

We capture the future.

Janich & Klass

D P U

scanner value pack



PlugIn zu DpuScan

# Klassifizierung

Ergänzung zum DpuScan Referenzhandbuch

## Copyrights

© 1997 bis 2011 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

Das DPU Logo ist eingetragenes Warenzeichen der Janich & Klass Computertechnik GmbH.

DpuScan ist Warenzeichen von J&K Imaging, Marietta/USA. Alle anderen 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 und 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 Handbuch und dem Produkt entstanden sind.

## Aktualität

Es ist möglich, dass im Internet eine neuere Version dieser Dokumentation zum DpuScan verfügbar ist. Wir empfehlen deshalb, die Version an Hand 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.

Die aktuelle Version dieses Anhangs zum DpuScan Referenzhandbuch finden Sie im Web unter:

<http://www.jkimaging.com/pdf/PlugIns/Klassifizierung.pdf>

© 2011 Janich & Klass Computertechnik GmbH, Wuppertal, Germany

21. März 2011

## Inhaltsverzeichnis

1	Übersicht.....	4
2	Klassen .....	5
3	Feld-Typen für die Klassifizierung .....	6
3.1	Klassifizierungsbedingung.....	12
3.2	Barcode .....	12
3.3	Patchcode.....	12
4	Feld-Typen für die Datenextraktion .....	13
4.1	Textfeld.....	13
4.2	4.2 Listenfelder .....	14
4.3	4.3 Relative Suche .....	16
5	Konfiguration des Plugins .....	6
6	Verwendung des Plugins in DpuScan .....	17
7	Reguläre Ausdrücke .....	18
7.1	7.1 Syntax.....	18
7.2	7.2 Beispiele für reguläre Ausdrücke.....	22

# 1 Übersicht

Das vorliegende PlugIn dient der Klassifikation von Bildern, die von DpuScan bereitgestellt werden.

Ziel ist es, ein Bild einer Dokumentenklasse zuzuordnen.

Dokumentklassen werden im PlugIn durch ein oder mehrere Klassifikationsbedingungen definiert. Eine solche Definition vergleicht die OCR-Ergebnisse eines bestimmten Bildbereiches mit hinterlegten Daten. Diese Daten können feste Texte oder auch reguläre Ausdrücke sein.

Neben der Klassifizierung kann das PlugIn außerdem noch Daten aus einem Bild extrahieren und diese mit benutzerdefinierten Variablen im weiteren Prozess für DpuScan verfügbar machen. Die Feldtypen sind daher in zwei Gruppen zu unterteilen.

- Klassifizierungsfelder und
- Extraktionsfelder.

Für jede zu erkennende Klasse lassen sich ein oder mehrere Klassifizierungsbedingungen definieren, die logisch miteinander verknüpft werden können.

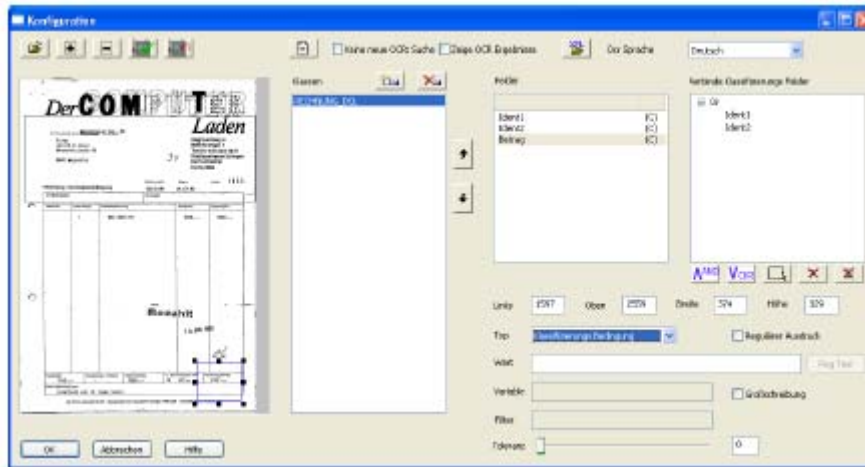
## 2 Klassen

Als Dokumentklasse wird ein Satz von logisch miteinander verknüpften Merkmalen bezeichnet. Die Merkmale beziehen sich auf die Ergebnisse einer zuvor ausgeführten OCR. So können beispielsweise Schlüsselworte, wie Rechnung, Lieferschein etc. in bestimmten Bereichen auf Belegen gesucht werden, um die Klasse eines Dokumentes zu bestimmen. Um die Erkennungssicherheit zu erhöhen, können diese Merkmale logisch miteinander verknüpft werden.

Eine weit öfter genutzte Möglichkeit ist es, alle Rechnungen eines Lieferanten als zu einer Klasse zugehörig anzusehen, weil auf diesen zu extrahierende Inhalte fest definierten Positionen auf dem Bild zu finden sind.

So dienen beispielsweise der Absender und das Schlüsselwort Rechnung zur Identifikation eines Beleges. Nach der Identifikation können dann die benötigten Daten an bekannten Stellen des Blattes ausgelesen werden, beispielsweise der Rechnungsbetrag oder die Rechnungsnummer.

### 3 Konfiguration des PlugIns



*Einstelldialog*



Lädt eine Datei von der Festplatte. Es werden alle gängigen Dateiformate unterstützt.



Erzeugt einen neuen Rahmen. Sie können den Rahmen in seiner Größe verändern oder an eine andere Position schieben. Halten Sie die linke Maustaste gedrückt während Sie sich im Rahmen befinden, um den Rahmen zu verschieben. Wenn Sie auf den Rand des Rahmens klicken, können Sie die Größe des Rahmens verändern.



Löscht das aktive Feld aus dem Vorschauenfenster und aus der Felder Liste.

Achtung: Alle definierten Operationen werden gelöscht, wenn ein Feld gelöscht wird, unabhängig davon, ob dieses gelöschte Feld Bestandteil einer Operation war.



Verbindet das aktuelle Bild mit dem gerade aktiven Rahmen. Dieses Bild kann erneut eingelesen werden, wenn Sie die nächste Schaltfläche drücken.



Liest das mit dem aktiven Rahmen verbundene Bild wieder ein. Sie können auf diese Weise das Bild wieder herstellen, mit dem Sie den gerade aktiven Rahmen definiert haben. Das ist nützlich zur Überprüfung der Definition.

Vorschauenfenster

Zeigt ein Dokument an, das mit Datei laden ausgewählt werden kann. Es werden alle üblichen Bildformate unterstützt. Mit den Maustasten lässt sich das Bild zoomen, mit der rechten Maustaste kann es verschoben werden, wenn das Bild größer als das Fenster ist. Rahmen, die mit Feld hinzufügen erzeugt wurden, werden ebenfalls angezeigt. Sie lassen sich mit der Maus verschieben oder in der Größe verändern.



Führt einen Klassifizierungstest für das im Vorschaufenster angezeigte Bild durch. Das Ergebnis des Tests wird in einer Message Box angezeigt.



Ein Klassifizierungstest kann einige Zeit in Anspruch nehmen, da eine OCR Volltextsuche Bestandteil davon ist.

Normalerweise ändert sich die Volltextsuche aber nicht, nachdem sie bereits einmal ausgeführt wurde. Wenn zu Testzwecken noch Änderungen an der Felddefinition durchgeführt werden, kann es sinnvoll sein, den Test zu verkürzen, indem das Ergebnis der letzten OCR Suche erneut verwendet wird. Aktivieren Sie in diesem Fall diese Schaltfläche.

Wenn sich das Bild im Vorschaufenster ändert muss, diese Option aber unbedingt wieder ausgeschaltet werden, da ansonsten der Test ein falsches Ergebnis suggeriert.



Wird diese Checkbox aktiviert und der Cursor im Vorschaufenster über den selektierte Rahmen gezogen, wird der Inhalt der OCR Suche innerhalb dieses Rahmens angezeigt. Voraussetzung ist, dass vorher die Testschaltfläche für das gerade aktive Bild gedrückt worden ist. Das OCR Ergebnis kann aus der angezeigten Dialogbox kopiert werden und zur Definition einer Text oder Klassifizierungsbedingung verwendet werden.

Ein Klassifizierungstest legt das Ergebnis der zugehörigen OCR Suche in einer Datei im PlugIn Verzeichnis ab. Diese Datei wird mit dem nächsten Klassifizierungstest wieder überschrieben. Wenn der Inhalt der OCR Datei eingesehen werden soll, sollte diese Schaltfläche gedrückt werden.

OCR Sprache

Diese Box dient der Definition des Klassifizierungs-Tests. Es kann die OCR Sprache eingestellt werden, die bei einem Test für die OCR Suche verwendet wird.

Klassen

Alle definierten Klassen werden in dieser Liste aufgeführt. Wird eine dieser Klassen selektiert werden alle definierten Elemente dieser Klasse in den entsprechenden Dialogelementen dargestellt.



Legt eine neue Klasse an. Eine neu definierte Klasse hat zunächst einmal kein definiertes Feld.



Löscht die gerade ausgewählte Klasse.



Schiebt die selektierte Klasse um eine Zeile nach oben.

Die Reihenfolge der Klassen kann wichtig sein. Ein Bild wird der ersten Klasse zugeordnet, für die die Klassifizierungs-Bedingung erfüllt ist. Durch die Reihenfolge kann man daher auch das Klassifizierungsergebnis beeinflussen.



Schiebt die selektierte Klasse um eine Zeile nach unten.

Die Reihenfolge der Klassen kann wichtig sein. Ein Bild wird der ersten Klasse zugeordnet, für die die Klassifizierungs-Bedingung erfüllt ist. Durch die Reihenfolge kann man daher auch das Klassifizierungsergebnis beeinflussen.

Alle definierten Felder (egal ob Extraktionsfelder oder Klassifizierungsfelder) werden in dieser Liste dargestellt. Jedes Feld hat einen Namen, der beim Anlegen definiert wurde. Mit einem Rechtsklick auf ein Element in dieser Liste kann dieser Name geändert werden. Wird ein Element in dieser Liste selektiert, wird der zugehörige Rahmen auch im Vorschaufenster selektiert. Ist dieses Element auch Teil einer Klassifikations-Verknüpfung, wird es auch im Baum selektiert (nur bei Klassifizierungs-Typen möglich).

Verbinde Klassifizierungs-Felder

Die definierten Operationen werden auch in einer Baumansicht dargestellt.

Nur die Klassifizierungs-Typen können mit logischen Operationen verknüpft werden, die die Zugehörigkeit zu einer Klassifizierungs-Klasse bestimmen. Extraktionstypen erscheinen nicht in dieser Baumansicht.

Sie können die fünf Schaltflächen unter der Baumansicht auch mit der rechten Maustaste in der Baumansicht simulieren. Wählen Sie zur Definition die entsprechenden Menüeinträge.



Wenn ein Element in der Baumansicht selektiert ist, wird die Und Verknüpfung unterhalb dieser Selektion eingefügt. Ist der Baum leer, wird die Und Verknüpfung als erstes Element eingefügt.



Wenn ein Element in der Baumansicht selektiert ist, wird die Oder Verknüpfung unterhalb dieser Selektion eingefügt. Ist der Baum leer, wird die Oder Verknüpfung als erstes Element eingefügt.



Alle verfügbaren Klassifikations-Elemente werden in einem Menü angezeigt. Voraussetzung ist, dass im Baum eine logische Verknüpfung selektiert ist. Unterhalb von Klassifikationselementen können keine weiteren Klassifizierungselemente eingefügt werden.

Es werden hier keine Extraktions-Elemente angezeigt. Dieser Baum dient nur der Verknüpfung von Klassifizierungstypen (Klassifikations-Bedingung, Barcodes und Patchcodes).



Löscht das selektierte Element aus der Baumstruktur. Ist kein Element selektiert, wird nichts gelöscht.



Löscht alle definierten Verknüpfungen. Die Baumansicht wird dadurch geleert.



Links	Stellt die linke Koordinate des gerade selektierten Rahmens in Zehntelmillimeter dar.
Oben	Stellt die obere Koordinate des gerade selektierten Rahmens in Zehntelmillimeter dar.
Breite	Stellt die Breite des gerade selektierten Rahmens in Zehntelmillimeter dar
Höhe	Stellt die Höhe des gerade selektierten Rahmens in Zehntelmillimeter dar.
<b>Typ</b>	Wählen Sie zwischen Klassifizierungs-Bedingung, Texttyp, Barcodetyp, Patchcodetyp, und Listenfeld
<b>Klassifizierungs-Bedingung:</b>	<p>Das definierte Feld dient zur Klassifizierung des Dokuments. Hiermit ist ein Texttyp als regulärer oder nicht-regulärer Ausdruck gemeint.</p> <p>Klassifizierungstypen können kombiniert werden (durch logische Operationen) und definieren gemeinsam die Kennzeichen einer Klassifizierungs-Klasse.</p>
<b>Barcodes</b>	<p>Auch der Barcode ist ein Klassifizierungs-Typ. Im Gegensatz zur Klassifizierungs-Bedingung handelt es sich nicht um einen Texttyp, der mit einem OCR Ergebnis verglichen wird, sondern um einen erkannten Barcode im Bereich des definierten Feldes. Sie können unter Wert noch einen Barcodewert definieren, der gefunden werden muss.</p> <p>Hierzu eignet sich ein regulärer Ausdruck besonders, mit dem man auch Wertebereiche definieren kann. Wird kein Wert definiert, erfüllt jeder gefundene Barcode im Bereich des Feldes die Klassifizierungs-Bedingung.</p>
<b>Patchcodes</b>	Ebenfalls ein Klassifizierungs-Typ. Beim Patchcode ist der Bereich eigentlich unerheblich, da die Position des gefundenen Patchcodes nicht überprüft wird. Da es nur sechs verschiedene Patchcodetypen gibt, sind diese auch fest über eine Combobox zu wählen. Es ist zur Zeit nicht möglich bei beliebigem Patchcode die Klassifizierungs-Bedingung zu erfüllen.
<b>Text</b>	<p>Hierbei handelt es sich um einen Extraktionstyp, das heißt, die hier zu definierende Bedingung bestimmt nicht die Klassifizierungs-Klasse. Stattdessen können bestimmte Textbereiche extrahiert werden. Extraktionsbereiche werden nur dann beachtet, wenn die zugehörige Klassifikations-Klasse erkannt wurde. Wird zufällig auf einem Bild ein Text gefunden, der einer Extraktions-Bedingung genügt, das Bild aber nicht dieser Klasse zugeordnet, wird der Extraktionstext auch nicht zurückgegeben.</p>
<b>Listenfeld</b>	Das Listenfeld ist ebenfalls ein Extraktionstyp. Da die Definition einer Liste sehr umfangreich ist, kann sie mit einem zusätzlichen Dialog vorgenommen werden. Weitere Einzelheiten in der <a href="#">Listendefinition</a> .

☐ Regulärer Ausdruck

Schalten Sie diese Option ein, wenn Sie als Vergleichswert für ihren gefundenen Text oder Barcode einen regulären Ausdruck verwenden wollen. Tragen Sie diesen regulären Ausdruck dann in das Editfeld Wert ein. Lassen Sie diese Option ausgeschaltet, wenn Sie anstelle eines regulären Ausdrucks mit einem fixen Wert vergleichen wollen.

Sie können den regulären Ausdruck mit der TestReg Schaltfläche überprüfen.

**Wert**

Definieren Sie hier den Vergleichsstring. Falls die Option Regulärer Ausdruck aktiviert ist, muss hier jetzt ein korrekter regulärer Ausdruck definiert werden. Sie können die Korrektheit eines regulären Ausdrucks über die Test Schaltfläche für reguläre Ausdrücke testen.

Ist regulärer Ausdruck nicht aktiviert ist hier ein Vergleichstext anzugeben. In diesem Fall ist die Angabe einer Toleranz zu empfehlen. Ein einziges falsch erkanntes Zeichen innerhalb der OCR führt sonst zu einem negativen Vergleichsergebnis.

Bei Patchcodes wird dieses Editfeld zu einer Combobox, in der Sie nur zwischen den sechs bekannten Patchcodetypen wählen können. Die Toleranz ist dann inaktiv.

**Variable**

Um DpuScan das Ergebnis einer Text-Extraktion mitzuteilen, kann eine Variable definiert werden. Das Ergebnis der Extraktion wird in dieser Variable gespeichert und kann dann innerhalb von DpuScan weiter verwendet werden.

Die Variable für ein Listefeld wird nicht an dieser Stelle sondern in der Listendefinition festgelegt. Sie ist für Listentypen daher inaktiv.

☐ Großschreibung

Macht nur bei Text Typen Sinn. Schalten Sie die Groß-Kleinschreibungs-Unterscheidung aus (Checkbox muss dazu aktiviert werden), wenn Sie es wünschen.

**Filter**

Dieses Feld ist nur dann aktiv, wenn ein regulärer Ausdruck definiert wurde. Sie können Teile eines regulären Ausdrucks, die in runden Klammern stehen, über einen Parameter gezielt auswählen.

\1 ist das Ergebnis der ersten runden Klammer im regulären Ausdruck, \2 das Ergebnis der zweiten runden Klammer im regulären Ausdruck.

**Toleranz**

Falls Sie mit einem Text Typ oder Klassifizierungs-Typ vergleichen, ist es sinnvoll eine Toleranz vorzugeben. Falls die OCR Suche nur ein Zeichen falsch erkannt hat, wäre ansonsten der Vergleich fehlgeschlagen. Diese Option ist nur für Text Typen verfügbar und wenn kein regulärer Ausdruck verwendet wird. Bei Patchcodes kann keine Toleranz angegeben werden, bei Listentypen kann eine Toleranz innerhalb der Listendefinition angegeben werden (an dieser Stelle hier ist sie dann nicht verfügbar).

Empfohlene Werte für eine Toleranz liegen zwischen 5 und 20, je nach Qualität des Bildes. Eine 0 bedeutet, der Text muss vollständig übereinstimmen.

<b>OK</b>	Schließt die PlugIn Konfiguration und speichert alle Einstellungen.
<b>Abbrechen</b>	Schließt die PlugIn Konfiguration ohne zu speichern.
<b>Hilfe</b>	Öffnet die PlugIn Hilfe

## 4 Feld-Typen für die Klassifizierung

Klassifizierungsfelder dienen dazu, festzustellen, um welche Dokumentenklasse es sich bei dem Beleg handelt. Auf Basis der Text-, Barcode- oder Patchcode-Erkennung wird für das Feld geprüft, ob die formulierte Bedingung zutrifft.

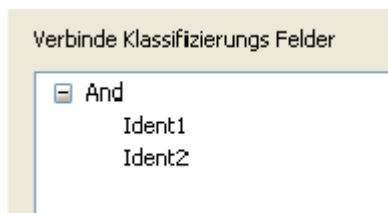
Definition für ein Klassifizierungsfeld mit Textauswertung

Der Feld-Typ wird über die Drop-Down-Liste **Typ** bestimmt.

Zur Auswahl stehen

- Klassifizierungsbedingung (Textvergleich)
- Barcode und
- Patchcode

Wenn mehr als ein Klassifizierungsmerkmal notwendig erscheint, um eine Klasse zu identifizieren, sind diese Merkmale logisch zu verknüpfen. Die Verknüpfungen werden in einer Baumdarstellung angezeigt.



Verknüpfung von Klassifizierungsbedingungen

Innerhalb einer Konfiguration können beliebig viele Klassen definiert werden. Die Prüfung der der Felder erfolgt jeweils nacheinander für jede der Klassen, wobei die Liste von oben nach unten abgearbeitet wird.

Sobald die Bedingungen bzw. deren Verknüpfung für eine Klasse wahr sind, gilt die Klasse als erfolgreich bestimmt und die weitere Prüfung für die restlichen Klassen wird nicht mehr durchgeführt. Daher kann die Reihenfolge der Klassen in der Liste beliebig geändert werden.

Der Name der ermittelten Klasse wird in die Variable `%(I.CLASSIFY)` eingetragen und evtl. vorhandene Felder für die Datenextraktion werden bearbeitet.

Danach werden alle Variablen an DpuScan übertragen und können dort weiter verarbeitet werden, beispielsweise in Ereignisregeln, Protokolldateien etc..

## 4.1 Klassifizierungsbedingung

Die Auswahl **Klassifizierungsbedingung** kennzeichnet ein Feld, wo das OCR-Ergebnis in dem Bereich analysiert wird, der durch den zugehörigen Rahmen gekennzeichnet ist.

Der Vergleich mit dem Suchmuster kann dabei als reiner Textvergleich oder als Anwendung eines regulären Ausdrucks konfiguriert werden.

Ein **Textvergleich** sucht den gegebenen Text in dem Text, der von der OCR erkannt wurde. Hierbei kann eine **Toleranz** angegeben werden, so dass ein Wort auch dann noch als zutreffend eingeordnet wird, wenn beispielsweise ein oder zwei Buchstaben falsch sind. Den Toleranzwert kann man mit dem Schieberegler am unteren Fensterrand einstellen. 0 bedeutet hierbei, dass der Wert exakt der hier beschriebenen Weise auftaucht. Schreib- oder Erkennungsfehler werden nicht akzeptiert.

Mit zunehmender Toleranz werden auch mehr Abweichungen vom gegebenen Text akzeptiert.

Eine weitere Vergleichsmöglichkeit ist die Anwendung eines **regulären Ausdrucks**. Erzielt dieser Ausdruck einen Treffer, so gilt das Feld als erfolgreich gefunden.

Mehr zu regulären Ausdrücken findet sich im Kapitel 7 [Reguläre Ausdrücke](#) .

## 4.2 Barcode

Zur Klassifizierung wird kann auch der Barcodewert aus der DpuScan Variable `%(S.BAR1)` verwendet werden. Der Barcodewert kann analog zum Feld-Typ Klassifizierungsbedingung ausgewertet werden. Dabei kann ein Textvergleich mit Toleranz oder ein regulärer Ausdruck verwendet werden.

Mehr zu regulären Ausdrücken findet sich im Kapitel 7 [Reguläre Ausdrücke](#) .

## 4.3 Patchcode

Zur Klassifizierung wird kann auch der Patchcodewert aus der DpuScan Variable `%(S. PATCH)` verwendet werden. Der Patchcodewert kann analog zum Feld-Typ Klassifizierungsbedingung ausgewertet werden. Dabei kann ein Textvergleich mit Toleranz oder ein regulärer Ausdruck verwendet werden.

Mehr zu regulären Ausdrücken findet sich im Kapitel 7 [Reguläre Ausdrücke](#) .

## 5 Feld-Typen für die Datenextraktion

Das Auslesen von Zeichen oder Werten wird nachfolgend als Datenextraktion bezeichnet.

Im Gegensatz zur einfachen Klassifikation soll hier entsprechend den definierten Bedingungen der gelesene Wert an das Scan-Programm zurückgeliefert werden.

Beispielsweise kann hier nach einem Datum oder einer Rechnungsnummer gesucht werden, wobei es nicht ausreicht, dass ein solcher Ausdruck gefunden wurde, es soll auch der Wert zurückgeliefert werden.

Um diese Aufgabe zu erfüllen stehen drei verschiedene Feldtypen zur Verfügung:

Im Textfeld kann eine Suche nach festem Text mit einstellbarer Fehlertoleranz oder eine Suche mit regulären Ausdrücken durchgeführt werden.

### 5.1 Textfeld

Das **Textfeld** dient zur einfachen Suche nach Textelementen in einem bestimmten Bereich des Bildes. Der Bereich des Suchfeldes wird durch einen Rahmen auf dem Bild festgelegt.

Konnte ein solches Element gefunden werden, so wird das Ergebnis der Suche in einer frei definierbaren Variablen gespeichert. Der Variablenname wird in dem Eingabefeld **Variable** angegeben.

Wenn der Suchtext nicht gefunden werden konnte, ist die Variable leer.

Die Suche kann dabei als reiner Textvergleich oder als Anwendung eines regulären Ausdrucks konfiguriert werden.

Bei diesem Feldtyp sucht das PlugIn den Eintrag aus dem Eingabefeld **Wert** in dem Text, der von der OCR erkannt wurde.

Ist ein reiner Textvergleich gewählt, d.h. wenn das Kontrollkästchen **Regulärer Ausdruck** keinen Haken enthält, dann ist der Schieberegler **Toleranz** aktiv und es kann ein Toleranzwert angegeben werden, so dass ein Wort auch dann noch als zutreffend eingeordnet wird, wenn beispielsweise ein oder zwei Buchstaben falsch sind. 0 bedeutet hierbei, dass der Wert exakt der hier beschriebenen Weise auftaucht. Schreib- oder Erkennungsfehler werden nicht akzeptiert. Mit zunehmender Toleranz werden auch mehr Abweichungen vom gegebenen Text akzeptiert.

Eine weitere Vergleichsmöglichkeit ist die Anwendung eines regulären Ausdrucks. Dieses Verfahren wird ausgewählt, indem das Kontrollkästchen **Regulärer Ausdruck** markiert wird. Erzielt der Ausdruck aus dem Eingabefeld **Wert** einen Treffer, so gilt das Feld als erfolgreich gefunden. Wenn nur ein Teilausdruck in der Variablen gespeichert werden soll, dann ist dieser im Eingabefeld **Filter** anzugeben.

Mehr zu regulären Ausdrücken findet sich im Kapitel 7 [Reguläre Ausdrücke](#) .

## 5.2 Listenfelder

Ein Listenfeld enthält eine Liste mit beliebig vielen Zeilen. In jeder der Zeilen sind ein oder zwei Bedingungen angegeben, die jeweils wie das bereits beschriebene **Textfeld** arbeiten. Sind zwei Bedingungen in einer Zeile angegeben, so sind diese logisch miteinander zu verknüpfen.

Jede einzelne Zeile der Liste wird mit den OCR-Ergebnissen aus dem zur Liste gehörenden Suchbereich verglichen. Der Vergleich erfolgt in der Reihenfolge, in der die Zeilen in der Liste aufgeführt sind.

Sobald die Bedingungen einer Zeile zutreffen wird die Variable mit dem in der Zeile festgelegten Wert gefüllt und die weitere Prüfung der restlichen Zeilen in der Liste nicht mehr durchgeführt. Der Rückgabewert ist frei wählbar und wird pro Zeile der Liste angegeben.

Folgende Verknüpfungsarten sind möglich:

- Und-Verknüpfung
- Oder-Verknüpfung
- UndNicht-Verknüpfung

Jede Zeile in dieser Liste kann einzeln definiert werden. Sie besteht aus dem Rückgabewert und bis zu zwei Vergleichsstrings, die mit einer logischen Operation verknüpft werden.

### Dialogelemente

Zeile	<p>Eine Zeile wird durch die Schaltflächen <b>Hinzufügen</b> oder <b>Ändern</b> definiert und besteht aus dem Rückgabewert, dem ersten Vergleichsstring, dem ersten Vergleichstyp der logischen Operation und dem zweiten Vergleichsstring und Vergleichstyp.</p> <p>Der Rückgabewert ist ein beliebiger Text, der bei erfolgreichem Vergleich für die definierte Zeile zurückgegeben werden soll.</p> <p>Der Typ ist entweder ein regulärer Ausdruck oder ein Toleranzwert. Diese Begriffe sind in der Konfiguration des PlugIns bereits erklärt worden. Als Toleranz ist hier ein Wert zwischen 0 und 100 zulässig.</p> <p>Bei den logischen Operationen gibt es außer den bekannten <b>And / Or</b> nun auch noch <b>No</b> (wenn nur der erste der beiden Strings ausgewertet werden soll) und <b>And Not</b>.</p>
Variable	<p>Definiert den Namen der Rückgabevariable für DpuScan. Der Variablenname ist für alle Zeilen gleich, es ändert sich nur der Rückgabewert, der für jede Zeile separat definiert werden kann.</p>
Hinzufügen	<p>Fügt eine neue Zeile an das Ende der Liste ein. Es öffnet sich ein weiterer Dialog um die einzelnen Elemente einer Zeile zu definieren.</p>
Ändern	<p>Voraussetzung für das Ändern einer Zeile der Liste ist, dass überhaupt schon eine definiert wurde und dass diese dann auch selektiert ist. Es öffnet sich dann der Dialog, der auch beim Hinzufügen einer Zeile verwendet wird. Die einzelnen Elemente einer Zeile haben ihren momentanen Wert und können geändert werden.</p>

Löschen	Löscht die selektierte Zeile aus der Liste.
Lösche alles	Löscht alle definierten Zeilen aus der Liste.
Importieren	<p>Ermöglicht das Öffnen einer Textdatei. Dadurch kann eine komplette Listendefinition in einem Schritt eingefügt werden. Die Importdatei muss einer gewissen Syntax genügen:</p> <p>Zeilen müssen durch Carriage Return/Linefeed getrennt sein. Die einzelnen Zeilenelemente sind durch # getrennt. Toleranzen müssen einen Wert zwischen 0 und 100 haben und geben damit an, dass kein regulärer Ausdruck verwendet werden soll. Jeder Wert, der länger als drei Zeichen ist, führt zu einer Interpretation des Eintrags als regulärer Ausdruck, unabhängig vom Inhalt der Zeichenkette. Bei den Operationen ist die Groß-Kleinschreibung irrelevant, es werden aber nur die Werte And, Or, And Not sowie No akzeptiert.</p>

### 5.3 Relative Suche

Als relative Suche wird eine Methode bezeichnet, die zwei Suchkriterien derart verknüpft, dass

- die Ausgangs-Position durch den ersten Suchbegriff bestimmt wird und
- dann in relativem Abstand zu dem ersten gefundenen Begriff der zweite Begriff gesucht wird.

So kann beispielsweise erreicht werden, dass in einem definierbaren Bereich unterhalb des Begriffs **"Datum"** mit einem regulären Ausdruck nach einer Ziffernstruktur gesucht wird, die einer Datumsangabe entspricht.

*Relative Suche nach einem Datumswert*

Kann der erste Begriff gefunden werden und auch die Suche nach dem zweiten Begriff ist erfolgreich, dann kann der Wert des zweiten Begriffs in einer Variablen gespeichert und an DpuScan übergeben werden.



## 6 Verwendung des Plugins in DpuScan

Das Plugin benötigt sowohl die Daten aus einer OCR Ganzseitensuche mit exportiertem XML Format als auch Informationen über gefundene Bar- und Patchcodes. Damit diese Daten beim Plugin-Aufruf auch zur Verfügung stehen muss im OpenJob Modus gearbeitet werden und OCR Ganzseitensuche, Bar- und Patchcodesuche vor dem Aufruf des Klassifizierungs-Plugins ausgeführt werden. Die zwingend nötigen Taskschritte sind also:

```
Lade Basisprofil (Laden der DpuScan Klasse)
Lade Stapel (OpenJob)
  Lade vom Scanner (Scanprozess)
  Barcodes suchen
  Patchcode suchen
  Pfadnamen setzen
  Dateinamen setzen
  OCR ausführen
  Plugin für jedes Bild aufrufen (Aufruf des Classify Plugins)
  Bild speichern
```

## 7 Reguläre Ausdrücke

Ein regulärer Ausdruck (abgekürzt auch RegExp oder Regex) ist ein Suchmuster, das verwendet wird, um in einem Text eine bestimmte Zeichenfolge, z.B. ein Wort, zu finden. Das Suchmuster wird mit einer komplexen Syntax aufgebaut.

Die zu verwendende Syntax-Variante kann für das jeweilige Feld eingestellt werden.

Dazu dient die Schaltfläche .

### 7.1 Syntax

Anmerkung: Zur besseren Lesbarkeit in diesem Text sind im Folgenden einige regulären Ausdrücke in Anführungszeichen " " angegeben; die Anführungszeichen sind aber in jedem Fall *nicht* Bestandteil des Suchausdrucks und müssen bei der Suche weggelassen werden.

Ein regulärer Ausdruck besteht im Allgemeinen aus Platzhaltern und ggf. aus Zeichen, welche in dem zu suchenden Text exakt vorkommen müssen, sog. Literalen.

#### Literale

Alle Zeichen außer ".", "\*", "?", "+", "(", ")", "{", "}", "[", "]", "^", "\$" und "\" sind Literale. Sollen Punkt, Stern, Fragezeichen, Pluszeichen, Dachzeichen, Dollarzeichen, runde, eckige, und geschweifte Klammern sowie der inverse Schrägstrich trotzdem als Literal verwendet werden, müssen sie „maskiert“ werden, d.h. es wird ihnen ein inverse Schrägstrich vorangestellt.

Soll z.B. nach dem Dateinamen `C:\temp\Datei.txt` gesucht werden, so müssen die inversen Schrägstriche im regulären Ausdruck doppelt angegeben werden. Ebenso muss der Punkt vor dem Dateityp maskiert werden:

```
"C:\\Temp\\Datei\\.txt"
```

**Im Klassifikation-PlugIn wird bei der Suche immer die Großkleinschreibung beachtet.**

#### Platzhalter, Wildcards

Der Punkt "." steht für ein beliebiges einzelnes Zeichen.

#### Wiederholungen, Geschweifte Klammern

Ein Wiederholungsausdruck ist ein Ausdruck, der angibt, wie oft der vorangestellte Ausdruck vorkommen darf.

"\*" Der Stern hinter einem Ausdruck bedeutet, dass dieser Ausdruck beliebig oft oder auch gar nicht vorkommen darf.

"+" Das Plus bedeutet, dass dieser Ausdruck beliebig oft, aber mindestens einmal vorkommen muss.

"?" Das Fragezeichen gibt an, dass ein Ausdruck höchstens einmal (oder gar nicht) vorkommen darf.

"{x,y}" Die Werte in den geschweiften Klammern geben an, wie oft ein Ausdruck mindestens x bzw. höchstens y vorkommen darf. Der Mindestwert x muss dabei immer angegeben werden, der Höchstwert y darf weggelassen werden.

Ist kein Höchstwert angegeben, darf der Suchausdruck beliebig oft vorkommen.

### Beispiele:

"Ab\*" "

Suche nach einem einzelnen A vielleicht gefolgt von b's, findet "A" , "Ab" , "Abbb" usw.

"Ab+" "

Suche nach einem A gefolgt von mindestens einem b, findet "Ab" oder "Abbbb" aber nicht "A" .

"Ab?" "

Suche nach einem A gefolgt von höchstens einem b , findet "A" oder "Ab" .

"Ab{ 2 , 4 }" "

Suche nach einem A gefolgt von mindestens 2 und höchsten 4 b's, findet

"Abb" , "Abbb" und "Abbbb" .

"Ab{ , 5 }" "

Suche nach einem A gefolgt höchsten fünf b's

"Ab{ 3 , }" "

Suche nach einem A gefolgt mindestens drei b's

In diesen Beispielen beziehen sich die Wiederholungsausdrücke jeweils nur ein einzelnes Zeichen, das „b“. Will man Buchstabenfolgen wiederholen, kann man Klammern verwenden.

### Zusammenfassungen, Runde Klammern

Runde Klammern dienen der Zusammenfassung von Suchmustern und der Gruppierung des Ergebnisses.

**Im Klassifikation-Plugin muss ein Suchmuster immer in Klammern eingeschlossen werden.**

Der Einstelldialog bietet dort auch die Möglichkeit, einzelne Ergebnisgruppen abzuholen. Dazu wird dort als Ergebnis-Filter "\1" als Platzhalter für das erste Teilergebnis verwendet.

"(Ab)\*" "

sucht nach der Zeichenfolge Ab mit beliebig viele Wiederholungen, findet Ab AbAb usw.

".\*(Betrag).\*" "

sucht nach dem Wort Betrag, wobei davor und dahinter beliebig viele andere Zeichen sein dürfen. Das Ergebnis wird als ersten gemerkter Text gespeichert und kann mit "\1" abgerufen werden.

**Wenn im Klassifikation-Plugin eine Zeichenkette in einem Text gesucht wird, so müssen stets Suchmuster für den Text vor und nach dem zu suchenden Teil angegeben werden.**

Ist man an den Texten davor und dahinter interessiert, so kann man diese ebenfalls klammern

"(.\*) (Betrag) (.\*)" "

In diesem Fall findet sich das Wort Betrag im zweiten Teil und kann mit "\2" abgerufen werden.

Fängt - im Gegensatz dazu - der geklammerte Ausdruck "?:" an, so wird der Unterausdruck zwar ermittelt, aber nicht zurückgegeben.

"(?:.\*) (Betrag) (?:.\*)" "

Jetzt findet sich das Wort Betrag wieder im ersten Teil "\1". Bemerkenswerterweise verhält sich die Suche nun nicht mehr gefräßig.

### Alternativen, Senkrechte Striche

Senkrechte Striche können verwendet werden, um zwischen jeweils zwei möglichen Ausdrücken zu wählen:

```
" ( Jan | Januar ) "
```

Sucht in einem Text nach Januar oder Jan.

### Klassen, eckige Klammern, Dach

Klassen sind Zusammenfassungen von Zeichen, nach denen gesucht werden soll. Sie werden durch eckige Klammern eingeschlossen und enthalten die die Zeichen, die erlaubt sein sollen.

```
" ( [ UuMmWmSsTt ] + ) "
```

Sucht in einem Text nach einem Wort, das aus den Buchstaben UMWST, groß oder klein besteht, und findet z.B. UST, Ust, MWST, MWSt aber auch stumm.

Statt einer aufzählenden Schreibweise kann auch ein Bereich angegeben werden:

```
" ( [ A-Z ] + ) "
```

sucht in einem Text nach allen Großbuchstaben.

```
" ( [ A-z ] + ) "
```

sucht in einem Text nach einem Wort, das aus Buchstaben, aber nicht aus z.B. Zahlen besteht.

Das Dach steht für eine Umkehrung, d.h. es werden nur Zeichen genommen, die *nicht* aufgezählt sind

```
" ( [ ^a-z ] + ) "
```

sucht in einem Text nach einem Wort, das keine Kleinbuchstaben enthält.

Statt der Aufzählung der Buchstaben kann man auch Klassennamen verwenden. Dafür sind diese Konstanten vorgesehen:

" [[:alnum:]] "	Buchstaben oder Zahlen
" [[:word:]] "	Buchstaben, Zahlen oder Unterstriche
" [[:alpha:]] "	Buchstaben
" [[:digit:]] "	Zahlen
" [[:punct:]] "	Interpunktionszeichen, (Punkt, Komma, Semikolon, alle Klammern, ...)
" [[:print:]] "	Beliebige druckbares Zeichen
" [[:space:]] "	Leerzeichen, Tabulatoren, Zeilenumschaltungen, Seitenvorschübe
" [[:upper:]] "	Großbuchstaben
" [[:lower:]] "	Kleinbuchstaben

Für häufig verwendete Klassen gibt es Kurzschreibweisen

"\d" Zahlen

"\w" Buchstaben, Zahlen oder Unterstriche

"\s" Leerzeichen, Tabulatoren, Zeilenumschaltungen, Seitenvorschübe

"\u" Großbuchstaben

"\l" Kleinbuchstaben

"([[:alpha:]]+)"

sucht nach einer Folge von Buchstaben, also einem Wort.

"(\d{1,4})"

sucht nach einer ein- bis vierstelligen Zahl, und findet 123 oder 5

## 7.2 Beispiele für reguläre Ausdrücke

### Umsatzsteuer ID (Deutschland)

In Deutschland beginnt die Umsatzsteuer ID mit der Zeichenfolge "DE", gefolgt von 9 Ziffern.

"DE\d{9}"

DE	Sucht nach einer genauen Übereinstimmung im Text mit der Zeichenfolge DE.
\d{9}	Sucht nach genau 9 Ziffern

### Betrag

Gesucht wird die Zahl hinter dem Wort Summe oder Zwischensumme.

Zum Beispiel 123,45 oder 67,--

"(Summe|Zwischensumme) (\s\*) (\d{1,} [.,] {1} [-|\d]{2}) "

(Summe Zwischensumme)	Sucht nach einer genauen Übereinstimmung im Text mit dem Wort <b>Summe</b> oder dem Wort <b>Zwischensumme</b>
(\s*)	Sucht nach einer beliebigen Anzahl an Leerzeichen
(\d{1,} [.,] {1} [- \d]{2})	Sucht nach mindestens einer Ziffer, gefolgt von genau einem Punkt oder Komma, gefolgt von zwei Minuszeichen oder Ziffern. Dieser letzte Teil findet zum Beispiel 122,18 aber auch 150, --

Um die Zahl zu erhalten, muss der dritte Teilausdruck referenziert werden, also liefert der Eintrag \3 im Feld "Filter" den Betrag.

### Bestellnummer

Gesucht wird die Zeichenkette hinter dem Wortteil **ummer** oder **nr**. Die Bestellnummer kann – je nach Lieferant – bis 30 Stellen lang sein und Trennzeichen enthalten

"(ummer|nr\.) (\s\*) ([[:digit:]][:punct:]]{1,30}"

(ummer nr\.)	Sucht nach einer genauen Übereinstimmung im Text mit dem Wortteil <b>ummer</b> oder dem Kürzel <b>nr</b>
(\s*)	Sucht nach einer beliebigen Anzahl an Leerzeichen
([[:digit:]][:punct:]]{1,30})	Sucht nach mindestens einer Ziffer oder Interpunktionszeichen, höchstens jedoch 30 solcher Zeichen Dieser letzte Teil findet zum Beispiel 122.18.4432.1234 oder auch 122/18/4432.1234

Um die Zahl zu erhalten, muss der dritte Teilausdruck referenziert werden, also liefert der Eintrag \3 im Feld "Filter" den Betrag.

*Hinweis:*

Falls ein Ausdruck kein Ergebnis liefert, obwohl im Fenster zum Testen von regulären Ausdrücken Ergebnisse angezeigt wurden, kann das daran liegen, dass die OCR Zeichen vor der eigentlichen Struktur gefunden hat. Probieren Sie, Muster in der Art von (.\*) dem eigentlichen Suchmuster voranzustellen bzw. anzufügen.



Janich & Klass Computertechnik GmbH  
Zum Alten Zollhaus 24  
42281 Wuppertal  
Deutschland

Tel.: +49 (0)202 2708-0  
Fax: +49 (0)202 700 625  
<http://www.janichklass.com>  
<http://www.JKImaging.com>

408.201002.049 7