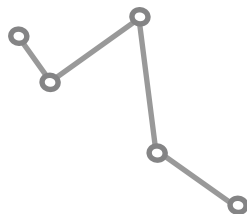
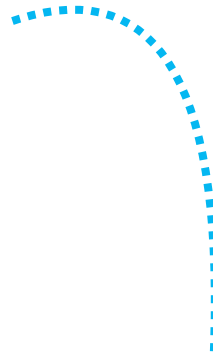
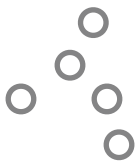




DATENSCHULE

# PDF - SCRAPING

## MIT TABULA



OPEN  
KNOWLEDGE  
FOUNDATION  
DEUTSCHLAND

Mehr zu unseren Projekten & Workshops:  
[datenschule.de](http://datenschule.de)

E-Mail: [info@datenschule.de](mailto:info@datenschule.de) |

Telefon: 030-57703666-2

# PDF-SCRAPING IN TABULA

## INHALTSVERZEICHNIS

1. My Files : Daten Importieren
2. Get the Info
3. Preview of Extracted Data
4. Export Data

Tabula ist ein Tool, mit dem Informationen aus PDF-Dokumenten systematisch extrahiert werden können. Die Daten stehen anschließend in Form von Tabellen zum Download bereit.

# 1. MY FILES: DATEN IMPORTIEREN

Wird Tabula gestartet gelangt man zunächst zu My Files. Hier können PDF-Dokumente ausgewählt und anschließend in Tabula importiert werden.

New version! Tabula 1.1.1 Release (1.1.1.17021118) is available (you have 1.0.1) ×

Import one or more PDFs

---

Imported PDFs

File Name	Size	Pages	Date Added	Remove	Process
Unterszeichner_Hamburger_Appell_2005.pdf	18 kB	5	19 Jun 2017 12:10	✕	<input type="button" value="Extract Data"/>
1807280_Spenden_Dezember_2015.pdf	229 kB	2	19 Jun 2017 12:03	✕	<input type="button" value="Extract Data"/>
Wahkampfslogan.pdf	75 kB	6	14 Jun 2017 08:46	✕	<input type="button" value="Extract Data"/>
Interessenvertreter.pdf	62 kB	3	14 Jun 2017 08:43	✕	<input type="button" value="Extract Data"/>
Gewerkschaftsorganisation.pdf	85 kB	3	13 Jun 2017 22:26	✕	<input type="button" value="Extract Data"/>
list_signatories_18012008.pdf	33 kB	6	31 May 2017 17:14	✕	<input type="button" value="Extract Data"/>

## 2. GET THE INFO

Im nächsten Schritt können die Informationen ausgewählt werden. Fortlaufende Tabellen, die mehrere Seiten umfassen müssen nicht einzelnen angewählt werden. Klickt man auf Repeat this Section, wird der ausgewählte Bereich auf die folgenden Seiten übertragen.

The screenshot shows a document viewer interface. At the top, there is a light blue header bar with the document title 'Unterzeichner\_Hamburger\_Appell\_2005...' on the left and three buttons: 'Autodetect Tables', 'Clear All Selections', and 'Preview & Export Extracted Data'. The main content area displays a table with a red dashed border. The table title is 'Unterzeichner des "Hamburger Appells" am 07.07.2005'. The table contains a list of names and their affiliations.

Unterzeichner des "Hamburger Appells" am 07.07.2005	
Prof. Dr. Frank Achtenhagen,	Universität Göttingen
Prof. Dr. Michael Albrecht,	Universität Hohenheim
Prof. Dr. Max Albert,	Universität des Saarlandes
Prof. Dr. Erwin Amann,	Universität Eisen
Prof. Dr. Peter Anker,	Universität Duisburg
Prof. Dr. Thomas Apolte,	Universität Münster
Prof. Dr. Gerhard Aminger,	Universität Wuppertal
Prof. Dr. Lutz Arnold,	Universität Regensburg
Prof. Dr. Jürgen G. Backhaus,	Universität Erfurt
Prof. Dr. Peter Baeis,	Universität Hohenheim
Prof. Dr. Tamis Bauer,	Universität Frankfurt
Prof. Dr. Dieter Bender,	Universität Bochum
Prof. Dr. Siegfried Berninghaus,	Universität Karlsruhe
Prof. Dr. Norbert Berthold,	Universität Würzburg <a href="#">siehe auch</a>
Prof. Dr. Helmut Besten,	Freie Universität Berlin
Prof. Dr. Michael Binder,	Universität Frankfurt
Prof. Dr. Charles Blankart,	Humboldt-Universität Berlin
Prof. Dr. Matthias Blonski,	Universität Frankfurt
Prof. Dr. Ulrich Blum,	Institut für Wirtschaftsforschung, Halle
Prof. Dr. Stephan Brandmüller,	ifo, Business School of Finance & Management, Frankfurt
Prof. Dr. Friedrich Breyer,	Universität Konstanz
Prof. Dr. Johannes Bröcker,	Universität Kiel
Prof. Dr. Udo Broll,	Technische Universität Dresden
Prof. Dr. Dieter Bräunerhoff,	Universität Rostock
Prof. Dr. Walter Bahr,	Universität Siegen
Prof. Dr. Michael Bards,	Humboldt-Universität Berlin
Prof. Dr. Hans-Peter Barchhof,	Universität Hohenheim
Prof. Dr. Thies Blümel,	Universität München und ifo
Prof. Dr. Rolf Caspar,	Universität Hohenheim
Prof. Dr. Dieter Casel,	Universität Duisburg

# 3. PREVIEW OF EXTRACTED DATA

Anschließend werden die gescrapten Informationen im Preview-Modus angezeigt.

The screenshot shows a web interface for previewing extracted data. On the left, there is a sidebar with the following sections:

- 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.  
[← Revise selection(s)]
- 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.  
[Stream] [Lattice]
- Stream looks for whitespace between columns, while Lattice looks for boundary lines between columns.**
- Still look wrong?**  
Contact the developers and tell us what you tried to do that didn't work.

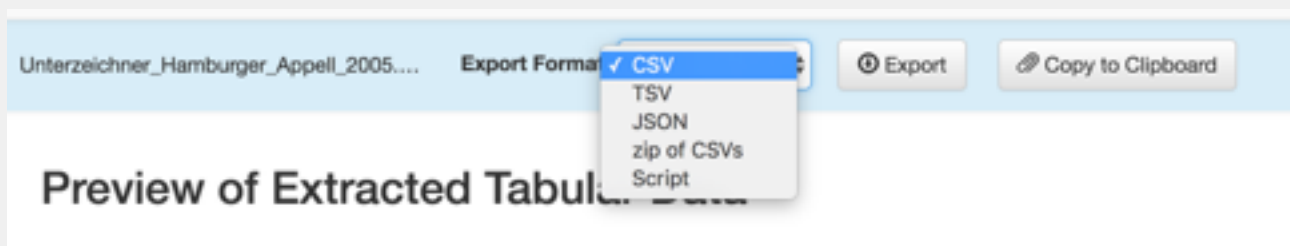
The main area is titled "Preview of Extracted Tabular Data" and shows a list of names and universities:

Prof. Dr. Frank Achtenhagen, Universität Göttingen
Prof. Dr. Michael Ahlheim, Universität Hohenheim
Prof. Dr. Max Albert, Universität des Saarlandes
Prof. Dr. Erwin Amann, Universität Essen
Prof. Dr. Peter Anker, Universität Duisburg
Prof. Dr. Thomas Apolte, Universität Münster
Prof. Dr. Gerhard Armingier, Universität Wuppertal
Prof. Dr. Lutz Arnold, Universität Regensburg
Prof. Dr. Jürgen G. Backhaus, Universität Erfurt
Prof. Dr. Peter Bareis, Universität Hohenheim
Prof. Dr. Tamás Bauer, Universität Frankfurt
Prof. Dr. Dieter Bender, Universität Bochum
Prof. Dr. Siegfried Beringhaus, Universität Karlsruhe
Prof. Dr. Norbert Berthold, Universität Würzburg siehe auch

Standardmäßig wird die *Stream-Methode* zur Extraktion von Informationen angewendet. Hierbei wird nach Leerzeilen zwischen den Spalten gesucht um diese voneinander abzutrennen. Als zweite Methode steht die *Lattice-Methode* bereit. Hier wird nach Trennlinien Ausschau gehalten, um Spalten identifizieren zu können. Die Qualität des PDFs bestimmt dabei maßgeblich die Qualität der Resultate.

## 4. Export Data

Wenn alle Informationen ordentlich erkannt wurden, können die Daten anschließend exportiert werden. Hierfür stehen die Formate CSV, gezippte CSVs, TSV, JSON oder ein Script zur Verfügung. Außerdem können die Daten über den button „Copy to Clipboard“ in die Ablage kopiert werden.





*Die Datenschule vermittelt gemeinnützigen Organisationen die nötigen Fähigkeiten, Daten und Technologien zu verstehen, um sie zielgerichtet für ihre gesellschaftlichen Aufgaben einzusetzen.*



OPEN  
KNOWLEDGE  
FOUNDATION  
DEUTSCHLAND

*Die Open Knowledge Foundation Deutschland ist ein gemeinnütziger Verein, der sich für offenes Wissen, offene Daten, Transparenz und Beteiligung einsetzt.*

Mehr zu unseren Projekten &  
Workshops: [datenschule.de](https://datenschule.de)  
E-Mail: [info@datenschule.de](mailto:info@datenschule.de) |  
Telefon: 030-57703666-2