DATENSCHULE

PDF - SCRAPING MIT TABULA

OPEN KNOWLEDGE FOUNDATION DEUTSCHLAND

Mehr zu unseren Projekten & Workshops:
datenschule.de
E-Mail: info@datenschule.de |
Telefon: 030-57703666-2

PDF-SCRAPING IN TABULA

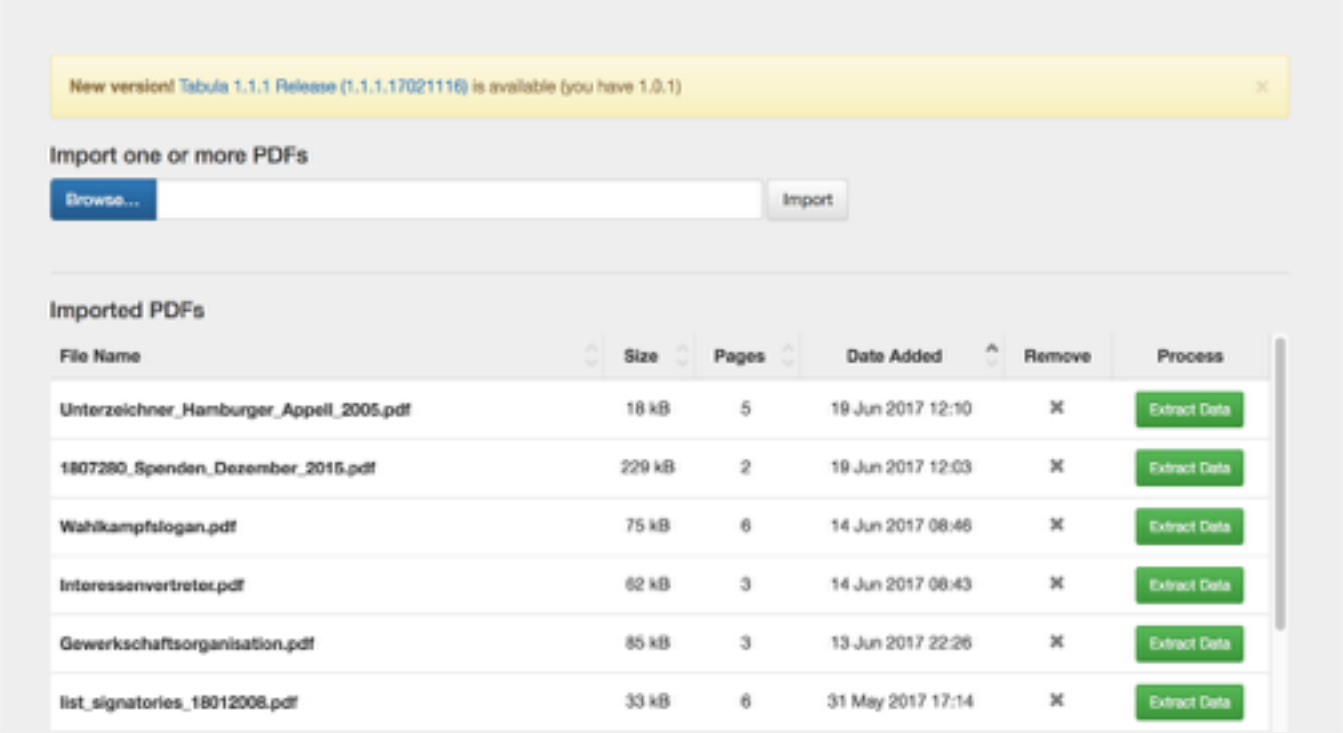
INHALTSVERZEICHNIS

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.

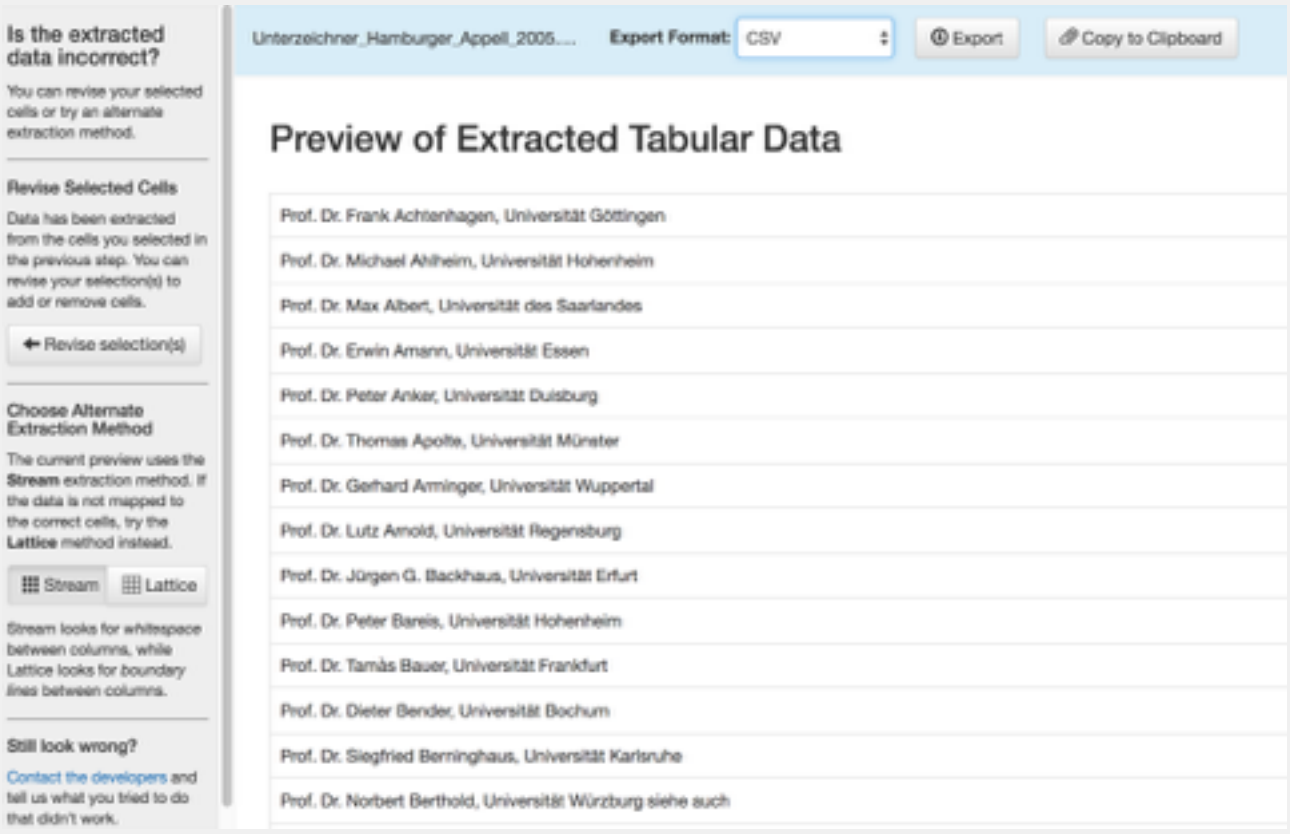


2. GET THE INFO

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.

3. PREVIEW OF EXTRACTED DATA

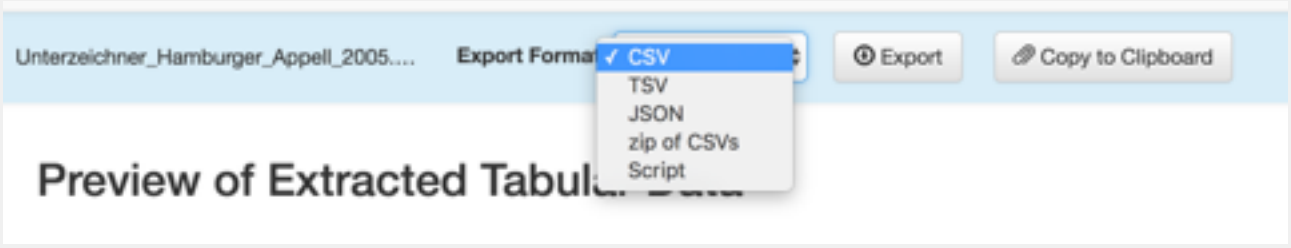
Anschließend werden die gescrapten Informationen im Preview-Modus angezeigt.



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, gezipte CSVs, TSV, JSON oder ein Script zur Verfügung. Außerdem können die Daten über den button „Copy to Clipboard“ in die Ablage kopiert werden.



DATENSCHULE

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
E-Mail: info@datenschule.de |
Telefon: 030-57703666-2