Tipsheet: Extracting data tables from PDF files

Kaas & Mulvad, DataHarvest 2014

www.kaasogmulvad.dk

https://www.facebook.com/pages/Kaas-Mulvad/227948559464

Here we list some of the tools we have tried. Pdf files are tricky and don't expect the same tool to extract everything perfect. It's good to know several tools.

www.cometdocs.com

Online conversion. Free or several subscription models. From 30 days: 10$. To lifetime: 130$

Works fine (most of the time), but best if you use the login, upload your pdf-file, start the conversion and download the spreadsheet. Converts multipage pdf-files.

The free account has a limit on max 5 conversion per week. No limits if you subscribe.

Full list of the difference between free and paid services: http://www.cometdocs.com/user/subscriptions

NB: Subscription is free, if you are a member of Investigative Reporters and Editors, IRE.

Other tools we have tried – they might work on most files and still fail on some.

Able2extract http://www.investintech.com/

Free 7-day trial. Runs on macOS, Windows and Linux. 30-day version: 35$. Full version: 100$

PDF2XL http://www.cogniview.com/

Free 7-day trial. Runs on Windows. From 82$.

http://www.foolabs.com/xpdf

Xpdf is an opensource project, which includes a tool for converting pdf-files to text files. After that some work is required to change text files to spreadsheets again. The program must be run from the command line.

Tabula tabula.nerdpower.org

Tabula, created by a group of journalists and developers at ProPublica and the Knight-Mozilla Fellowship, is a free, open code application that allows users to upload their files and select the tables from the PDF they want to extract into CSV files. Does not (yet) work with multipage files. Runs on all platforms. We have had problems when we did tests, so the application might be unstable.

The tools mentioned above works by extracting the text – like you would do manually by copy/paste.

Another approach is to use an OCR tool. OCR works by "reading" the pdf even if the pdf is created like a picture. OCR also works with picture file formats like jpg, tiff or png. It works if it contains typewritten or printed text and it's changed into machine-encoded/computer-readable text.

Many applications claim to work, but the quality differs.

Wikipedia has created a comparison of Optical Character Recognition software:

http://en.wikipedia.org/wiki/Comparison_of_optical_character_recognition_software

A free OCR tool which works fine, if you don't have to convert large documents, is this:

http://www.onlineocr.net/

We have had great success using especially ABBYY FineReader:

http://www.abbyy.com/

The learning curve is not too steep. You can try it for free for 30 days – max 100 pages. A full version is 129$.

Other tools to manipulate pdf files

Adobe Acrobat XI

XI Reader http://www.adobe.com/dk/products/reader.html

With the free reader you can open and read pdf files. You can't change them, but you can copy text from tables in pdf files (unless they are created by scanning a page) and paste the text into a spreadsheet. If you press down the Alt-key on the keyboard, while using the mouse, you can highlight single columns in the data area and copy those one at a time.

Adobe Acrobat XI Standard (or Pro) http://www.adobe.com/dk/products/acrobat.html

With the full version you have lots of possibilities. Especially one feature is valuable for a datajournalist: When you highlight a table and right-click, you can open the table in a spreadsheet or save as a table. Usually the result is very close to what you want.

With the full version you can merge and split pdf files and lots of other stuff. But if you just need to split a document – or isolate a page with a table from a single report – you can just use free tools like http://www.ilovepdf.com/ or https://pdfmerge.w69b.com/

Tools for unlocking pdf files

To protect files some might choose to secure a pdf file with a password. If you receive a file like that from a source it means, that you perhaps can't copy content and paste it elsewhere if you haven't got the password.

A file like that can be unlocked. One way is to use this application, which also works fine, if you have many locked files. You can download a free trial version – the free version can only unlock two pages in the document. The paid version is only 9$ for a single user version.

http://www.pdf1.org/