Table 3 Overview of text and table extraction tools used in our study. Key extraction capabilities include extraction of Image (I), Text (T), Metadata (M), Table of Contents (TOC), and Table (TB). Most tools use rule-based (RB) technology, with some offering Optical Character Recognition (OCR) capabilities. However, Nougat and Table Transformers were not the primary focus of this study. Tool Version Extraction **Technology** Output PyPDF I. T. M RBTXT4.3.0pdfminer.six I, T, TOC TXT, HTML, hORC, JPG 20240706 RBPvMuPDF I, T, TB RB(MuPDF), OCR TXT, HTML, SVG, JSON 1.24.7 pdfplumber I, T, TB RB(pdfminer) TXT, HTML, hORC, JPG 0.11.2pvpdfium2 4.30.0 \overline{TXT} RBUnstructured 0.14.10T. TB RB, OCR TXTDataFrame, CSV, JSON Tabula 2.9.3 TB RBCamelot 0.11.0TB DataFrame, CSV, JSON, HTML RBNougat base(350M)Transformer Markdown Table Transformer TATR-v1.1-All TB Transformer Image

Image (I), Text (T), Metadata (M), Table of Contents (TOC), and Table (TB). Most tools use rule-based (RB) technology, with some offering Optical Character Recognition (OCR) capabilities. However, Nougat and Table Transformers were not the primary focus of this study.

Table 3 Overview of text and table extraction tools used in our study. Key extraction capabilities include extraction of

1001	version	Extraction	Technology	Output
PyPDF	4.3.0	I, T, M	RB	TXT
pdfminer.six	20240706	I, T, TOC	RB	TXT, HTML, hORC, JPG
PyMuPDF	1.24.7	I, T, TB	RB(MuPDF), OCR	TXT, HTML, SVG, JSON
pdfplumber	0.11.2	I, T, TB	RB(pdfminer)	TXT, HTML, hORC, JPG
pypdfium2	4.30.0	T	RB	TXT

RB. OCR

Transformer

Transformer

RB

RB

TXT

Image

Markdown

DataFrame, CSV, JSON

DataFrame, CSV, JSON, HTML

palminer.six	20240706	1, 1, 100	RB	TXT, HTML, nORC, J.
PyMuPDF	1.24.7	I, T, TB	RB(MuPDF), OCR	TXT, HTML, SVG, JS
pdfplumber	0.11.2	I, T, TB	RB(pdfminer)	TXT, HTML, hORC, J.
pypdfium2	4.30.0	T	RB	T

T. TB

TB

TB

TB

0.14.10

2.9.3

0.11.0

base(350M)

TATR-v1.1-All

Unstructured

Table Transformer

Tabula

Camelot

Nougat