# **Landscape of existing Tools for PDF scraping**

**R Packages for PDF Table Extraction:**

1. **Tabulizer and tabulaR:**
   * Purpose: Extraction of tables from PDFs.
   * Category: R packages with Tabula Java library integration.
2. **pdftools:**
   * Purpose: Extraction of text and structured information from PDFs, including tables.
   * Category: R package for general PDF content extraction.
3. **tab:**
   * Purpose: Extraction of tables from PDFs.
   * Category: R package for PDF table extraction.
4. **readpdf:**
   * Purpose: Reading PDF files, including extracting tables, text, and metadata.
   * Category: R package for PDF content extraction.
5. **tabularaster:**
   * Purpose: Extraction of tabular data from scanned or OCR-processed PDFs.
   * Category: R package for tabular data extraction from scanned PDFs.

**Python Libraries for PDF Table Extraction:**

1. **Tabula (tabula-py):**
   * Purpose: Extraction of tables from PDFs.
   * Category: Python library with Java integration for PDF table extraction.
2. **pdftabextract:**
   * Purpose: Extraction of tables from scanned and OCR-processed PDFs.
   * Category: Python library for extracting tables from scanned PDFs.

**General PDF Content Extraction Tools:**

1. **PDFMiner:**
   * Purpose: Extraction of text, images, and metadata from PDFs.
   * Category: General-purpose PDF content extraction in Python.
2. **PyPDF2:**
   * Purpose: Extraction of text and basic information from PDFs.
   * Category: Python library for basic PDF content extraction.
3. **PyMuPDF:**
   * Purpose: Extraction of text, images, and more from PDFs.
   * Category: Python library for advanced PDF content extraction.
4. **pdf2image:**
   * Purpose: Conversion of PDF pages to images, useful for extracting data from images in PDFs.
   * Category: Python library for PDF page-to-image conversion.

**Additional PDF Extraction Tools:**

1. **Tika:**
   * Purpose: Content extraction framework for various document formats, including PDFs.
   * Category: General-purpose content extraction tool.
2. **Slate:**
   * Purpose: Extraction of text from PDFs, useful for text extraction from PDF images.
   * Category: Python library for text extraction from PDFs.

## **ANNEX: Additional information on existing packages and how to access them**

1. **Tabula (Java):**
   * Tabula is a popular open-source tool for extracting tables from PDFs.
   * Website: <https://tabula-py.readthedocs.io/en/latest/tabula.html>
2. **Camelot (Python):**
   * Camelot is a Python library for extracting tables from PDFs.
   * Website: <https://camelot-py.readthedocs.io/en/master/>
3. **PDFMiner (Python):**
   * PDFMiner is a Python library for extracting text, images, and metadata from PDF files.
   * Website: <https://github.com/pdfminer/pdfminer.six>
4. **PyPDF2 (Python):**
   * PyPDF2 is a Python library for extracting text and basic information from PDF files.
   * Website: <https://pythonhosted.org/PyPDF2/>
5. **Tabula Rasa (R):**
   * Tabula Rasa is an R package that provides an interface to the Tabula Java library for extracting tables from PDFs.
   * GitHub: <https://github.com/dianagold/tabula-py>
6. **pdftools (R):**
   * pdftools is an R package for extracting text and metadata from PDFs.
   * CRAN: <https://cran.r-project.org/web/packages/pdftools/index.html>
7. **PyMuPDF (Python):**
   * PyMuPDF is a Python library for extracting text, images, and more from PDFs.
   * GitHub: <https://github.com/pymupdf/PyMuPDF>
8. **pdf2image (Python):**
   * pdf2image is a Python library that converts PDF pages to images, which can be useful for extracting data from images in PDFs.
   * GitHub: <https://github.com/Belval/pdf2image>
9. **PDFPlumber (Python):**
   * PDFPlumber is a Python library built on top of pdfminer.six for extracting text and tables from PDFs.
   * GitHub: <https://github.com/jsvine/pdfplumber>
10. **Tika (Java/Python):**
    * Apache Tika is a content extraction framework that can handle various document formats, including PDFs.
    * Website: <https://tika.apache.org/>
11. **Slate (Python):**
    * Slate is a Python library for extracting text from PDFs. While it doesn't handle tables and images directly, it can be useful for text extraction as a preprocessing step.
    * GitHub: <https://github.com/timClicks/slate>
12. **Tabulizer (R Package):**

* Tabulizer is an R package that is used for extracting tables from PDF documents.
* It is built on top of the Java-based Tabula library and provides an R interface for working with Tabula's functionality.
* GitHub: <https://github.com/ropensci/tabulizer>

1. **tabulaR:**
   * tabulaR is an R package that provides an interface to the Tabula Java library, similar to Tabulizer, for extracting tables from PDFs.
   * GitHub: <https://github.com/ropensci/tabulizer>
2. **pdftools:**
   * pdftools is an R package mentioned earlier that can extract text and some structured information from PDFs, including tables.
   * CRAN: <https://cran.r-project.org/web/packages/pdftools/index.html>
3. **tab:**
   * The "tab" package for R can be used to extract tables from PDFs.
   * GitHub: <https://github.com/leeper/tab>
4. **readpdf:**
   * The "readpdf" package is designed for reading PDF files, including extracting tables, text, and metadata.
   * CRAN: <https://cran.r-project.org/web/packages/readpdf/index.html>
5. **tabularaster:**
   * The "tabularaster" package is designed for extracting tabular data from scanned or OCR-processed PDFs.
   * GitHub: <https://github.com/sboysel/tabularaster>
6. **tabula (Python):**
   * Although Tabula is primarily a Java library, there is a Python wrapper called "tabula-py" that can be used in conjunction with R through the **reticulate** package.
   * GitHub: <https://github.com/chezou/tabula-py>
7. **pdftabextract (Python):**
   * If you're open to using Python, "pdftabextract" is a Python library for extracting tables from scanned and OCR-processed PDFs.
   * GitHub: <https://github.com/wolfgangB33r/pdftabextract>