

Tabula Guide

Tabula website: <https://tabula.technology/>

Benefits: Tool that can be used to select specific text and data from pdf and extract it manually into files such as CSV and TSV

Downloading/Setting up:

1. Tabula website: <https://tabula.technology/>
2. In this website, scroll down and download the software package for the computer operating system you currently use

Starting Tabula/Setting up PDF:

1. Save pdf file locally (to your laptop)
 - a. Make sure file is saved as PDF
2. Open tabula → search for Tabula on your laptop (click on the actual tabula software icon when the tabula folder opens on your screen)
3. Will open a browser page with Tabula software loaded onto it
4. Under “MyFiles” tab click on the blue “browse” button to select a file from your computer
 - a. Find the file that you are looking for
5. Wait for PDF file to be uploaded fully into the program (may take a couple of minutes)
6. Once, ready click on the green “Extract Data” button

Extracting data:

1. Using your mouse highlights certain text on each page that you would like to extract text/data from.
 - a. This includes graphs and bar graphs that have data labels
 - b. Can also use the “autodetect tables” feature however this sometimes has more errors than manually selecting them. If using this feature, I recommend double checking table selection after using the feature.
2. Once done highlighting, click on the green “Preview & Export Extracted Data” button
 - a. This will take you to another page with options to export as CSV (excel), TSV(text) files. If you click on “Export” it will download a file onto your laptop
3. You may have to do some editing/reorganizing of the data in your file after selection to get it into the format you want it in.
4. Can also create templates for pdfs with similar text structure and data tables

Tabula Help: <http://127.0.0.1:8080/help>

- Optional: <https://github.com/tabulapdf/tabula-java>
 - Has a sample code that you could use to test out Java code to capture all tables and data within tables (different than using only this software, using only Tabula may be easier just a bit more time consuming)