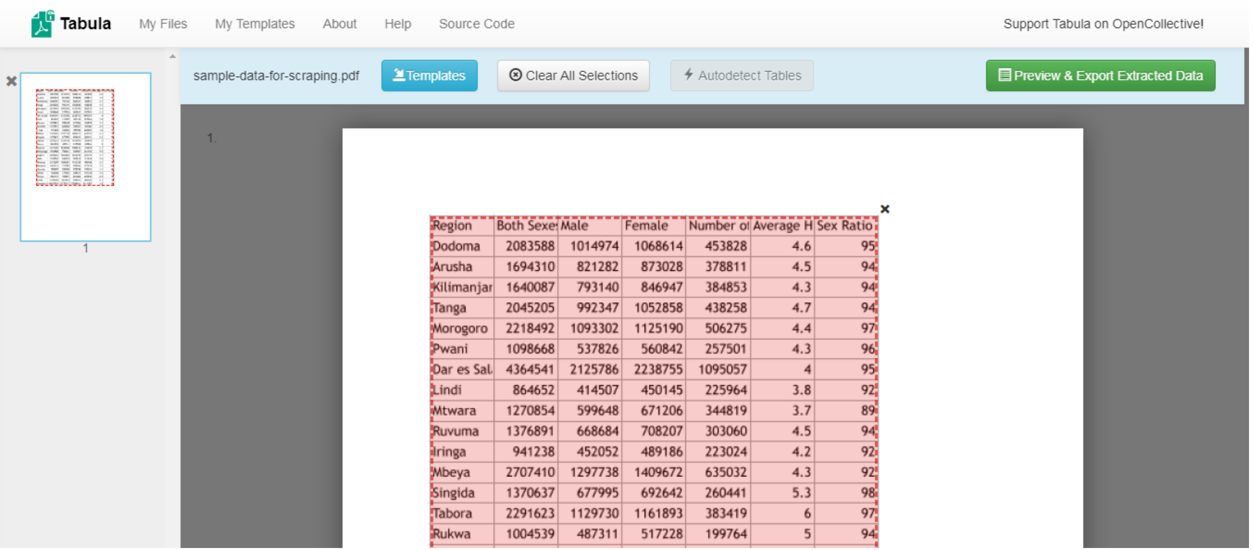
sample-data-for-scraping.pdf

Region	Both Sexe	Male	Female	Number ol	Average H	Sex Ratio
Dodoma	2083588	1014974	1068614	453828	4.6	95
Arusha	1694310	821282	873028	378811	4.5	94
Kilimanjar	1640087	793140	846947	384853	4.3	94
Tanga	2045205	992347	1052858	438258	4.7	94
Morogoro	2218492	1093302	1125190	506275	4.4	97
Pwani	1098668	537826	560842	257501	4.3	96
Dar es Sal	4364541	2125786	2238755	1095057	4	95
Lindi	864652	414507	450145	225964	3.8	92
Mtwara	1270854	599648	671206	344819	3.7	89
Ruvuma	1376891	668684	708207	303060	4.5	94
Iringa	941238	452052	489186	223024	4.2	92
Mbeya	2707410	1297738	1409672	635032	4.3	92
Singida	1370637	677995	692642	260441	5.3	98
Tabora	2291623	1129730	1161893	383419	6	97
Rukwa	1004539	487311	517228	199764	5	94
Kigoma	2127930	1028994	1098936	374479	5.7	94
Shinyanga	1534808	750841	783967	261720	5.9	96
Kagera	2458023	1205683	1252340	524770	4.7	96
Mara	1743830	840020	903810	312424	5.6	93
Mwanza	2772509	1360381	1412128	486166	5.7	96

Depending on the size of the PDF, it may take some time for the file to upload.
Be patient!

Tabula in Action

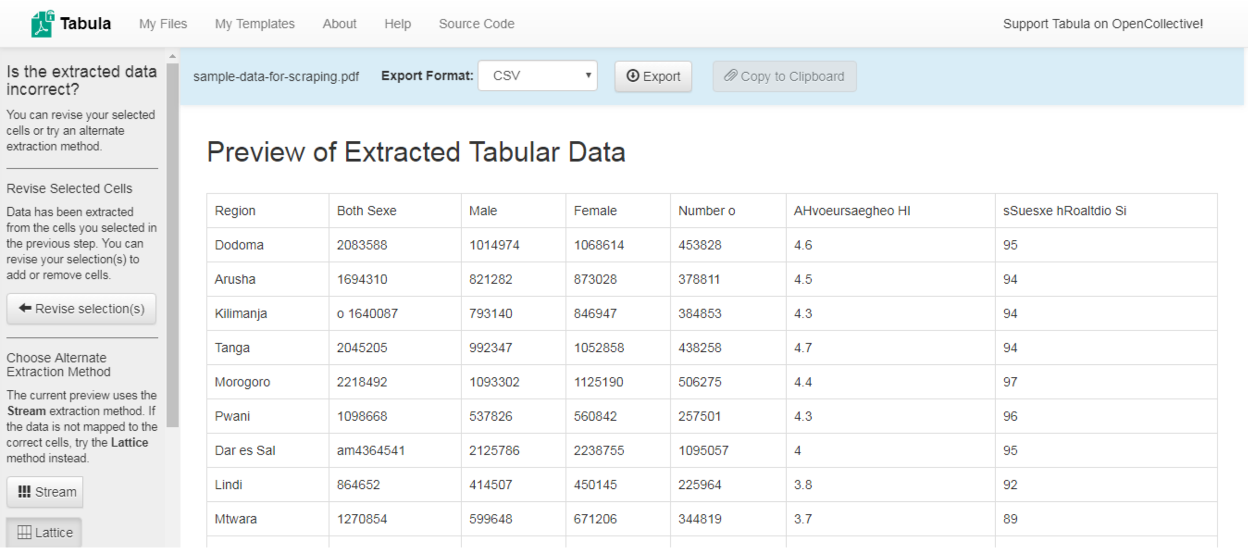
2. Make table selection: Now you need to move to the page from where you want to extract table data, then select the table by clicking and dragging to draw a box around the table. *You can also click on “Autodetect Tables” option which will select the tabular data automatically.*



Region	Both Sexe	Male	Female	Number o	Average H	Sex Ratio
Dodoma	2083588	1014974	1068614	453828	4.6	95
Arusha	1694310	821282	873028	378811	4.5	94
Kilimanjar	1640087	793140	846947	384853	4.3	94
Tanga	2045205	992347	1052858	438258	4.7	94
Morogoro	2218492	1093302	1125190	506275	4.4	97
Pwani	1098668	537826	560842	257501	4.3	96
Dar es Sal	4364541	2125786	2238755	1095057	4	95
Lindi	864652	414507	450145	225964	3.8	92
Mtwara	1270854	599648	671206	344819	3.7	89
Ruvuma	1376891	668684	708207	303060	4.5	94
Iringa	941238	452052	489186	223024	4.2	92
Mbeya	2707410	1297738	1409672	635032	4.3	92
Singida	1370637	677995	692642	260441	5.3	98
Tabora	2291623	1129730	1161893	383419	6	97
Rukwa	1004539	487311	517228	199764	5	94

Tabula in Action

3. Preview Data: After table selection “Preview & Export Extracted Data” button will be enabled. Click on that button to preview the data. Preview is shown in the below screenshot:



sample-data-for-scraping.pdf **Export Format:** CSV

Is the extracted data incorrect?

You can revise your selected cells or try an alternate extraction method.

Revise Selected Cells

Data has been extracted from the cells you selected in the previous step. You can revise your selection(s) to add or remove cells.

Choose Alternate Extraction Method

The current preview uses the **Stream** extraction method. If the data is not mapped to the correct cells, try the **Lattice** method instead.

Preview of Extracted Tabular Data

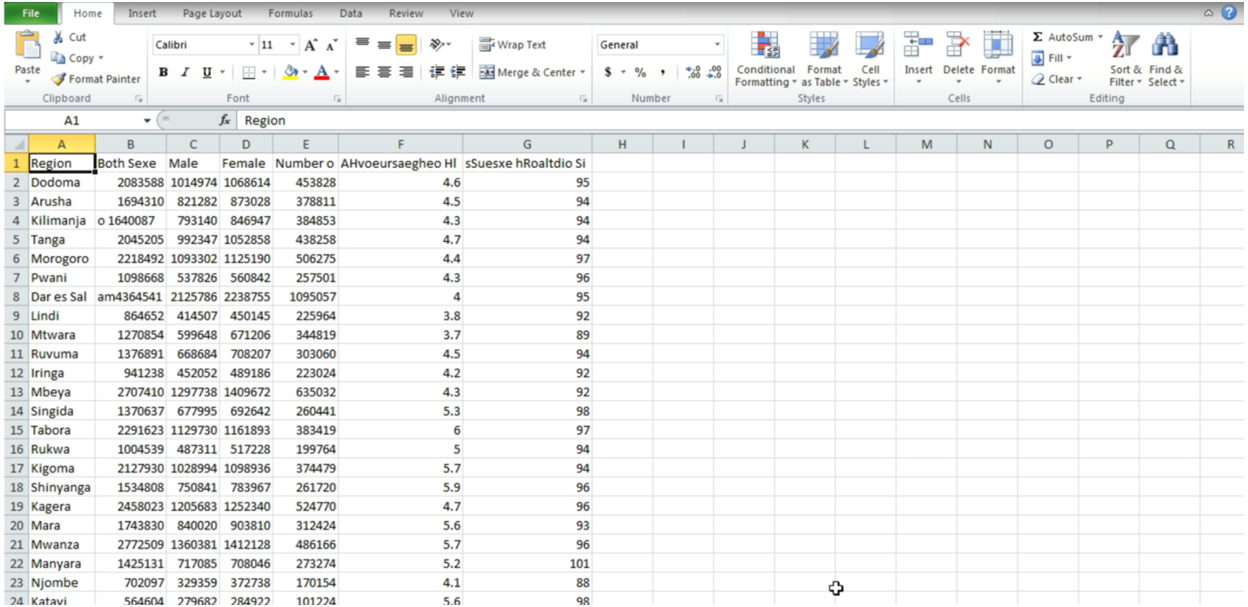
Region	Both Sexe	Male	Female	Number o	AHvoeursaegheo HI	sSuesxe hRoaltdio Si
Dodoma	2083588	1014974	1068614	453828	4.6	95
Arusha	1694310	821282	873028	378811	4.5	94
Kilimanja	o 1640087	793140	846947	384853	4.3	94
Tanga	2045205	992347	1052858	438258	4.7	94
Morogoro	2218492	1093302	1125190	506275	4.4	97
Pwani	1098668	537826	560842	257501	4.3	96
Dar es Sal	am4364541	2125786	2238755	1095057	4	95
Lindi	864652	414507	450145	225964	3.8	92
Mtwara	1270854	599648	671206	344819	3.7	89

Double-check your data by cross-referencing your table

Double-check your Tabula preview of your table with the original PDF. We use another program, like Preview or Adobe Acrobat, to compare. This way, you’ll make sure no data has been lost or misread.

Tabula in Action

4. Export Data: Now you can copy data to the clipboard and paste it to anywhere you want or you can export the data to a variety of file formats like CSV, TSV, JSON etc. The following screenshot shows exported CSV file:



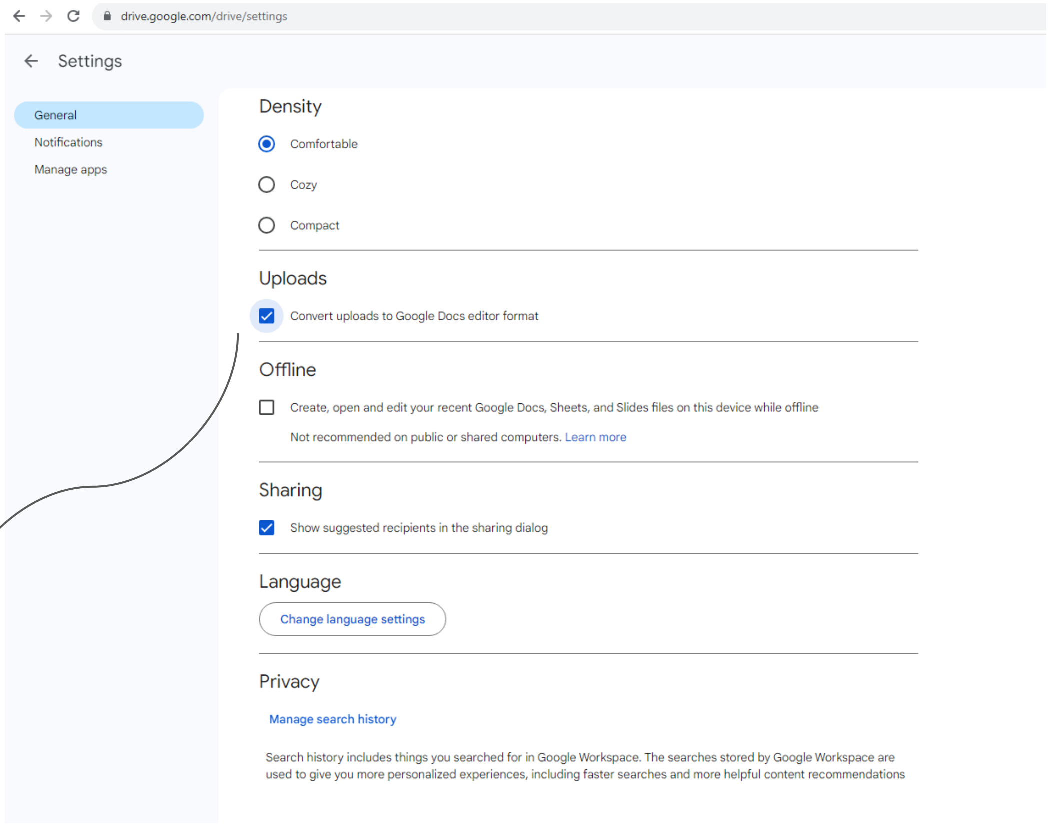
	A	B	C	D	E	F	G
1	Region	Both Sexe	Male	Female	Number o	AHvoeursaegheo Hl	sSuesxe hRoaltdio Si
2	Dodoma	2083588	1014974	1068614	453828	4.6	95
3	Arusha	1694310	821282	873028	378811	4.5	94
4	Kilimanja	o 1640087	793140	846947	384853	4.3	94
5	Tanga	2045205	992347	1052858	438258	4.7	94
6	Morogoro	2218492	1093302	1125190	506275	4.4	97
7	Pwani	1098668	537826	560842	257501	4.3	96
8	Dar es Sal	am4364541	2125786	2238755	1095057	4	95
9	Lindi	864652	414507	450145	225964	3.8	92
10	Mtwara	1270854	599648	671206	344819	3.7	89
11	Ruvuma	1376891	668684	708207	303060	4.5	94
12	Iringa	941238	452052	489186	223024	4.2	92
13	Mbeya	2707410	1297738	1409672	635032	4.3	92
14	Singida	1370637	677995	692642	260441	5.3	98
15	Tabora	2291623	1129730	1161893	383419	6	97
16	Rukwa	1004539	487311	517228	199764	5	94
17	Kigoma	2127930	1028994	1098936	374479	5.7	94
18	Shinyanga	1534808	750841	783967	261720	5.9	96
19	Kagera	2458023	1205683	1252340	524770	4.7	96
20	Mara	1743830	840020	903810	312424	5.6	93
21	Mwanza	2772509	1360381	1412128	486166	5.7	96
22	Manyara	1425131	717085	708046	273274	5.2	101
23	Njombe	702097	329359	372738	170154	4.1	88
24	Katavi	564604	279682	284922	101224	5.6	98

Extract text from image in Google Drive

Google Docs has a powerful Optical Character Recognition feature built right in..

Enabling the feature

Open up Google Drive and then click on the gear icon. From the drop-down, click Settings. In the resulting window, make sure Convert uploads is checked.



drive.google.com/drive/settings

Settings

General

Notifications

Manage apps

Density

● Comfortable

○ Cozy

○ Compact

Uploads

☑ Convert uploads to Google Docs editor format

Offline

□ Create, open and edit your recent Google Docs, Sheets, and Slides files on this device while offline

Not recommended on public or shared computers. Learn more

Sharing

☑ Show suggested recipients in the sharing dialog

Language

Change language settings

Privacy

Manage search history

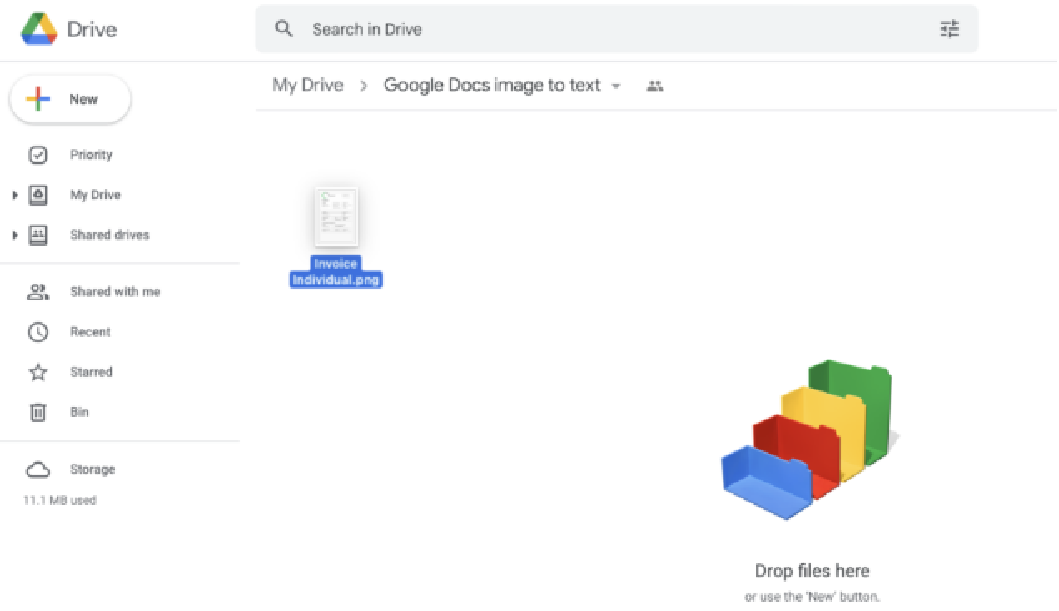
Search history includes things you searched for in Google Workspace. The searches stored by Google Workspace are used to give you more personalized experiences, including faster searches and more helpful content recommendations

Extract text from image in Google Drive

Using the feature

Upload either an image to Google Drive. The uploaded image doesn’t automatically convert. In fact, it will remain exactly as uploaded.

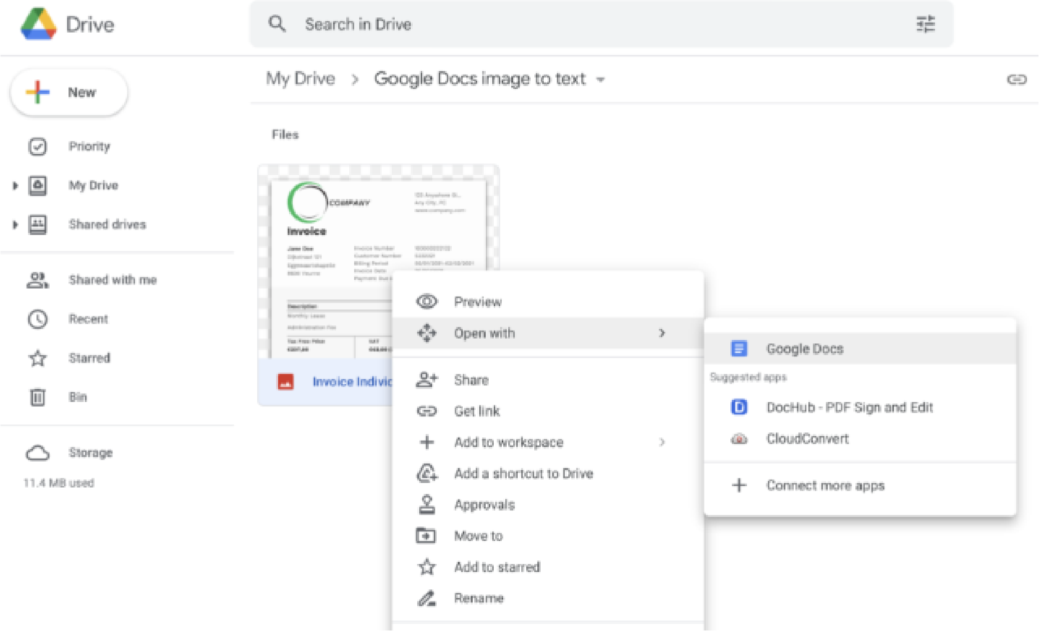
It is important to remember that Google Drive supports documents with .jpg, .png, .gif extensions and PDF files with a maximum size of 2 MB.



Extract text from image in Google Drive

Using the feature

Once the file is in your Google Drive account, right-click it and select Open with | Google Docs.



After selecting to open your file with Google Docs, wait for a bit and a new document will open, containing the newly converted text.