Convert-Poppler Configuration Documentation

Overview

Convert-Poppler is a TypeScript utility for converting PDF files to text using the Poppler (pdftotext) command-line tool. It provides extensive configuration options for both the PDF extraction process and post-processing of the extracted text.

Configuration File Structure

The configuration is provided via a JSON file with the following structure:

```
json
{
    "inputDir": "string",
    "outputDir": "string",
    "processSubDirs": boolean,
    "popplerPath": "string",
    "pdfTextExtractOptions": {},
    "postProcessingOptions": {}
}
```

Main Configuration Options

(required)

- **Type**: (string)
- **Description**: Path to the directory containing PDF files to convert
- **Example**: ("C:\\docs\\transcripts\\pdf")

$ig({f output Dir} ig) ig({f required} ig)$

- Type: string
- **Description**: Path to the directory where converted text files will be saved
- **Example**: ("C:\\docs\\transcripts\\txt")

processSubDirs (optional)

- **Type**: (boolean)
- **Default**: false

- **Description**: When true, processes subdirectories one level deep within inputDir
- Behavior:
 - (false): Processes PDFs directly in (inputDir)
 - true: Processes PDFs in subdirectories of (inputDir), creating matching subdirectories in (outputDir)

popplerPath (optional)

- **Type**: string
- **Default**: Uses system PATH
- **Description**: Directory containing the (pdftotext) executable
- **Example**: ("C:\\Program Files\\poppler\\bin")

PDF Text Extract Options

Options passed directly to the (pdftotext) command. Place these in the (pdfTextExtractOptions) object.

Layout Options

layout) (boolean)

- Description: Maintain original physical layout
- Recommended: (true) for legal transcripts

simple (boolean)

• **Description**: Simple one-column page layout

simple2 (boolean)

• **Description**: Simple one-column page layout, version 2

table (boolean)

• **Description**: Similar to layout, optimized for tables

lineprinter (boolean)

• **Description**: Use strict fixed-pitch/height layout

raw) (boolean)

• **Description**: Keep strings in content stream order

phys (boolean)

• **Description**: Physical layout mode

Resolution and Spacing

r (number)

• **Description**: Resolution in DPI (dots per inch)

• **Default**: 72

• Recommended: 300 for accurate spacing detection

• **Example**: ("r": 300)

fixed (number)

• Description: Assume fixed-pitch (or tabular) text

• **Recommended**: 0 for variable-width fonts

• **Example**: ("fixed": 0)

(linespacing) (number)

• **Description**: Fixed line spacing for LineePrinter mode

• **Example**: ("linespacing": 1.2)

Page Selection

f (number)

• **Description**: First page to convert

• Example: ("f": 1)

(number)

• **Description**: Last page to convert

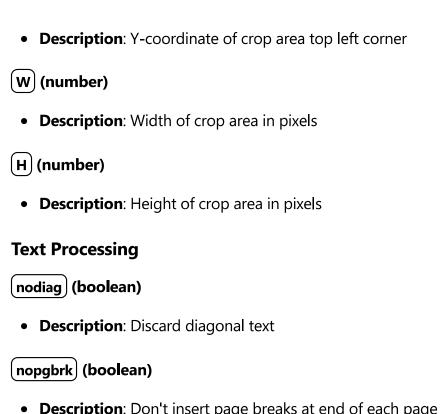
• Example: ("I": 10)

Area Selection

x (number)

• **Description**: X-coordinate of crop area top left corner

y (number)



Description: Don't insert page breaks at end of each page

Recommended: (true) for continuous text

bbox (boolean)

Description: Output bounding box information

bbox-layout (boolean)

• **Description**: Output bounding box information with layout

tsv (boolean)

Description: Output in TSV (Tab-Separated Values) format

Encoding Options

enc (string)

Description: Output text encoding

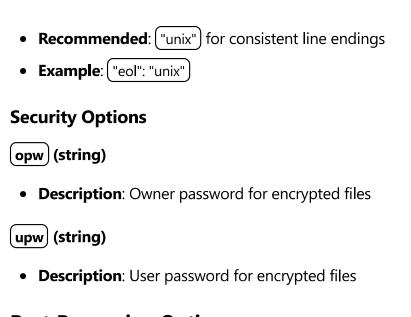
Recommended: ("UTF-8")

Example: ("enc": "UTF-8"

eol (string)

Description: End-of-line convention

Options: ["unix"] "dos" "mac"



Post-Processing Options

Options for cleaning and normalizing the extracted text. Place these in the (postProcessingOptions) object.

fixTranscriptSpacing) (boolean)

- Default: false
- **Description**: Fixes common spacing issues in legal transcripts
- Fixes:
 - Line numbers touching Q/A markers: ("9Q") → ("9 Q")
 - Attorney names: ("BYMR.SMITH") → ("BY MR. SMITH")
 - Time stamps: ("9:30a.m.") → ("9:30 a.m.")
 - Page/line references: ("page42") → ("page 42")

fix Transcript Quotes (boolean)

- **Default**: false
- **Description**: Normalizes quotes and special characters
- Fixes:
 - Smart quotes → straight quotes: (""") → ("""
 - Em dashes → double hyphens: ("—") → ("--")
 - En dashes → single hyphens: ("-") → ("-")
 - Ellipsis character → three dots: ("...") → ("..."

normalizeWhitespace (boolean)

• **Default**: (false)

- **Description**: Comprehensive whitespace normalization
- Actions:
 - Replaces multiple spaces with single space
 - Removes trailing whitespace from lines
 - Ensures consistent spacing after punctuation
 - Trims lines

normalizeLineNumberWhitespace (boolean)

- **Default**: (false)
- **Description**: Removes leading whitespace before line numbers
- Behavior:
 - Removes spaces before 1-2 digit numbers at start of lines
 - Preserves lines without line numbers (headers, etc.)
 - Specifically designed for legal transcript formatting

Example Configurations

Basic Configuration

```
json
{
  "inputDir": "./pdfs",
  "outputDir": "./txt",
  "pdfTextExtractOptions": {
    "layout": true,
    "r": 300
  }
}
```

Legal Transcripts with Full Processing

```
json
```

```
"inputDir": "C:\\transcripts\\pdf",
 "outputDir": "C:\\transcripts\\txt",
 "processSubDirs": true,
 "pdfTextExtractOptions": {
  "layout": true,
  "r": 300,
  "nopgbrk": true,
  "enc": "UTF-8",
  "eol": "unix",
  "fixed": 0
 },
 "postProcessingOptions": {
  "fixTranscriptSpacing": true,
  "fixTranscriptQuotes": true,
  "normalizeLineNumberWhitespace": true
}
```

Custom Poppler Installation

```
json
{
    "inputDir": "F:\\docs\\pdf",
    "outputDir": "F:\\docs\\txt",
    "popplerPath": "C:\\tools\\poppler-24.07\\bin",
    "pdfTextExtractOptions": {
        "layout": true,
        "r": 300
     }
}
```

Usage

```
bash
npx ts-node convert-poppler.ts config.json
```

Tips for Legal Transcripts

1. **Use high resolution**: Set ("r": 300) for accurate spacing detection

- 2. **Preserve layout**: Use ("layout": true) to maintain structure
- 3. **Enable line number normalization**: Use <u>"normalizeLineNumberWhitespace"</u>: true for consistent formatting
- 4. **Fix spacing issues**: Enable ("fixTranscriptSpacing": true) for Q/A formatting
- 5. **Use consistent encoding**: Set ("enc": "UTF-8") and ("eol": "unix")

Troubleshooting

- Missing spaces between columns: Increase resolution with ("r": 300) or higher
- Smart quotes in output: Enable ("fixTranscriptQuotes": true)
- Inconsistent line numbers: Enable ("normalizeLineNumberWhitespace": true)
- Page breaks in output: Set ("nopgbrk": true)
- Can't find pdftotext: Specify full path with ["popplerPath"]