Stapelklasseneigenschaften OCR-Texterkennung

Allgemeines

Die OCR Extraktion ist ein elementarer Teil der Squeeze Software. Dieser Kernbereich der Software ist mit verschiedenen Einstellungen versehen, die das Ergebnis der Extraktion tangieren. Im folgenden Artikel gehen wir auf die Besonderheiten und die Anforderungen der unterschiedlichen Eigenschaften ein.

Welche Arten der OCR unterstützt Squeeze?

Grundsätzlich unterscheiden wir im Kontext von Squeeze zwischen dem Einsatz einer OCR basierend auf den Ressourcen der lokalen Maschine und dem Einsatz eines Remote-OCR-Dienstes.

Was beinhaltet meine Standardversion von Squeeze?

Im Auslieferungszustand ist Squeeze mit einer lokal verfügbaren OCR-Engine ausgestattet. Auf Kundenwunsch können unsere Berater bei einer Squeeze Installation ab der Version 2.4 eine Remote-OCR aktivieren, die mithilfe von AI bessere Ergebnisse liefern kann.

> Alle hier beschriebenen Eigenschaften können sowohl auf der **Stapelklasse** als auch auf der **Dokumentenklasse** gesetzt werden. Ist eine Eigenschaft auf beiden Ebenen gesetzt, gewinnt der Wert der Dokumentenklasse.

Allgemeine Stapelklassen-Eigenschaften

OCREngine (ab Squeeze 2.4)

Wird diese Stapelklassen-Eigenschaft nicht konfiguriert greift automatisch die lokale OCR-Engine `ocrmypdf` (Wert `default`).
Je nach Spezifikation und Lizensierung ihres Squeeze-Systems können folgende Optionen für die OCREngine verwendet werden:

Wert	Squeeze Version	Beschreibung
`default`	ab 2.4.0	Lokale OCR über ocrmypdf. Standard, wenn nichts konfiguriert ist.
`maxocr`	ab 2.5.0	Remote-OCR über die Dexpro Platform.
`proxy-ocr`	ab 2.6.0	Remote-OCR über den OCR-Proxy.
`squeeze`	ab 2.7.0	Lokale OCR auf Basis von Tesseract und Ghostscript, Seiten werden parallel verarbeitet.
`no-ocr`	ab 2.31.0	Es wird keine OCR durchgeführt. Wird u. a. im ZUGFeRD Strict Mode verwendet.

> Der frühere Wert `ai-ocr` wird nicht mehr unterstützt. Bestehende Konfigurationen wurden beim Update automatisch auf `proxy-ocr` umgestellt.

Voraussetzungen:

- default:
  - keine
- squeeze:
  - keine
- maxocr:
  - die konfigurierte Mandanten-Konfiguration/Server-Konfiguration für die Dexpro Platform Integration.
  - die MaxOCR Konfiguration.
- proxy-ocr:
  - die KI Proxy Konfiguration

OCRPageLimit – Begrenzung der verarbeiteten Seiten

Mit `OCRPageLimit` wird festgelegt, welche Seiten eines Dokumentes ausgelesen werden. Der typische Anwendungsfall sind sehr umfangreiche Belege mit vielen Positionsseiten, bei denen nicht alle Positionen benötigt werden – in der Regel die ersten Seiten mit den Kopfdaten sowie die letzten Seiten, auf denen die Rechnungssumme steht.

Syntax

Der Wert ist eine durch Komma getrennte Liste aus drei möglichen Bausteinen:

Form	Bedeutung	Beispiel
`n`	einzelne Seite	`3`
`n-m`	Seitenbereich von n bis m	`1-15`
`-n`	die **letzten** n Seiten des Dokumentes	`-2`

Beispiele:

- `1-3` – die ersten drei Seiten
- `1-5,-2` – die ersten fünf Seiten und die beiden letzten Seiten
- `1-15,-2` – die ersten 15 Seiten und die beiden letzten Seiten

Regeln der Eingabeprüfung

Die Eingabe wird beim Speichern geprüft. Folgendes führt zu einer Fehlermeldung:

- Leerzeichen, die Ziffer `0` oder andere Zeichen als Ziffern, Komma und Bindestrich
- ein Komma am Ende des Wertes
- mehr als eine Angabe der Form `-n` (die „letzten n Seiten" dürfen nur einmal angegeben werden)
- Angaben wie `-3-5`
- eine Einzelseite, die bereits in einem angegebenen Bereich enthalten ist (z. B. `1-15,7`)

Wirkung je OCREngine

> Die Eigenschaft wirkt nicht bei jeder OCREngine gleich. Vor dem Einsatz bitte die eingestellte `OCREngine` prüfen.

OCREngine	Wirkung	Hinweis
`default`	vollständig	Die OCR wird ausschließlich auf den angegebenen Seiten gerechnet. Spart Verarbeitungszeit.
`proxy-ocr`	vollständig (ab 2.10.0)	Die Seitenauswahl wird an den OCR-Proxy übergeben, es werden nur diese Seiten verarbeitet.
`squeeze`	eingeschränkt	Die OCR läuft weiterhin über alle Seiten, lediglich das anschließende Auslesen wird begrenzt. Es entsteht dadurch **kein** Zeitgewinn bei der Texterkennung.
`maxocr`	keine	Die Einstellung wird ignoriert, es wird immer das vollständige Dokument verarbeitet.
`no-ocr`	entfällt	Es findet keine OCR statt.

Vorrang gegenüber AttachmentBarcodePattern

> Ist auf der Stapel- oder Dokumentenklasse ein `AttachmentBarcodePattern` konfiguriert, bestimmt dieses Pattern die zu verarbeitenden Seiten und `OCRPageLimit` wird vollständig ignoriert. Beide Eigenschaften lassen sich nicht kombinieren.

Hinweis zum Auslesen des Textlayers

Beim lokalen Auslesen des Textlayers (PDFBox/PDFMiner) werden nur die erste und die letzte Seite der Auswahl als Rahmen verwendet. Bei einem Wert von `1-15,-2` auf einem 50-seitigen Dokument bedeutet das: Die OCR wird nur auf den Seiten 1–15 und 49–50 gerechnet, beim Auslesen eines digitalen PDFs kann der bereits vorhandene gedruckte Text der Zwischenseiten jedoch mitgelesen werden. Das ist kein Fehler, erklärt aber, warum in einzelnen Fällen Text außerhalb der gesetzten Seitenauswahl im Dokument erscheint.

Auswirkung auf die Seitenzählung

Die über `OCRPageLimit` gesetzte Auswahl bestimmt auch die Anzahl der verarbeiteten Seiten, die im Exportstatus eines Dokumentes protokolliert wird.

Stapelklassen-Eigenschaften für die lokale OCR Engine

OCRForce

Im Standard wird bei digitalen PDF´s der Textlayer genutzt und die Felderkennung darauf angewendet (`false`). Um aber eine OCR zu erzwingen ist dieser Schalter auf `true` zu setzen.

OCRLanguage

Im Standard werden die Sprachpakete Deutsch und Englisch verwendet. Für die deutsche Detektion wird der Wert `deu` eingetragen und für die englische Detektion der Wert `eng` eingetragen.

Hier können projektspezifisch auch weitere Sprachen oder abgewandelte Sprachpaket-Varianten angegeben werden, bei denen die OCR schneller/langsamer bzw. mit niedriger/höherer Qualität Ergebnisse liefert. Im folgenden eine Übersicht über die im Standard enthaltenen Sprachpakete:

Squeeze Version	Optionen
vor 2.4.0	• deu • eng

ab 2.4.0	• deu, deu_best, deu_fast, deu_std • lat_best, lat_fast, lat_std • eng • osd

PDFA-Conversion

Es wird ein PDFA kompatibles Dokument erzeugt. Eingabe 1|0 (`true`|`false`)

PDFProcessor

Hier gilt `PDFBox` als Standard. `PDFMiner` ist die Alternative.

PSM-Modes

Im Project bietet es sich an, die Modi 3, 4, 6 und 11 zu verwenden. Dabei gilt 3 als Standard.

3	Standardeinstellung liefert gute Ergebnisse.
4	Wortweise Segmentierung. Es wird nicht nach Zeilen geschaut sondern Worten. (verfügbar ab Version 2.0)
6	Gut für Positionsdaten. Hat aber Probleme bei Linien die sehr dicht am Text sind.
11	Gut bei vielen Grafiken auf den Dokumenten.

OCRRotationThreshold

Mit dieser Eigenschaft können Sie beeinflussen wie *agressiv* Seiten in der OCR gedreht werden. Nutzen Sie diesen Wert, wenn Dokumente falsch gedreht werden.

Geringe Werte führen dazu, dass mehr Dokumente gedreht werden. Die Software muss sich also nicht sehr sicher sein, dass eine Seite rotiert werden muss.

Hohe Werte führen dazu, dass Dokumente seltener gedreht werden, also nur wenn sich die Software sehr sicher ist, dass eine Seite rotiert werden muss.

Im Standard ist dieser Wert `9.0`

Stapelklassen-Eigenschaften für die MaxOCR/OCR-Proxy Engine

Die Remote-OCR-Dienste lassen sich über die Stapelklassen-Eigenschaften nur eingeschränkt beeinflussen. Verfügbar ist:

- `OCRPageLimit` bei der Engine `proxy-ocr` (ab Version 2.10.0) – siehe Abschnitt *OCRPageLimit – Begrenzung der verarbeiteten Seiten*. Bei `maxocr` hat die Eigenschaft keine Wirkung.

Die übrigen Eigenschaften dieses Artikels (OCRForce, OCRLanguage, PSM-Mode, OCRRotationThreshold, PDFA-Conversion, PDFProcessor) gelten ausschließlich für die lokalen OCR-Engines.

Fragen und Antworten?

1. Ich habe die `maxocr/proxy-ocr` als OCREngine Eigenschaft ausgewählt, jedoch funktioniert die Texterkennung nicht mehr?
   - Gehen Sie bitte sicher, dass Ihr Squeeze Berater die notwendigen Anmeldedaten zur Aktivierung der Remote OCR hinterlegt hat.
2. Ich habe mit der Remote-OCR ein Dokument verarbeitet, mehrere Dokumente liefen erfolgreich durch, jedoch bleibt dieses Dokument hängen.
   - Aufgrund der begrenzten Ressourcen kann die Remote-OCR maximal 100 Seiten pro Dokument verarbeiten. Überprüfen Sie daher die Anzahl der Seiten und nutzen bei nicht erfolgreicher Verarbeitung die lokale OCR. Alternativ kann bei der Engine `proxy-ocr` über `OCRPageLimit` die Anzahl der verarbeiteten Seiten reduziert werden.
3. Ich nutze die Remote-OCR und mein Dokument hat mehrere Seiten jedoch nicht mehr als 100 Seiten, trotzdem hängt das Dokument in der Verarbeitungskette fest.
   - Squeeze wartet insgesamt 3 Minuten auf die Verarbeitung des Dokumentes. Konnte der entfernte Dienst innerhalb dieser 3 Minuten das Dokument nicht verarbeiten, wird Squeeze eine Fehlermeldung mit einem Timeout Hinweis liefern. Schieben Sie das Dokument erneut über die technische Warteschlange in den Schritt "Texterkennung". Squeeze prüft in dem Fall, ob das bereits hochgeladene Dokument verarbeitet wurde.
4. Ich habe `OCRPageLimit` gesetzt, es werden aber trotzdem alle Seiten verarbeitet.
   - Prüfen Sie die eingestellte `OCREngine`: bei `maxocr` hat die Eigenschaft keine Wirkung, bei `squeeze` wird lediglich das Auslesen begrenzt.
   - Prüfen Sie, ob auf der Stapel- oder Dokumentenklasse ein `AttachmentBarcodePattern` konfiguriert ist – dieses hat Vorrang.
   - Prüfen Sie, ob die Eigenschaft auf der Dokumentenklasse mit einem abweichenden Wert gesetzt ist, da dieser den Wert der Stapelklasse übersteuert.

Revision #40
Created 2023-02-28 12:25:53 UTC by Vahdettin Balum
Updated 2026-11-02 10:08:05 UTC by Sofia Miletic