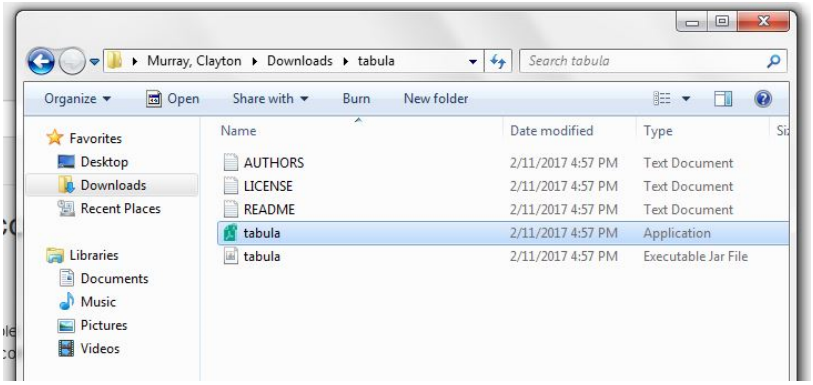
Using Tabula Extract to convert PDF tables back into Excel

A free alternative that beats Adobe products

Tabula Extract is free software that does not require installation and can be run on any Linux box or Windows machine. (However, I have not yet tried it on an IRS machine, as the site is blocked.)

Visit the website http://tabula.technology/ and download the proper distribution. We will be working with the Windows variation, as many analysts here use this predominantly.

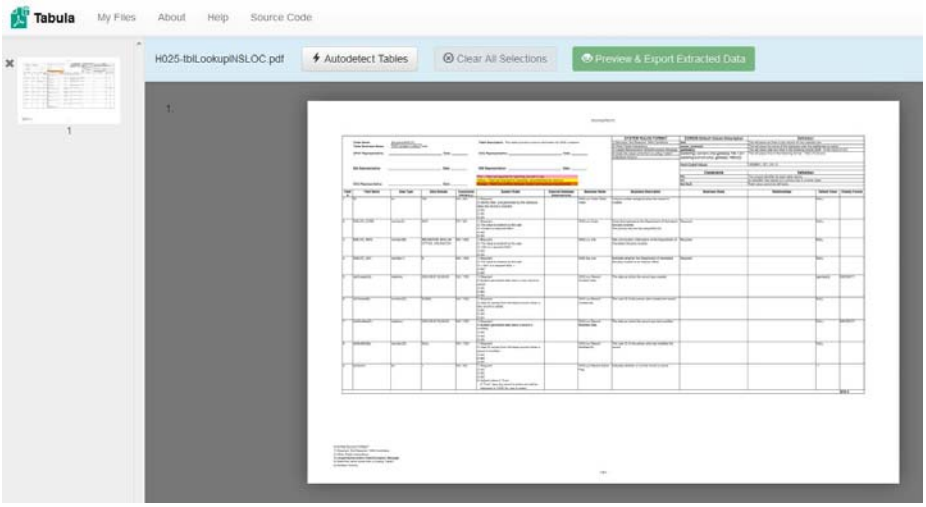
Step 1: Open the executable. It will run a script through the command prompt initializing the application. A window will pop up from your default browser, as this is the interface it uses. It is possible that stringent ad-blockers will prevent this; if there are issues here, double check the ad-block.



Step 2: Navigate to the browse field and select the file you want to convert. For this walk through, we are working with a file from the DOJ.

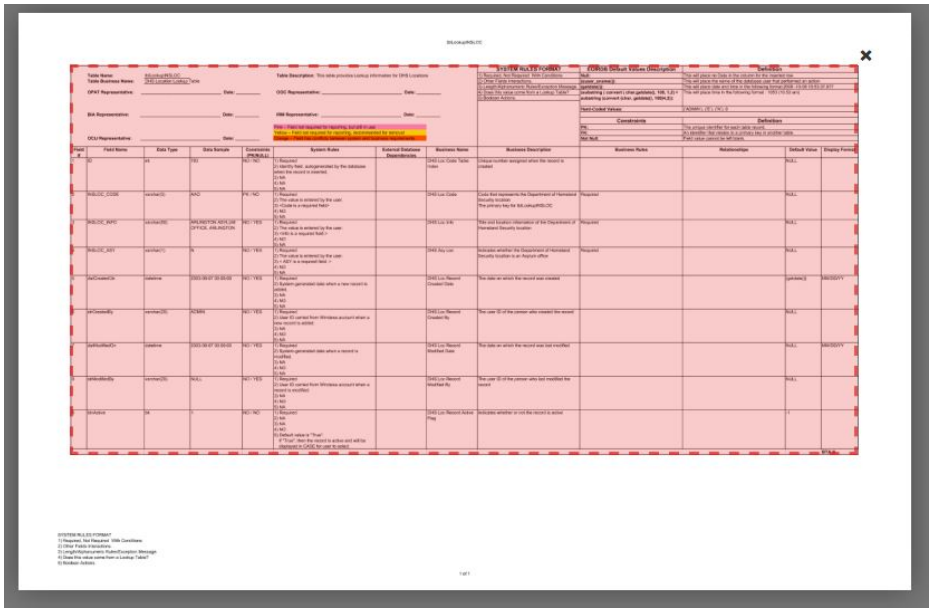
Step 3: After a loading bar indicating status reaches 100 percent, the file will appear in the window, page by page. Options above the page(s) should read 'Autodetect tables', 'Clear all selections', and 'Preview and export extracted data.' It is possible to also simply draw in conversion boxes (identifiable in red), but I have experienced consistency with autodetect.

Step 3



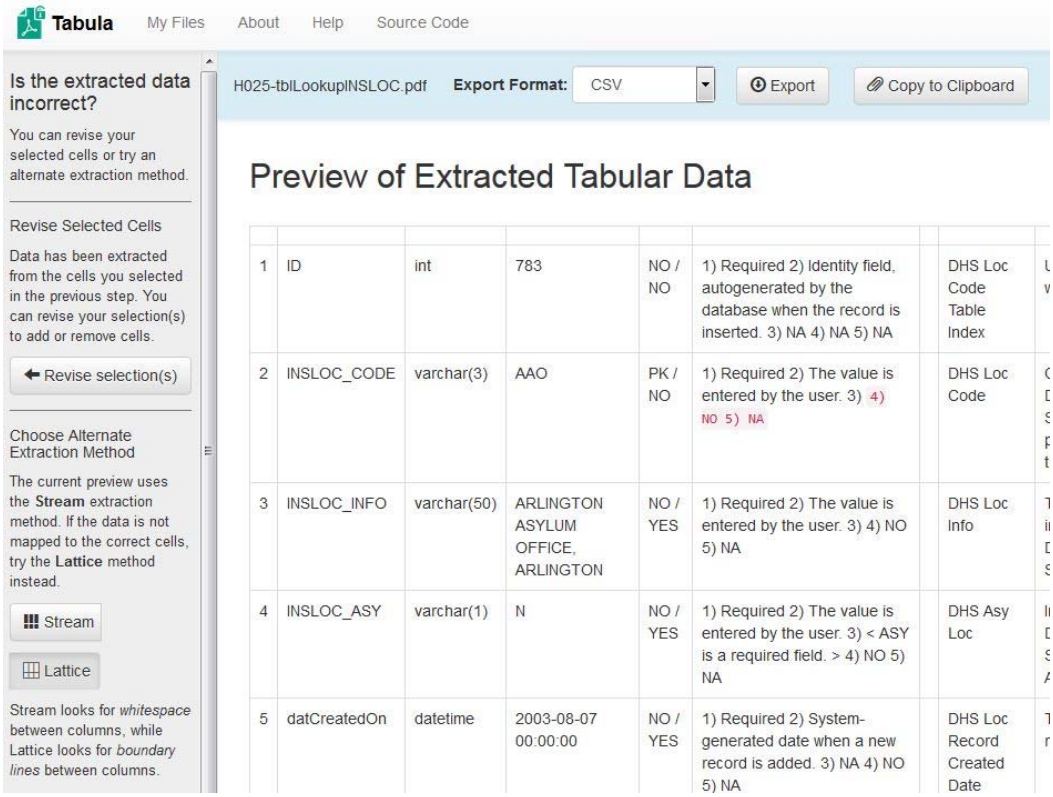
Step 4: When proceeding with AutoDetect, the application should do a good job of recognizing all tables. As shown, it also plays nicely with hanging text on the page – something I did not experience while trying to perform this task with Adobe products.

Note: The AutoDetected tables can be adjusted in size, and sometimes this *will* be necessary; experiment with some of the results. While extracting, I noticed entries touching the box boundaries would be omitted *unless* I slightly expanded the conversion box in their respective directions.



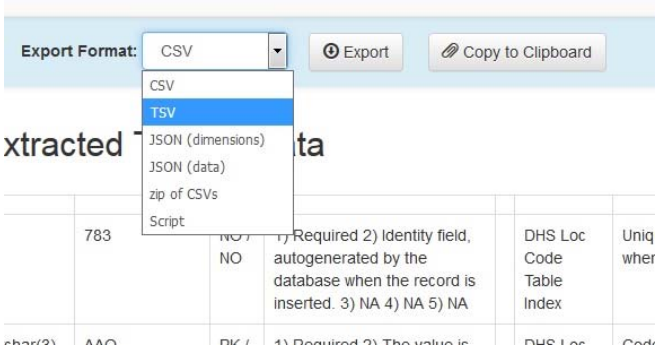
Step 5: Hitting preview and export should literally return a preview of the file to be exported. **You cannot change the table on this page.** If things do not appear to have converted properly, it might be required to return and reselect.

Note: Depending on the construction of the table itself, 'lattice' method could return more accurate results. Files I worked with included large, multi-line cells that confused the 'stream' method.

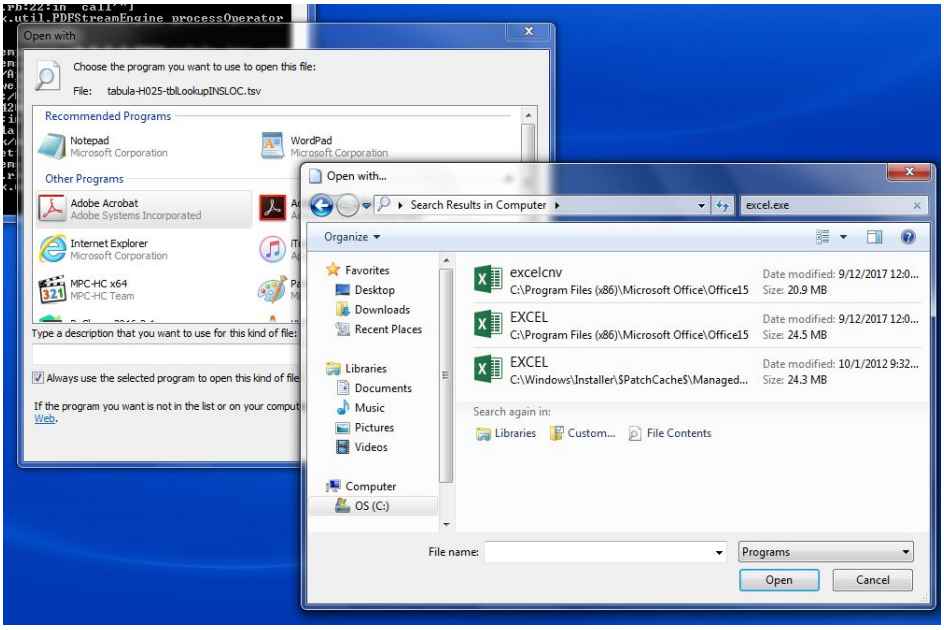


1	ID	int	783	NO / NO	1) Required 2) Identity field, autogenerated by the database when the record is inserted. 3) NA 4) NA 5) NA	DHS Loc Code Table Index
2	INSLOC_CODE	varchar(3)	AAO	PK / NO	1) Required 2) The value is entered by the user. 3) 4) NO 5) NA	DHS Loc Code
3	INSLOC_INFO	varchar(50)	ARLINGTON ASYLUM OFFICE, ARLINGTON	NO / YES	1) Required 2) The value is entered by the user. 3) 4) NO 5) NA	DHS Loc Info
4	INSLOC_ASY	varchar(1)	N	NO / YES	1) Required 2) The value is entered by the user. 3) < ASY is a required field. > 4) NO 5) NA	DHS Asy Loc
5	datCreatedOn	datetime	2003-08-07 00:00:00	NO / YES	1) Required 2) System-generated date when a new record is added. 3) NA 4) NO 5) NA	DHS Loc Record Created Date

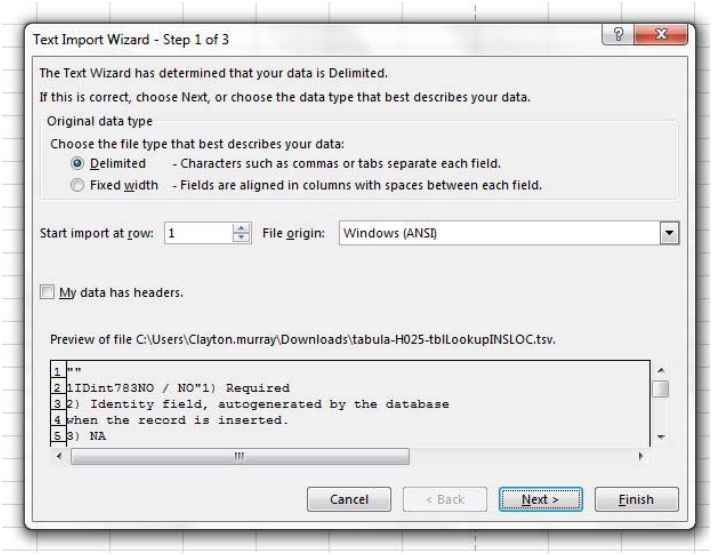
Step 6: There are file type options when exporting. I experimented with CSV and TSV; if your table is complicated with multiple values or lines in a cell, I recommend trying TSV. We will learn how to use these in Excel later. **If your file includes many commas, other punctuation marks, or possibly hidden characters due to JSON or other file formats, TSV is highly recommended.** This option is located in the drop down menu at the top. Hitting export to the right will generate the file. If you have worked with TSV files from this machine in the past, you should be done with the process here.



Step 7: Save the file here at the download prompt. (You can also paste it all to the clipboard.) Excel will sometimes struggle with files other than its proprietaries. Thankfully, the process dealing with them is straightforward. Navigate to the folder containing your TSV file, and right click to access 'open with'. Here you will select 'Choose program default.' If Excel is not listed under the 'other programs' list, click on browse, and on the left-hand side be sure to select 'Computer', as we will need to search the entire machine for Excel. (Sometimes the program is in different locations.) In the top right, type 'Excel.exe', and after results are returned select the Excel **associated with the Program Files, Office, or x86 folders.** Caches change and clear out, so we do not want to use that.



Other method: **This is an additional step you can perform if you are experiencing strange conversion behavior.** Excel can also attempt to best convert the file, but this is probably unnecessary as it seems Tabula handles most of the heavy lifting. This is done by having Excel explicitly open the file. With Excel open, hit ctrl + O and navigate to the proper folder. In the file type drop down, select 'All files' and proceed to open the TSV. This action should open import wizard that will require some knowledge of the file to select proper parameters.



The import wizard did not gracefully handle the DOJ file, as shown below, but it will likely work for simpler files.