



DpuScan

Janich & Klass
Computertechnik GmbH



DpuScan 7

PlugIn Tesseract OCR

Referenzhandbuch

Copyrights

© 1997 bis 2024 Janich & Klass Computertechnik GmbH. Alle Rechte vorbehalten. Gedruckt in Deutschland. Die in dieser Dokumentation enthaltenen Informationen sind Eigentum der Janich & Klass Computertechnik GmbH. Ohne schriftliche Genehmigung der Janich & Klass Computertechnik GmbH begründen weder der Empfang noch der Besitz dieser Informationen irgendein Recht auf Reproduktion oder Veröffentlichung irgendwelcher Teile davon.

Warenzeichen

Alle Produktnamen und Logos sind Warenzeichen oder eingetragene Warenzeichen der jeweiligen Eigentümer.

Haftungsausschluss

Die Anweisungen und Beschreibungen in diesem Handbuch waren zum Druckzeitpunkt zutreffend. Wir behalten uns jedoch das Recht vor, sowohl Beschreibung als auch Produkt jederzeit ohne Benachrichtigung zu ändern. Nach dem derzeitigen Stand der Softwaretechnik ist es nicht möglich, Programme zu entwickeln, die unter allen Bedingungen in jeder Konfiguration fehlerfrei arbeiten. Die Janich & Klass Computertechnik GmbH übernimmt keinerlei Haftung für Defekte, die direkt oder indirekt durch Fehler dieses Handbuches, Weglassen von Informationen oder durch Unstimmigkeiten zwischen diesem Referenzhandbuch und dem Produkt entstanden sind.

Aktualität

Es ist möglich, dass im Internet eine neuere Version dieses Handbuches verfügbar ist. Wir empfehlen deshalb, die Version anhand des auf dieser Seite abgedruckten Datums mit der Version auf dem Internet zu vergleichen. Falls die Version im Internet neueren Datums ist, sollten Sie diese herunterladen und ggf. selbst ausdrucken.

Inhaltsverzeichnis

1 Übersicht	4
1.1 Konfiguration im Basisprofil	5
1.2 Konfiguration des Plugins	8
1.2.1 Konfiguration der Feldsuche	10
1.2.2 Konfiguration der Volltextsuche	12
1.2.3 Sprachauswahl	14
1.2.4 Migration von FineReader-Subprofilen	15
1.3 Konfiguration im Taskprofil	18
1.4 Konfiguration als Kommando	19
1.5 Anzeige und Rückgabe	21
1.6 Zusammenfassung	23

1 Übersicht

Mit dem PlugIn Tesseract OCR kann in DpuScan eine Texterkennung, engl. Optical Character Recognition, von Teilbereichen oder ganzen Textseiten durchgeführt werden.

Bei der Teilbereichssuche oder Feldsuche werden Ausschnitte in einem Bild durchsucht und die dort gefundenen Texte in DpuScan-Variablen zurückgegeben.

Die Volltextsuche untersucht alle Bereiche eines Bildes und kann die Ergebnisse formatiert ausgeben. Als Exportformat stehen PDF, Text, HTML und XML zur Verfügung.

Voraussetzungen für den Einsatz des PlugIns

Das PlugIn kann in allen lizenzierten Versionen von DpuScan ab Version 6.12 eingesetzt werden. Es ist keine zusätzlich Lizenz für dieses PlugIn erforderlich.

Das PlugIn basiert auf der Tesseract (Apache 2.0) Bibliothek. Die Installation der Bibliothek erfolgt automatisch bei der Installation von DpuScan 6.12 oder höher.

Migration von FineReader-Profilen

Die in vorherigen Versionen von DpuScan verwendete OCR von FineReader wird ab Version 6.12 nicht mehr zur Installation angeboten. Das PlugIn Tesseract OCR bietet für die Weiterverwendung solcher Konfigurationen (Subprofile) aber eine Migrationsfunktion an.

Funktionsweise des PlugIns

Das PlugIn wird beim Scannen oder stapelweisen Einlesen von Dokumenten nach dem Erfassen der Bilddaten aufgerufen und führt eine Texterkennung durch. Die Ergebnisse werden in Variablen oder speziellen OCR-Dateien zwischengespeichert. Die Variablen können von DpuScan zur Steuerung oder Ausgabe verwendet werden, die OCR-Dateien werden zusammen mit dem Bild in das gewünschte Format überführt.

Im interaktiven Modus, d.h. der Pause nach dem Scannen, in der die Bilder angezeigt werden, kann das PlugIn gezielt auf ein einzelnes Bild angewendet werden.

Um das PlugIn zu verwenden sind verschiedene Konfigurationsschritte erforderlich:

[Konfiguration im Basisprofil](#)

[Konfiguration des PlugIn](#)

[Feldsuche](#)

[Volltextsuche](#)

[Migration](#)

[Konfiguration im Taskprofil](#)

[Konfiguration als Kommando](#)

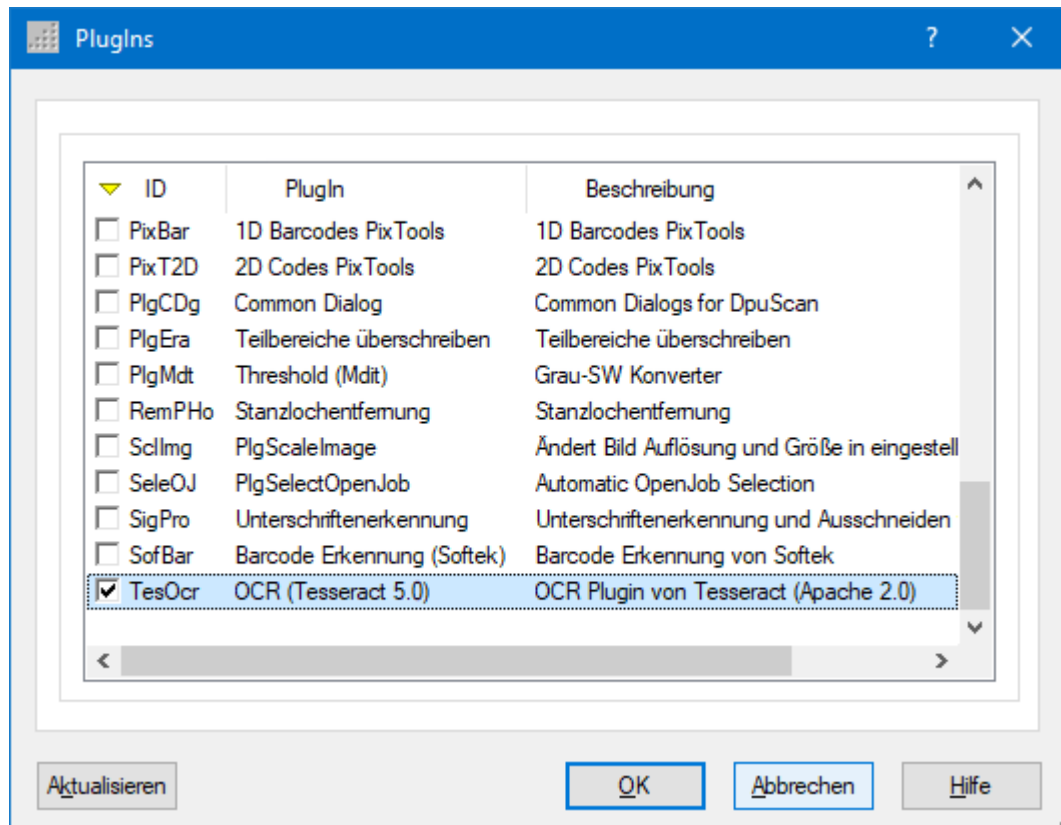
[Anzeige und Rückgabewerte](#)

[Zusammenfassung](#)

1.1 Konfiguration im Basisprofil

Das PlugIn ist innerhalb des Basisprofils zu laden und zu konfigurieren. Öffnen sie dazu die **Basisprofilkonfiguration**, wählen Sie dort die Registerkarte **Prozess** und klicken Sie auf die Schaltfläche **Plugins**.

Über die Schaltfläche Hinzufügen gelangen Sie zu der Auswahl der verfügbaren PlugIns.

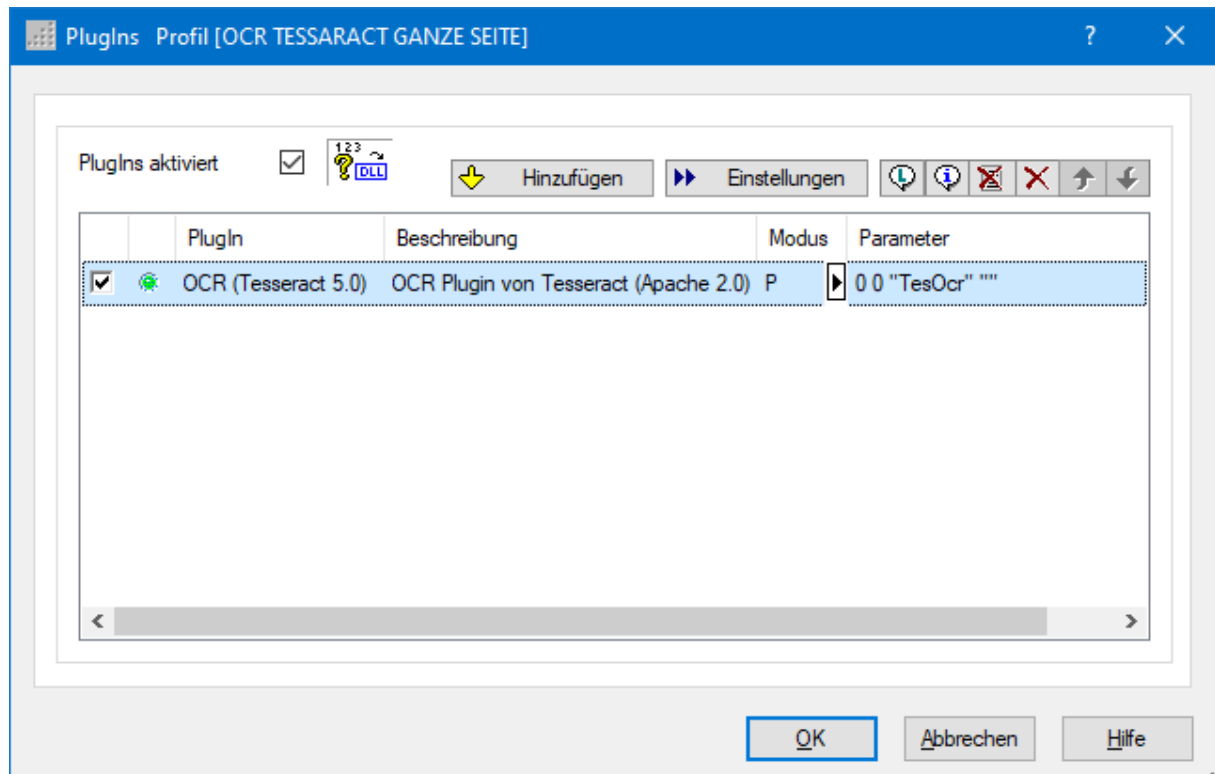


Auswahl des PlugIns

Wählen Sie die ID "TesOcr" für das PlugIn "OCR (Tesseract 5.0)" aus und bestätigen Sie mit OK. Das PlugIn wird nun für die Verwendung innerhalb des Basisprofil geladen.

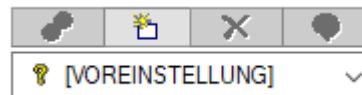
Bitte stellen Sie sicher, dass das Kontrollkästchen „PlugIns aktiviert“ mit einem Haken markiert ist, da ansonsten die PlugIns nicht verwendet werden. Das Kontrollkästchen kann erst aktiviert werden, wenn mindestens ein PlugIn geladen wurde.

Der grüne Punkt in der Liste der geladenen PlugIns zeigt an, dass das PlugIn einsatzbereit ist. Der Eintrag in der Spalte Modus zeigt die Zeichenfolge P. Das bedeutet, dass dieses PlugIn im Prozessmodus arbeitet.



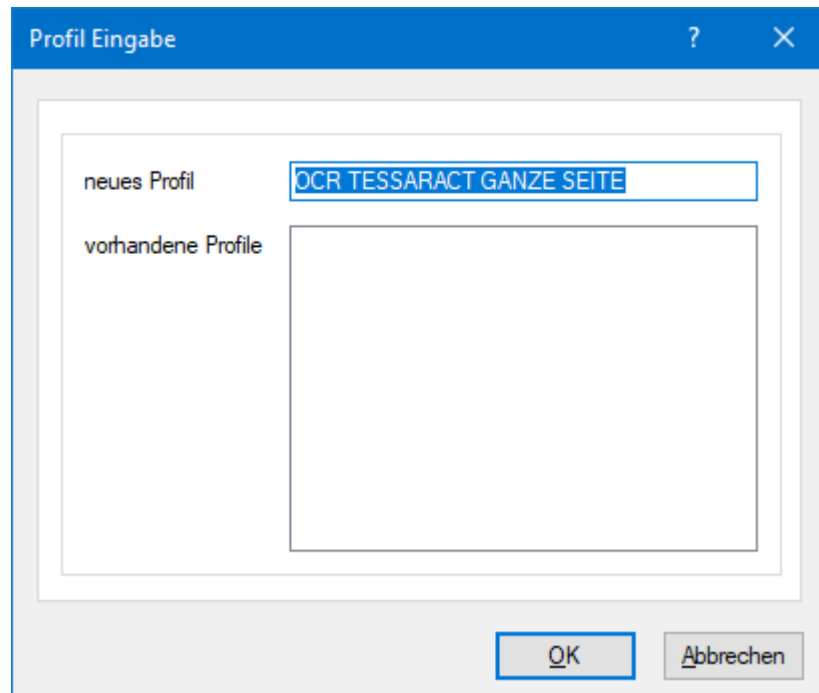
PlugIn in der Liste geladener PlugIns

Erzeugen Sie nun eine Konfiguration, indem Sie doppelt in die Zelle zur Spalte "Parameter" klicken. Es öffnet sich der Dialog für den Aufruf des PlugIns als Brokerereignis oder beim Selektionswechsel. Dort gibt es eine Reihe von Schaltflächen:



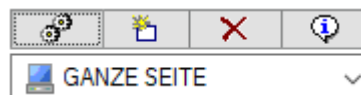
Anlegen eines neuen Subprofils

Die einzige verfügbare Schaltfläche ist die Neu-Taste; erstellen Sie damit ein neues Subprofil. Dabei werden Sie aufgefordert, einen Namen für die zu erstellende Konfiguration anzugeben.



Eingabe eines neuen Subprofils

Nach dem Anlegen des neuen Subprofils stehen nun die anderen Tasten zur Verfügung:



Verwalten von Subprofilen

Klicken Sie auf die Ändern-Taste mit den Rädchen. Für die eigentliche [Konfiguration des Plugins](#) steht ein eigener Dialog zur Verfügung. Nachdem Sie dort die Einstellungen vorgenommen haben, können Sie alle offenen Dialoge mit OK verlassen.

1.2 Konfiguration des PlugIns

Der Konfigurationsdialog für das PlugIn ist geteilt und zeigt links die Steuerelemente zum Einstellen der Suchparameter. Stellen Sie hier ein, wonach das PlugIn suchen soll.

Taskmodus

Gibt an, wie das PlugIn arbeiten soll:

Feldsuche: Jeder gesetzte Rahmen wird einzeln bearbeitet und das Ergebnis in zugeordneten Programm-Variable (%-Code) gespeichert.

Volltextsuche: Die ganze Seite wird erkannt und das Ergebnis in eine Datei gespeichert.

Auf der rechten Seite ist ein **Vorschaufenster**

In diesem Fenster werden Dateien angezeigt, um die eingestellten Parameter zu testen. Durch Halten und Ziehen der linken Maustaste kann der Bildausschnitt bewegt werden. Mit dem Mausekranz kann die Ansicht vergrößert bzw. verkleinert werden. Die rechte Maustaste hat hier keine Funktion.

Das Fenster akzeptiert Dateien, die mit Maus darauf gezogen werden. Bilddateien werden angezeigt, andere Dateien werden ignoriert.

Bitte beachten Sie, dass nur reine Bilddateien angezeigt werden können, Dateien mit gemischten Inhalten, z.B. durchsuchbares PDF, können nicht angezeigt werden.

Oberhalb des Vorschaufensters befinden sich diese Steuerelemente.



Datei öffnen

Lädt eine Datei von der Festplatte und stellt das Bild im Fenster dar. Es können alle gängigen Bildformate geladen werden, wie zum Beispiel Jpeg, Tiff, Bmp Bilder.



Rahmen hinzufügen

Erzeugt einen Rahmen auf dem geladenen Bild. Die Größe und die Position des Rahmens können noch im Nachhinein geändert werden.



Rahmen hinzufügen

Löscht den gerade aktiven Rahmen.



Scannen

Holt ein Bild vom Scanner



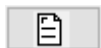
Scannereinstellungen

Öffnet die Scannereinstellungen



Test Feldsuche

Führt eine vollständige Erkennung aus und speichert das [Ergebnis in eine Datei](#).



Test Volltextsuche

Führt eine Suche nach Texten in allen angegebenen Rahmen durch und zeigt die [Ergebnisse in einem Dialog an](#).

Unten auf der linken Seite kann das Werkzeug zur Übernahme von FineReader-Konfigurationen gestartet werden.

Migration

Öffnet den [Dialog](#) zur Umwandlung von FineReader-Subprofilen.

Unten befinden sich die gewohnten Steuerelemente.

OK

Schließt die Dialogbox und speichert alle eingestellten Suchparameter.

Abbrechen

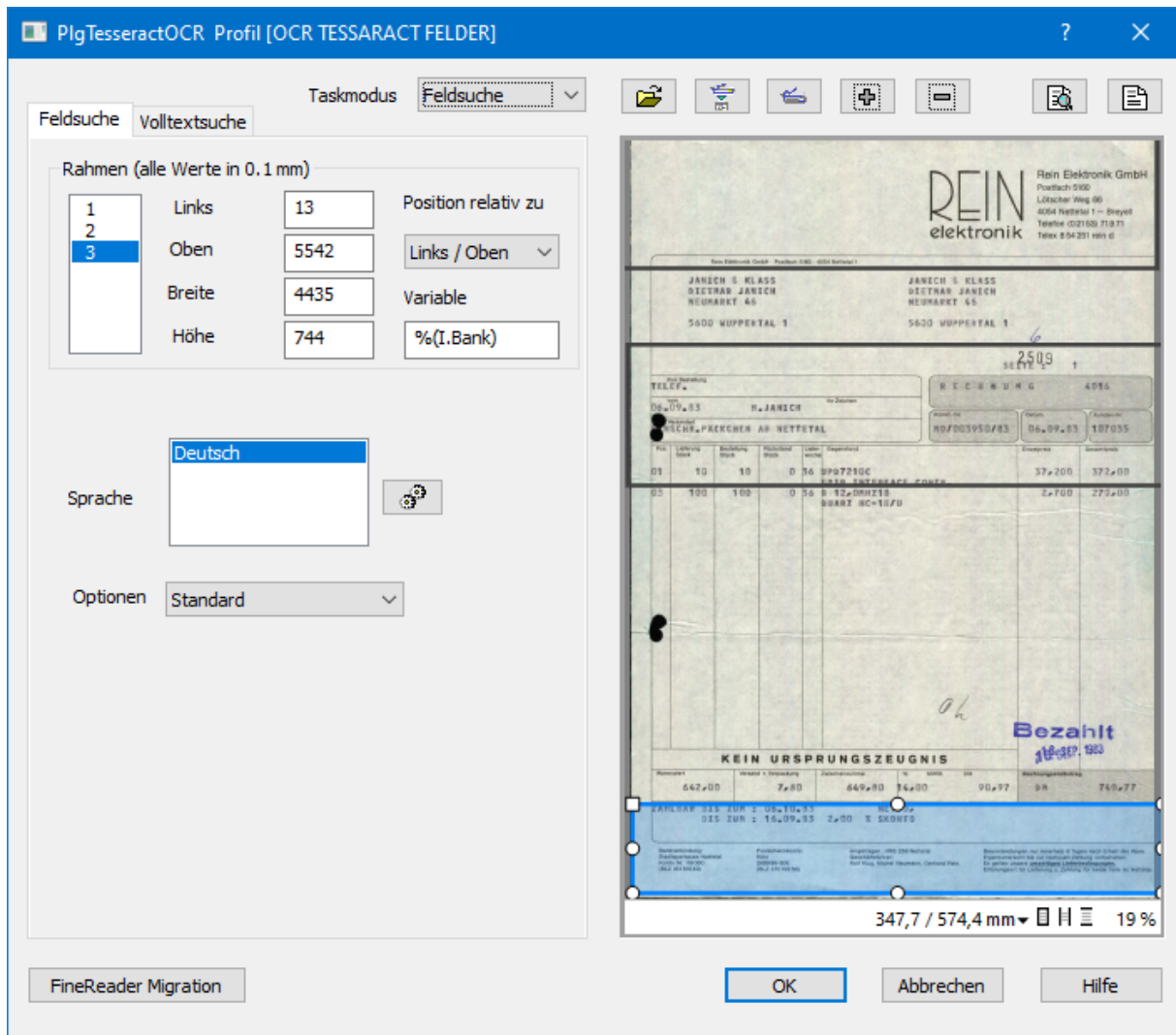
Verwirft alle Einstellungen und schließt die Dialogbox.

Hilfe

Öffnet die Hilfedatei zum PlugIn Tesseract OCR.

1.2.1 Konfiguration der Feldsuche

Bei der Feldsuche kann in einem oder mehreren Bereichen auf dem Bild eine Texterkennung stattfinden. Die Ergebnisse können einzelnen Programm-Variablen (%-Codes) zugeordnet werden um z.B. den weiteren Arbeitsablauf oder die Ablage zu steuern.



Suchparameter für die Felssuche

Suchparameter

Die Einstellungen zur Suche können dann auf der linken Seite vorgenommen werden:

Liste der Rahmen

Zeigt alle in dieser Konfiguration definierten Rahmen an. Ein Klicken auf den entsprechenden Eintrag in diesem Listefeld aktiviert den Rahmen und zeigt seine Suchparameter an.

Links/Oben

Eine Eingabe in dieses Eingabefeld verändert die linke Position des gerade aktiven Rahmens.

Die Änderung wird sofort aktiv und wird im Vorschaufenster angezeigt.

Breite/Höhe

Eine Eingabe in dieses Eingabefeld verändert die Breite gerade aktiven Rahmens.

Position relativ zu

Die Änderung wird sofort aktiv und wird im Vorschaufenster angezeigt.

Hier kann die Erkennung an den Texttyp angepasst werden:

Wählen Sie **Zahlen**, wenn in diesem Feld nur numerische Werte zu erwarten sind.

Wählen Sie **Tabelle**, wenn die Daten in einem rechteckigen Schema angeordnet sind.

In allen anderen Fällen liefert **Standard** die besten Ergebnisse.

Sprache

Auswahl und Reihenfolge der Wörterbücher, die bei der Suche verwendet werden sollen.

Sprachauswahl

Öffnet den Dialog zur [Auswahl der Wörterbücher](#) die verwendet werden sollen.

Optionen

Hier kann die Erkennung an den Texttyp angepasst werden:

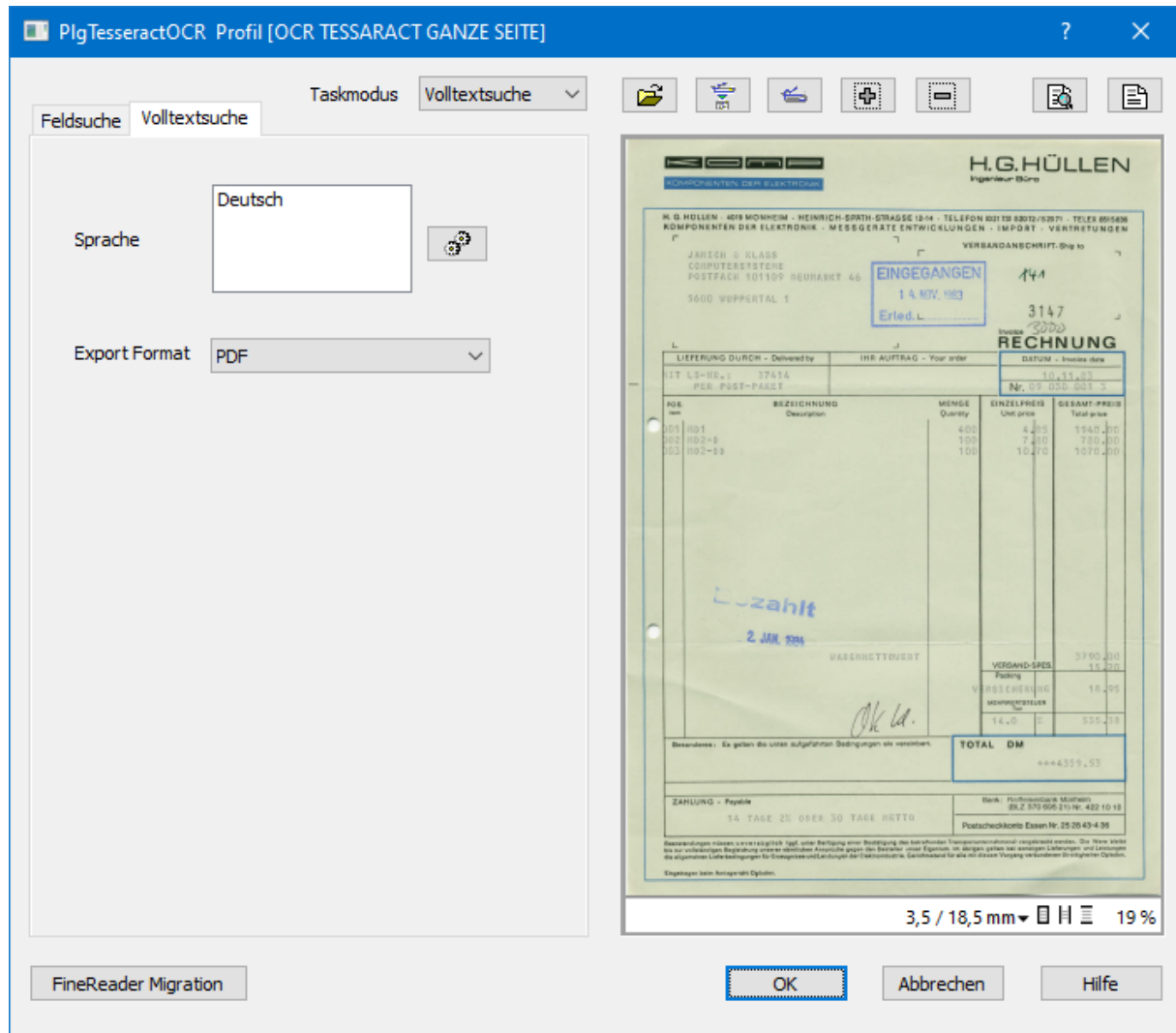
Wählen Sie **Zahlen**, wenn in diesem Feld nur numerische Werte zu erwarten sind.

Wählen Sie **Tabelle**, wenn die Daten in einem rechteckigen Schema angeordnet sind.

In allen anderen Fällen liefert **Standard** die besten Ergebnisse.

1.2.2 Konfiguration der Volltextsuche

Bei der Volltextsuche wird das gesamte Bild analysiert und in einen Text umgewandelt. Je nach gewähltem Speicherformat kann das Bild zusammen mit dem Text angezeigt werden.



Volltextsuche

Suchparameter

Sprache

Auswahl und Reihenfolge der Wörterbücher, die bei der Suche verwendet werden sollen.

Sprachauswahl

Öffnet den Dialog zur [Auswahl der Wörterbücher](#) die verwendet werden sollen.

Exportformat

Legen Sie hier fest, in welchem Format der Text, bzw. der Text und das Bild, gespeichert werden sollen.

PDF "durchsuchbares PDF", d.h. das Bild wird angezeigt, die Texte können markiert und kopiert werden.

Text Erzeugt eine einfache Textdatei mit Seitenumbrüchen, falls mehrere Bilder in einer Datei gespeichert werden

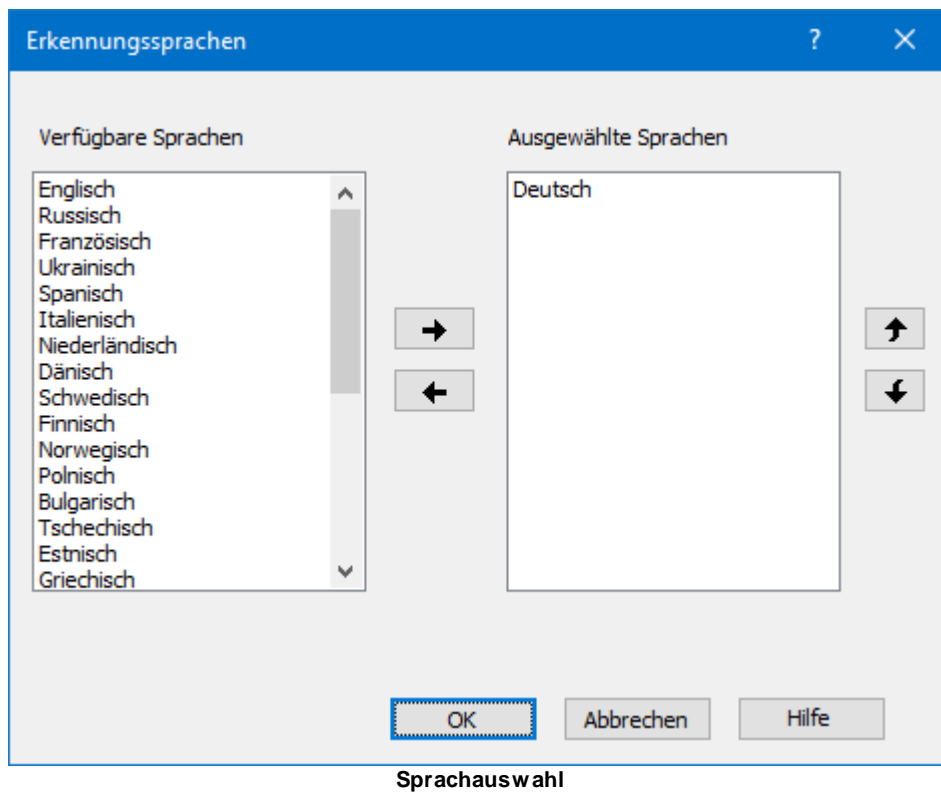
sollen.

HTML Erzeugt eine HTML-Datei.

XML Erzeugt eine XML-Datei.

1.2.3 Sprachauswahl

In diesem Dialog können die Sprachen angegeben werden und die Reihenfolge in welcher sie berücksichtigt werden sollen. Die Tesseract OCR Engine verwendet die verfügbaren Wörterbücher in dieser Reihenfolge um ein Wort zu identifizieren.



Die Steuerelemente dienen der Auswahl, bzw. der Sortierung der Sprachen:

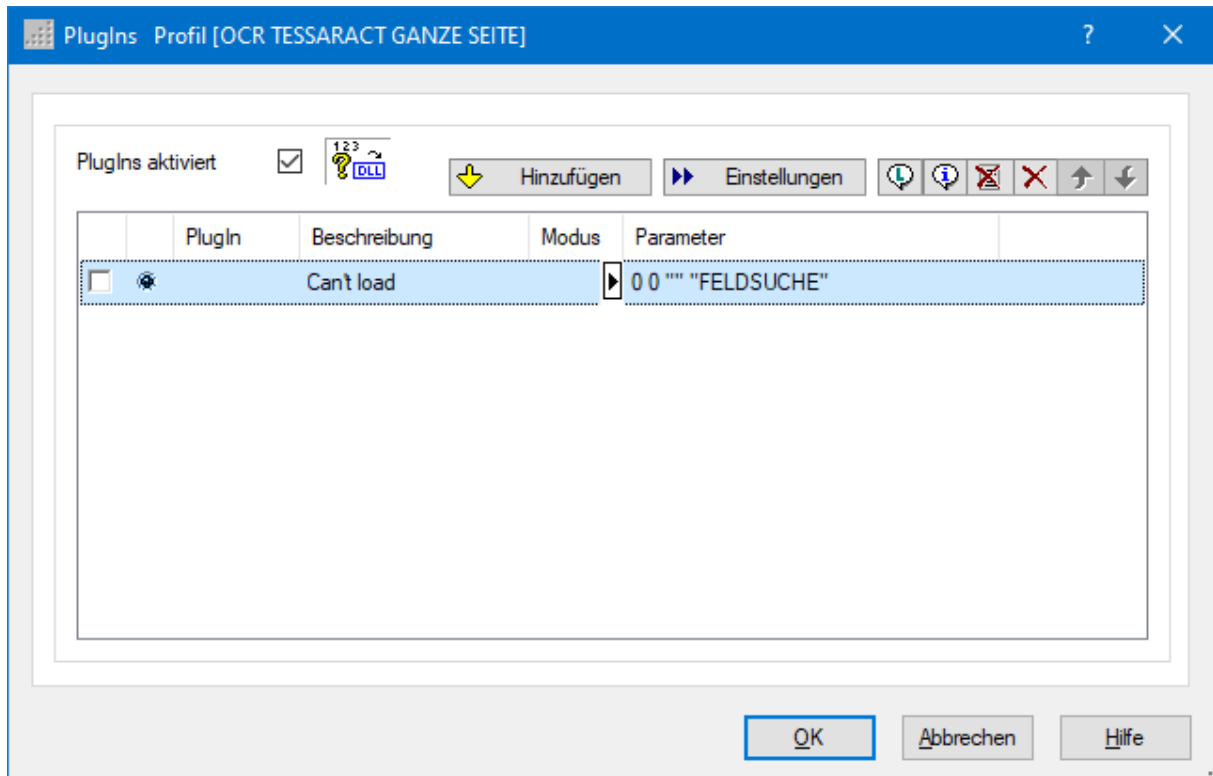
- **nach rechts** Fügt das Wörterbuch hinzu.
- ← **nach links** Entfernt das Wörterbuch.
- ↑ **nach oben** Verschiebt das Wörterbuch in der Suchreihenfolge nach oben.
- ↓ **nach unten** Verschiebt das Wörterbuch in der Suchreihenfolge nach unten.

Unten befinden sich die gewohnten Steuerelemente.

- OK** Schließt die Dialogbox und speichert alle eingestellten Suchparameter.
- Abbrechen** Verwirft alle Einstellungen und schließt die Dialogbox.
- Hilfe** Öffnet die Hilfedatei zum PlugIn Tesseract OCR.

1.2.4 Migration von FIneReader-Subprofilen

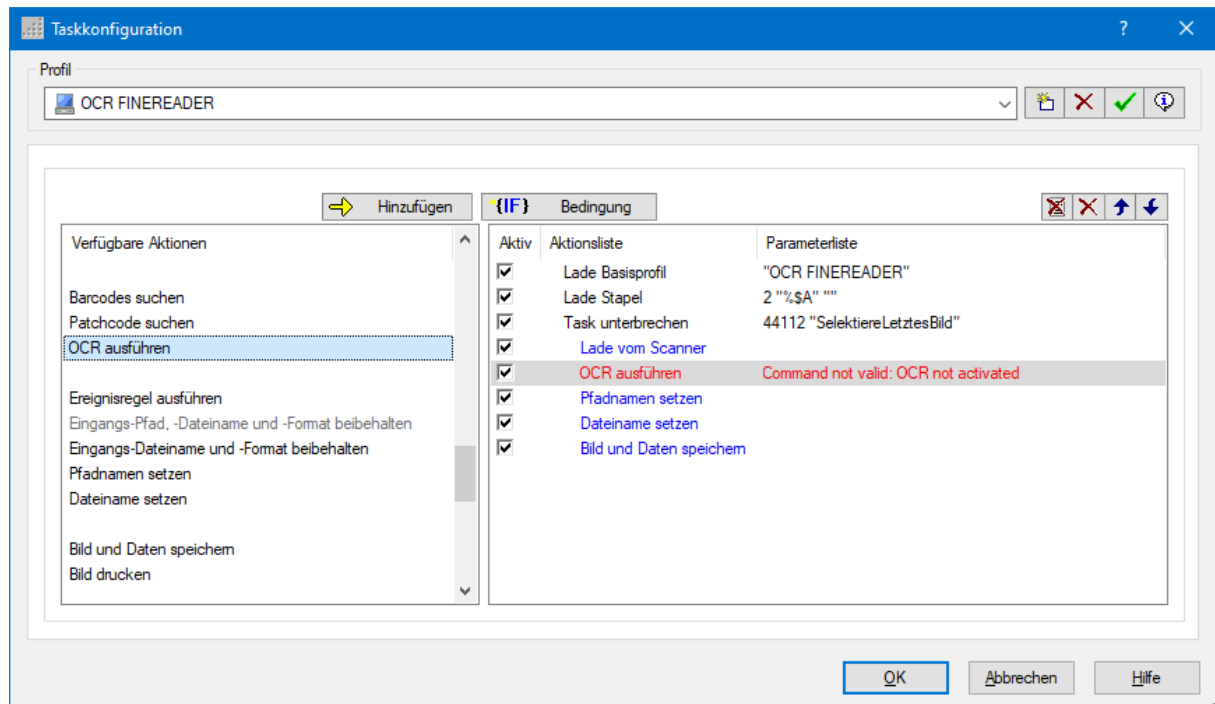
Mit dem Umstieg von Version 6.11 auf Version 6.12 des Scanprogramms, ist die Weiterverwendung der FineReader-Engine entfallen. Bei einem Import eines alten Profils erscheint in der Basisprofil-Konfiguration in der Liste der geladenen PlugIns diese Meldung:



Laden des FineReader-PlugIns nicht möglich

Verwenden Sie hier stattdessen das PlgTesseractOCR.

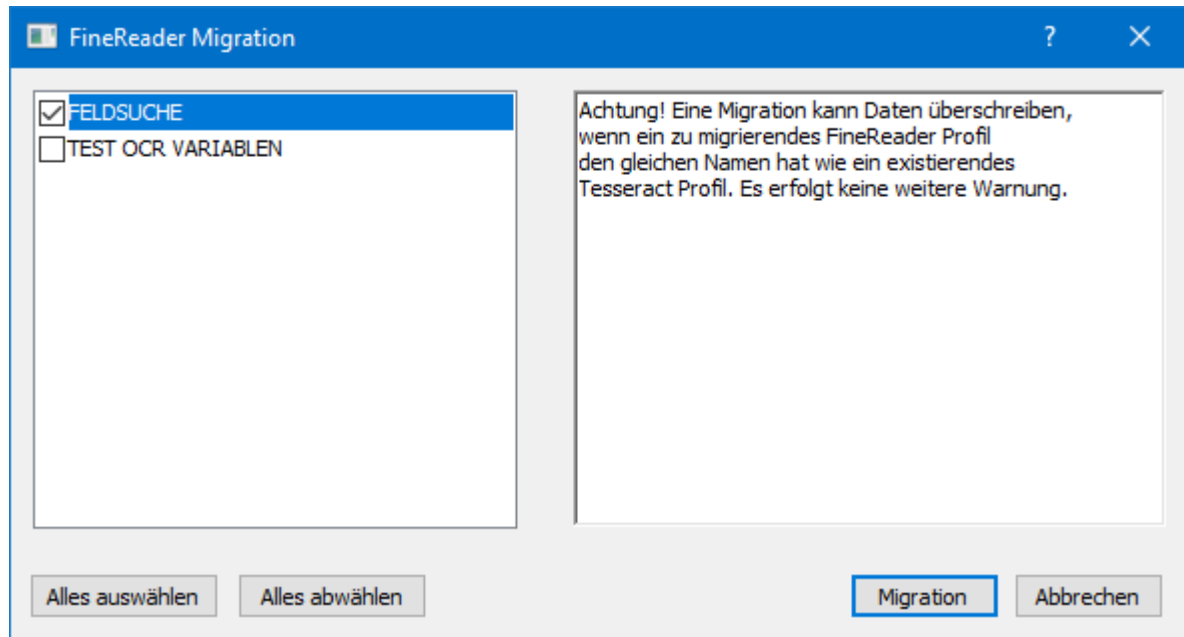
Im Taskprofil wird ein alter Aufruf des FineReader-PlugIns so angezeigt.



Verwendung alter OCR-Konfiguration nicht möglich

Verwenden Sie hier stattdessen die Anweisung "Plugin für jedes Bild aufrufen" mit dem **PlgTesseractOCR**.

Die Funktion **Migration** des PlgTesseractOCR ermöglicht die Umwandlung der Einstellungen in das Tesseract OCR Format. Es öffnet sich ein Dialog mit den Subprofilen, die migriert werden könnten:



Migration der FineReader-Profile

Bei der Migration wird ein neues gleichnamiges Tesseract-Profil erzeugt und die Einstellungen nach Möglichkeit übernommen. Die Rahmen bleiben stets erhalten, die Sprachauswahl in den meisten Fällen ebenfalls.

Es gibt keine Möglichkeit, ein FineReader-Konfiguration aus einer Tesseract-Konfiguration zu erzeugen.

Die Profilauswahl kann mit diesen Steuerelementen gesetzt werden:

Liste der FineReader-Subprofile

Liste der vorhandenen FineReader-Profile. Diese Liste sollte nach dem Update auf Version 6.12 oder höher und jedes Mal nach dem Import einer älteren Sicherung geprüft werden.

Warnhinweis

Wie angezeigt werden bereits vorhandene Tesseract-Profile überschrieben, wenn ein gleichnamiges FineReader-Profil migriert wird. Im Zweifelsfall sollten Sie vor der Migration eine Sicherung aller aktuelle Tesseract-Profile mit dem DpuEnterpriseManager durchführen.

Alles auswählen

Wählt alle Einträge aus.

Alles abwählen

Hebt die Auswahl auf.

Migration

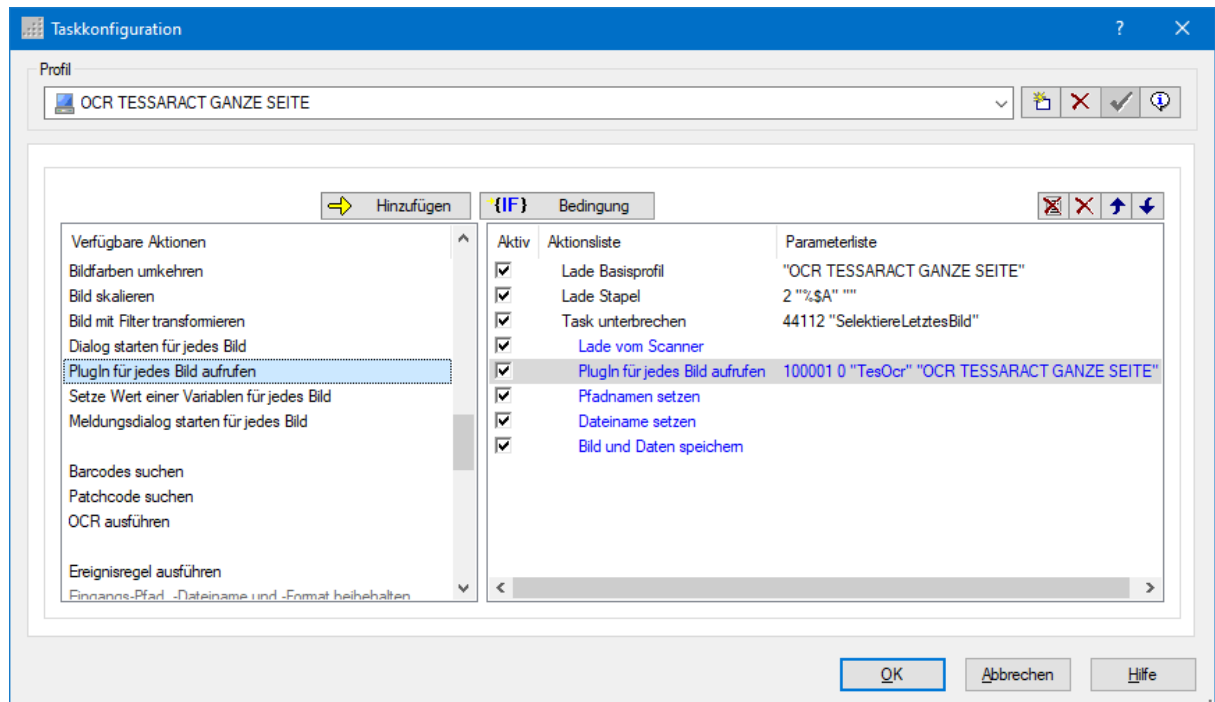
Führt die Umwandlung der FineReader-Profile in Tesseract-Profile aus.

Abbrechen

Verwirft alle Einstellungen und schließt die Dialogbox.

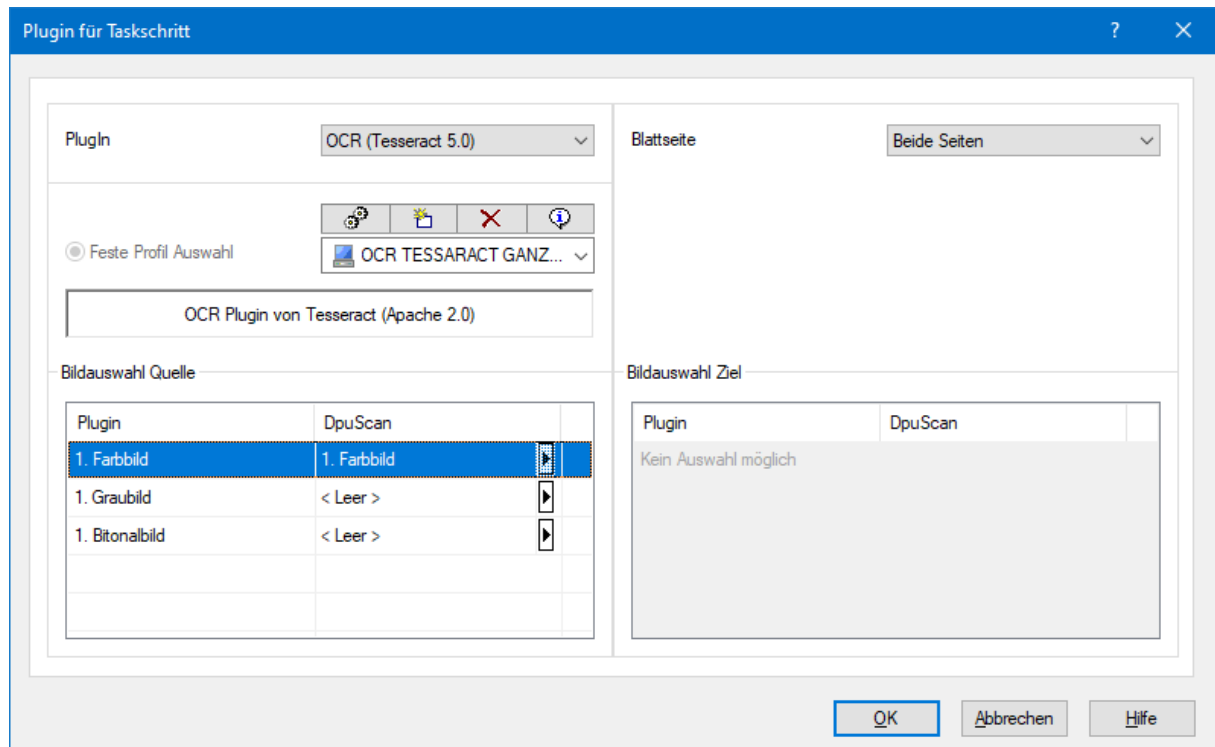
1.3 Konfiguration im Taskprofil

Im Taskprofil, d.h. in der Liste der Arbeitsanweisungen, kann das PlugIn mit dem Schritt "PlugIn aufrufen für jedes Bild" eingefügt werden. Achten Sie darauf, dass dieser Schritt nach dem Erfassen des Bildes, hier "Lade vom Scanner", erfolgt. Wenn Sie das Suchergebnis zur Steuerung des Ablaufes brauchen, so muss er vor "Ereignisregeln ausführen" stehen.



Aufruf des PlugIn im Taskprofil

Da dieses PlugIn mit Bildern arbeitet, müssen Sie dabei angeben, mit welchen Bildern gearbeitet werden soll:



Plugin für Taskschritt

Im Dialog links unten erfolgt die Zuordnung der Bilder. Das Plugin kann pro Aufruf jeweils nur ein Farb-, Grau- oder Schwarzweißbild bearbeiten. Tragen Sie auf der DpuScan-Seite ein, welches Bild übergeben werden soll. Tragen Sie die Bildtypen ein welche der Scanner liefert, im Beispiel ist dies nur das Farbbild.

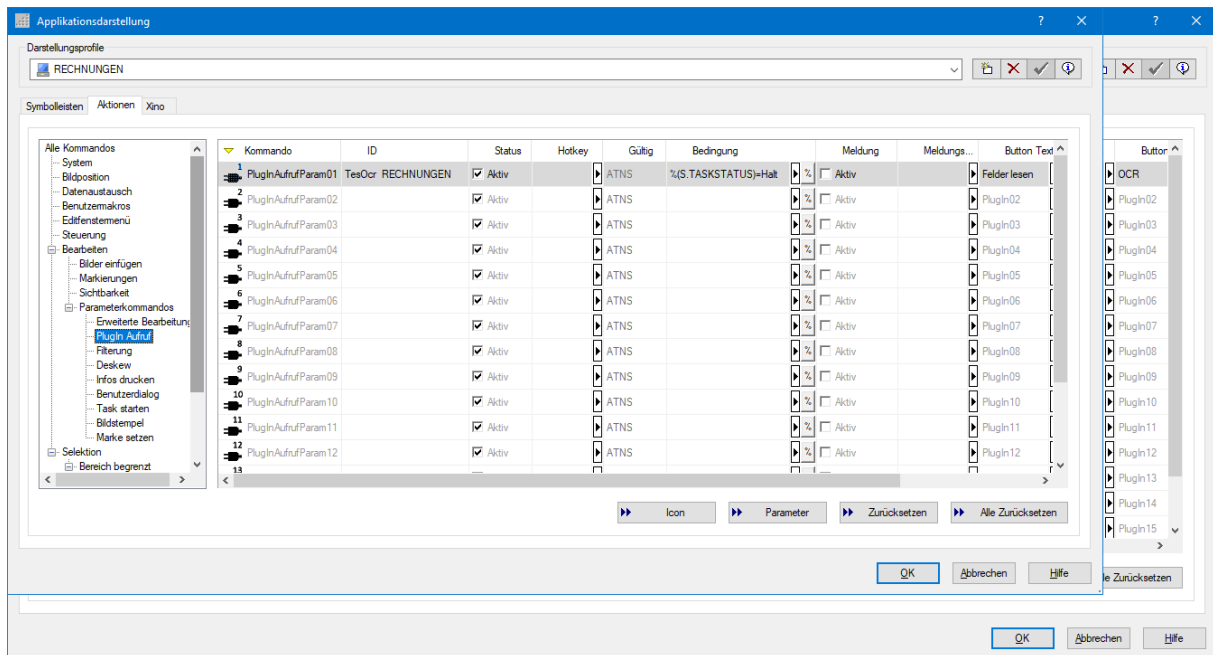
Die Nummer des Bildes gibt *nicht* die Position im Stapel an, sondern die Position innerhalb einer Bildgruppe. In den meisten Fällen muss das 1. Bild verwendet werden.

Die meisten Erkennungsprogramme arbeiten mit Schwarzweißbildern. Wenn der Scanner Farb- und Schwarzweißbildern liefert, sollten Sie für die Erkennung das Schwarzweißbild als "Bitonalbild" übergeben.

Rechts oben im Dialog kann man die Suche auf die Vorderseite einschränken, falls auf den Rückseiten nicht gesucht werden soll.

1.4 Konfiguration als Kommando

Das Plugin kann auch gezielt auf ein ausgewähltes Bild angewendet werden. Öffnen Sie dazu die Applikationsdarstellung und gehen Sie zu "Aktionen". Wählen Sie dort links in der Baumansicht den Zweig Bearbeiten -> Parameterkommandos -> Plugin Aufruf.



PlugIn Aufruf als Parameterkommando

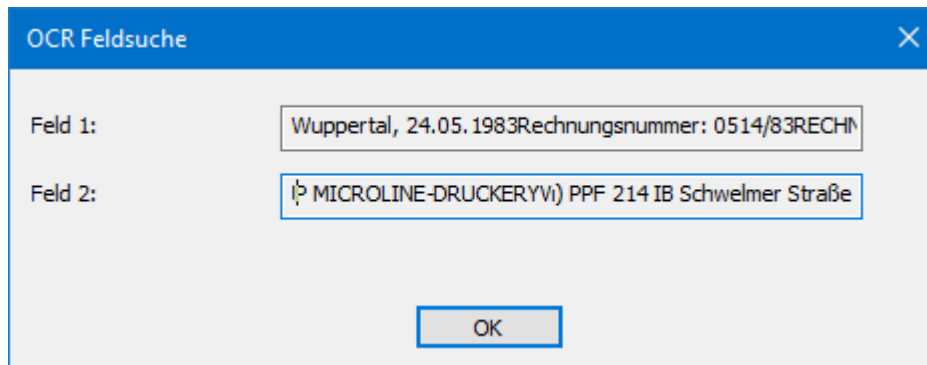
Ein Klick auf Parameter oder ein Doppelklick auf die Spalte Kommando öffnet den [bekannten Dialog](#) zur Auswahl von PlugIn, Subprofil und den zu übergebenden Bildern.

Nach der Angabe dieser Werte können Sie noch ein Symbolbild ein Tastaturkürzel und verschiedene Beschriftungen vergeben und die Schaltfläche auf der Symbolleiste platzieren. Wird das Tastaturkürzel eingegeben oder diese Schaltfläche gedrückt, so wird das PlugIn aufgerufen und die [Suchergebnisse](#) aktualisiert.

Wenn die Suche als **Makro**, d.h. als Teil einer Folge von Anweisungen, erfolgen soll, so wählen Sie Im Baum Benutzermakros und fügen den PlugIn-Aufruf als Kommando ein. Im selben Makro können Sie dann z.B. einen Merker setzen, der festhält, dass vor dem Finalisieren ein Neuaufbau des Stapels erfolgen soll.

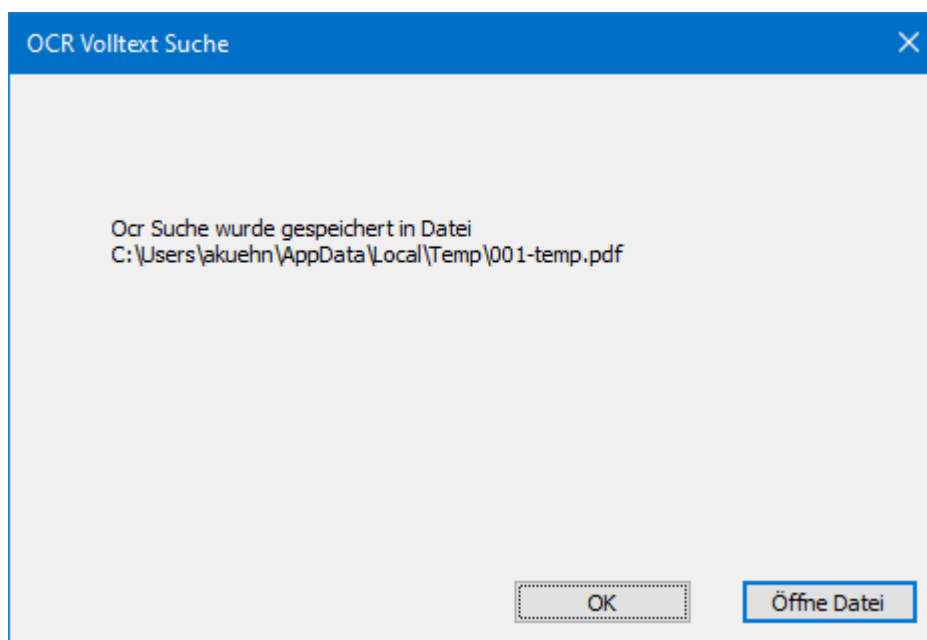
1.5 Anzeige und Rückgabe

Das PlugIn zeigt im laufenden Betrieb keine eigenen Fenster an. Nur wenn im Konfigurationsdialog die Suche ausgeführt wird, werden Ergebnisse als Liste angezeigt:



Anzeige des Ergebnisses bei der Feldsuche

Bei der Volltextsuche wird nur angegeben, wohin die Ergebnisse geschrieben wurden:



Anzeige des Ergebnisses bei der Volltextsuche

Diese Variablen werden zurückgegeben:

`%(S.OCR1n)`

Gefundener Text im Rahmen **n**. Diese Variable kann beliebig ersetzt werden, verwenden Sie vorzugsweise den Gültigkeitsbereich für Bildvariablen `%(I.xxx)` oder temporäre Variablen `%(V.xxx)`

%(I . IMAGE . XMLFILE)

Liefert den temporären Dateinamen mit den OCR-Ergebnissen für die Weiterverwendung im PlugIn Classify.

%(S . OCRFILE)

Liefert den Namen der Datei, in welcher die Erkennungsergebnisse gespeichert werden.

1.6 Zusammenfassung

Name des PlugIns	PlgTesseractOCR
Beschreibung	Erkennt Texte mit der Tesseract OCR-Engine
Stand	14.08.2024
DpuScan	Version 6.12 und höher
PlugIn Dateien	PlgTesseractOCR.dll, PlgTesseractOCR_07.Ing
Zusätzliche Engine	Tesseract (Apache 2.0)
Kostenpflichtig	Nein
Kann als Taskschritt verwendet werden	Ja
Kann als Makro-Kommando verwendet werden	Ja
Kann ein Fenster anzeigen	Nein
Reagiert auf Brokerereignisse	Nein
Reagiert auf Selektionswechsel	Nein
Eingangsvariablen	
keine	
Ausgangsvariablen	
%(S . OCR1 n)	Gefundener Text im Rahmen n . Diese Variable kann beliebig ersetzt werden, verwenden Sie vorzugsweise den Gültigkeitsbereich für Bildvariablen %(I.xxx) oder temporäre Variablen %(V.xxx)
%(I . IMAGE . XMLFILE)	Liefert den temporären Dateinamen mit den OCR-Ergebnissen für die Weiterverwendung im PlugIn Classify.
%(S . OCRFILE)	Liefert den Namen der Datei, in welcher die Erkennungsergebnisse gespeichert werden.

Index

- A -

Anzeige 21
Aufruf als Parameterkommando 19
Auswahl des Plugins 5

- B -

Bearbeiten eines Subprofils 5
Bilder auswählen 18

- D -

Deutsch 14

- E -

Englisch 14
Erzeugen eines Subprofils 5

- F -

Feldsuche 8
Festlegen des Taskmodus 8
Französisch 14

- K -

Konfiguration als Kommando 19
Konfiguration im Basisprofil 5
Konfiguration im Taskprofil 18

- M -

Migration 8

- R -

Reihenfolge der Sprachen 14
Rückgabewerte 21
Rückseiten 18

- S -

Schritt im Makro 19

- T -

Taskschritt 18

- U -

Übersicht 4

- V -

Volltextsuche 8
Vorschaufenster 8

- W -

weitere Sprachen 14
Wörterbücher 14

- Z -

Zusammenfassung 23