Starke-DMS®
Dokumenten-Management

Starke-DMS®

Dokumenten-Management für den Mittelstand

Belegerkennung / - training (Entwurf)

Version 3.1.0

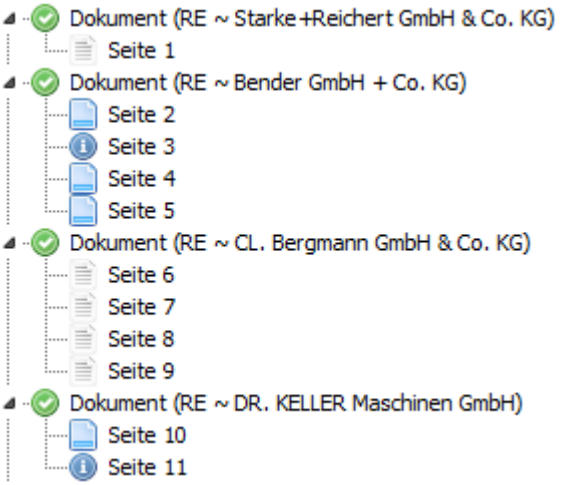
Inhaltsverzeichnis

Automatische Belegerkennung

Das SCAN-Modul kann um eine automatische Erkennung von Belegen erweitert werden. Neben der Zuordnung des Absenders (z.B. Lieferant) können nahezu beliebige Informationen auf der ersten Seite des Belegs extrahiert und in Indizes übertragen werden.

Zusammen mit der Stapelarchivierung kann ein sehr hohes Belegaufkommen schnell und mit geringem Aufwand verarbeitet werden, da eine manuelle Verschlagwortung weitestgehend entfällt.

Wenn alles korrekt konfiguriert ist, erscheinen Dokumente nach dem Scannen oder Importieren direkt als archivierungsbereite Dokumente im Stapel auf.



Das Belegtraining

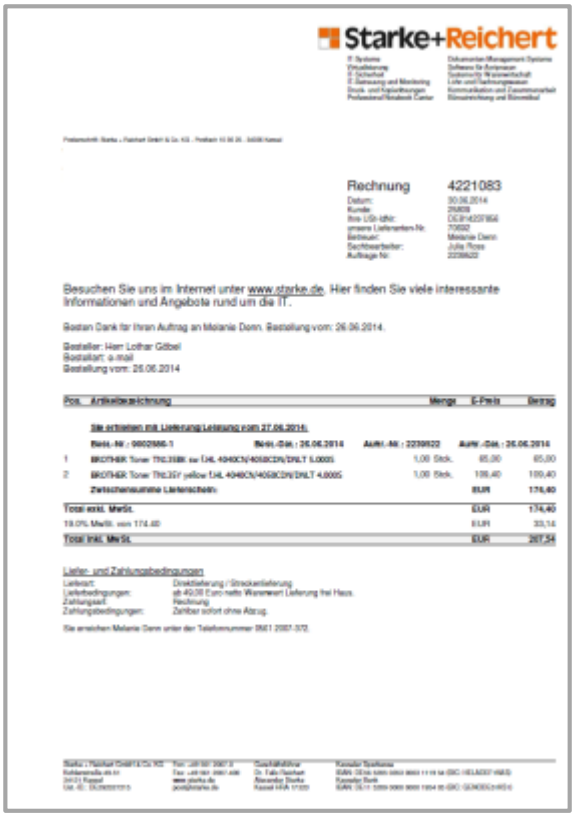
Damit das SCAN einen Beleg erkennt und verarbeiten kann, muss man dem Modul die Erkennung erst „antrainieren“. Das ist eine Vorgehensweise, bei dem Sie dem Programm erklären, wie Sie selbst den Beleg unterbewusst analysieren und aufbereiten.

Wenn sie z.B. eine Rechnungsnummer auf dem Blatt finden möchten, suchen Sie in erster Linie nach bestimmten Begriffen wie „Rechnungsnummer“ oder „RE-Nr.“ und erwarten daneben oder darunter einen Wert. Dieses für Menschen selbstverständliche Vorgehen wird durch das Training hinterlegt.

Die ersten Schritte: Belegnummer auslesen

In diesem Abschnitt wird Ihnen gezeigt, wie Sie die Belegnummer der Rechnung der Firma *Starke+Reichert GmbH & Co. KG* (im Folgenden mit S+R abgekürzt) trainieren und vollautomatisch auslesen.

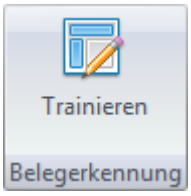
Sie können später nach Belieben weiterer Informationen wie Kundennummer, Sachbearbeiter usw. auslesen. Die Vorgehensweise ist in vielen Fällen ähnlich.


Starke+Reichert

Rechnung 4221083

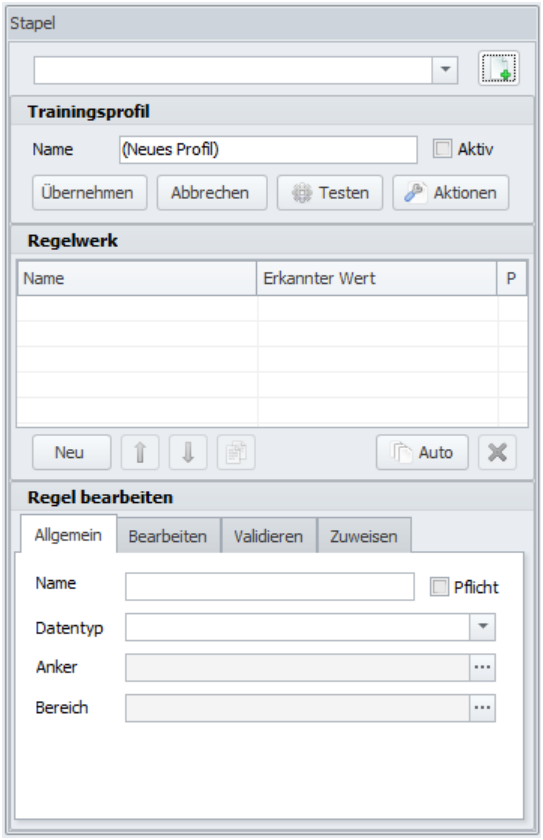
Training starten

Um das Training zu starten, fügen Sie ein Beleg der Firma S+R als Dokument ihrem Stapel hinzu. Das muss nicht zwingend eine Rechnung sein, dank gleichbleibendem Aufbau kann auch ein Angebot, Auftrag oder Lieferschein verwendet werden.



Wählen Sie die erste Seite des Dokuments im Stapel aus und öffnen Sie den Trainingsbereich durch einen Klick auf das Menü „Dokument | Trainieren“.

Die Stapelanzeige wird durch das Trainingsmodul ersetzt, alle anderen Fenster bleiben bestehen.



Der neue Bereich teilt sich in drei Abschnitt auf:

1. Trainingsprofil

   Hier sind die Kopfinformationen zu einem Profil hinterlegt.

2. Regelwerk

   Die Liste beinhaltet alle Regeln eines Profils. Ein Profil kann beliebig viele Regeln beinhalten.

3. Regel bearbeiten

   Sobald man im Regelwerk eine Regel auswählt, kann man in diesem Abschnitt die entsprechenden Einstellungen vornehmen.

Was ist ein Trainingsprofil?

Ein Profil ist vergleichbar mit einer Arbeitsanweisung und ist die Klammer um Regelwerk bestehend aus mehreren Arbeitsschritten. In der Praxis wird man pro Lieferant ein Profil hinterlegen.

Was ist eine Regel?

Eine Regel ist ein Arbeitsschritt und ein Teil eines Profils. Eine Regel kann immer nur einen Wert auslesen und einen Index füllen (z.B. die Belegnummer auslesen). Möchte man fünf Werte aus dem Beleg auslesen, braucht man mindestens genauso viele Regeln.

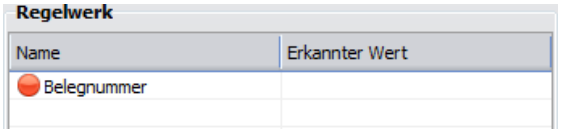
Profil erstellen

Wir werden im ersten Schritt dem Profil einen Namen geben und Platzhalterregeln (=Regeln ohne Konfiguration) anlegen, damit wir später den Überblick behalten, was wir alles schon auslesen können und noch bearbeiten müssen.

1. Wählen Sie im mittleren Bereich des SCAN-Moduls die Dokumentenart und Indexmaske aus, als würden Sie das Dokument manuell verschlagworten wollen.

2. Geben Sie im Bereich Trainingsprofil im Feld [Name] einen eindeutigen und aussagekräftigen Begriff ein, z.B. „RE Starke+Reichert GmbH“. Dieser Begriff wird später auch in Klammern hinter den erkannten Belegen im Stapel stehen.

3. Klicken Sie nun auf die Schaltfläche [Neu] unterhalb der Regelliste. Es erscheint ein neuer Eintrag in der Liste und Sie können im Abschnitt Regel bearbeiten im Feld [Name] die Bezeichung in „Belegnummer“ ändern. Sobald Sie wieder in die Liste des Regelwerks klicken, sollte die Anzeige mit dem Bild rechts übereinstimmen.



   Die rote Kugel vor dem Name bedeutet, dass die Regel nicht erkannt wurde und somit ungültig ist. Das ist nicht weiter verwunderlich, wir haben ja auch noch keine Einstellungen hinterlegt.

Beginnen wir nun, die Regel mit Leben zu füllen.

Regel bearbeiten

Nun werden wir die Belegnummer auslesen.

1. Klicken Sie die Regel „Belegnummer“ in der Regelliste an. Im unteren Bereich können wir nun die Einstellungen für die Regel vornehmen.

   Was ist ein Anker?

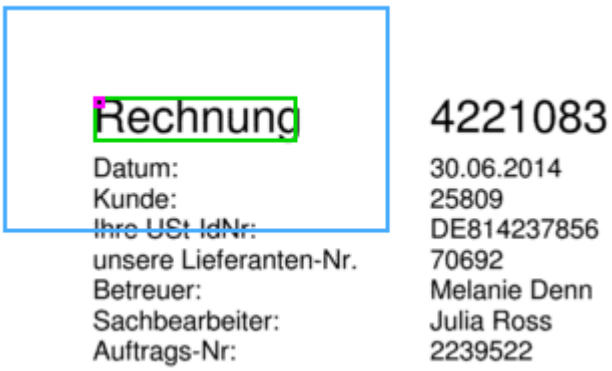
   Durch unterschiedliche Scanvorgänge kann es passieren, dass die Rechnungsnummer nicht immer exakt auf der Position steht, wo Sie sie in unserem Musterbeleg sehen. Die Nummer kann etwas in alle Richtungen verrutschen, je nachdem wie der Scanner das Blatt einzieht usw. Wenn wir nun einen engen Rahmen um die Nummer ziehen und sagen, „lies diese Nummer aus“, würden Belege mal mehr mal weniger gut erkannt werden.

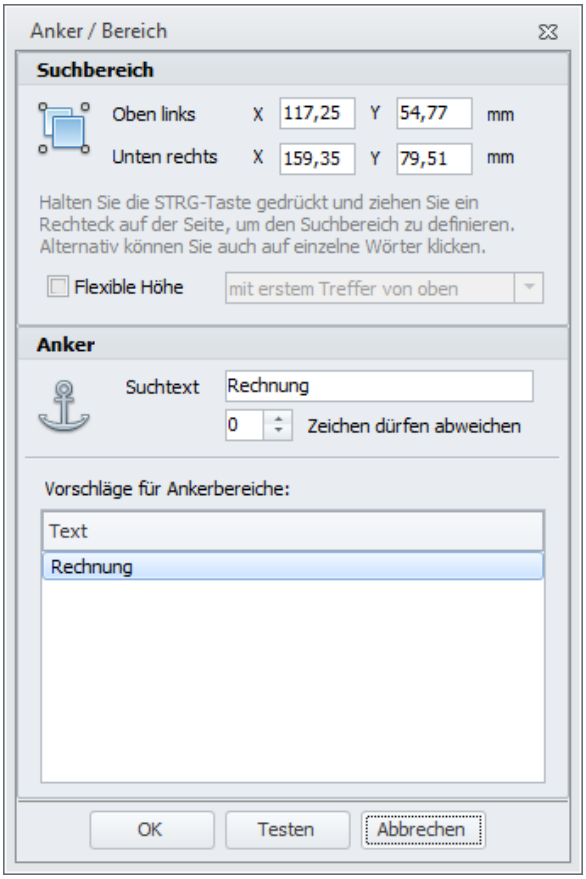
   Aus diesem Grund kann man einen Anker festlegen, der nicht ohne Grund ein Namensvetter des Schiffsankers ist. Wie in der Schifffahrt ist er dafür da, damit wir uns an einem bestimmen Punkt - auf dem Blatt Papier – festhalten können. Wenn der Scanner nun das Blatt verschiebt, schiebt sich der Anker mit und wir „bleiben auf Position“.

   Was ist ein Bereich?

   Der Bereich ist ein Viereck auf dem Blatt und stellt die Extraktionszone dar. Alle Absätze und Zeilen, die sich mit dem Bereich überschneiden, werden ausgelesen und stehen in der Regel zur Verfügung.

2. Wir legen einen Anker an. Klicken Sie dazu auf […] im Feld Anker und es öffne sich der Ankerdialog. Halten Sie nun die STRG-Taste gedrückt und klicken Sie mit der linken Maustaste auf das großgeschriebene Wort „Rechnung“ auf dem Beleg.





3. Durch den Klick hat der Ankerdialog den Abschnitt markiert und automatisch einen Anker gesetzt.

Was bedeuten die farbigen Rahmen?

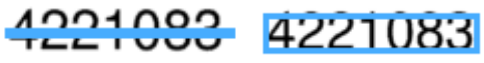
Der blaue Rahmen ist der Bereich, in dem das Wort „Rechnung“ auf dem Blatt gesucht wird. Bei automatischen Ankern wird der Bereich ca. 1-2cm um den gefundenen Abschnitt ausgedehnt. Steht das Wort „Rechnung“ nicht in dem blauen Bereich, wird der Anker später nicht funktionieren und die Erkennung der Rechnungsnummer scheitern.

Der grüne Rahmen umrandet den Abschnitt, der von der OCR als zusammenhängender Paragraf/Textzeile erkannt wurde. Das Training arbeitet nicht mit genaue Zonen oder Wörtern, sondern analysiert Textzeilen und Paragrafen. Dazu später mehr.

Der violette Punkt/Rahmen ist letztendlich der Punkt, an dem der Anker sitzt. Von diesem Punkt aus wird später relativ gesehen die Rechnungsnummer gesucht. Verschiebt sich der Begriff „Rechnung“ auf dem Beleg innerhalb der blauen Umrandung, verschiebt sich auch der Ankerpunkt mit.

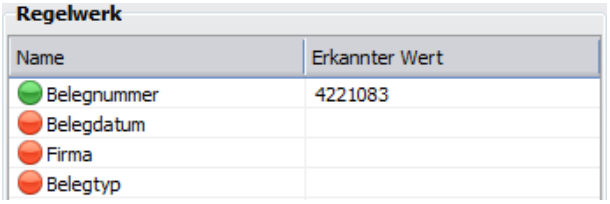
4. Schließen Sie den Dialog mit [OK].

5. Wir definieren nun, wo die Rechnungsnummer steht. Klicken Sie dazu auf […] im Feld Bereich. Es erscheint erneut der Ankerdialog, da dieser im oberen Abschnitt auch für den Bereich zuständig ist. Halten Sie erneut die STRG-Taste fest und ziehen sie einen Bereich auf der Rechnungsnummer. Der Bereich muss die Nummer nicht komplett umranden, eher im Gegenteil. Es geht nur darum, dass der Bereich sich mit der Nummer überschneidet. Hier einige gültige Beispiele:



Wichtig ist, dass Sie den Rahmen nicht so großen ziehen, das Sie Informationen aus anderen Zeilen mit erkennen. Das erschwert uns sonst später die Extraktion der Belegnummer.

6. Klicken Sie auf [OK], um den Dialog zu schließen. Damit ist das Profil bereits in der Lage, die Nummer auszulesen. Klicken Sie auf [Testen] oberhalb der Regelliste. Die Regel für die Belegnummer sollte nun grün werden und in der Spalte „Erkannter Wert“ die Nummer anzeigen.

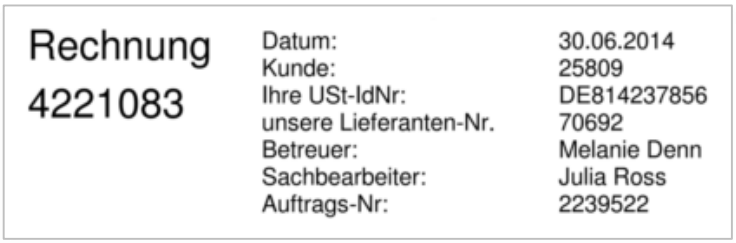


Gratulation, Sie haben soeben Ihre erste Regel erstellt und einen Wert ausgelesen.

Motto: „Immer mit dem schlimmsten Fall rechnen“

Sie haben sicherlich schon bemerkt, dass in der Einführung auf viele Optionen und Einstellungen gar nicht eingegangen wurde. Das Auslesen von Belegen kann einfach sein, aber auch schnell sehr komplex werden. In dem Beispiel oben sind wir einfach davon ausgegangen, dass die Texterkennung immer korrekt eine Zahl erkennt und durch Formularänderungen in Zukunft auch keine weiteren (unerwünschte) Informationen mit reinrutschen.

Was würde passieren, wenn der Lieferant den Aufbau des Formulars wie folgt ändert?



Gemäß der Regel wird der Bereich rechts des Begriffs „Rechnung“ ausgelesen. Die Regel wird den Wert

Datum: 30.06.2014

oder eventuell sogar

Datum: 30.06.2014Kunde: 25806

beinhalten und den Text in den Index stellen (sofern der Datentyp der Maske das zulässt). Man muss sich bewusst werden, dass das Belegtraining exakt das tut, was mit den Regeln vorgeben wird. Die Regel kann von sich aus nicht wissen, dass eine Zahl erwartet wird.

Deshalb ist es in nahezu allen Fällen sinnvoll, die Regeln mit weiteren Einstellungen genauer zu formulieren und so dem Computer bei der Erkennung zu unterstützen, was richtig oder falsch ist. Wie das funktioniert, wird in den nachfolgenden Kapiteln näher beschrieben.

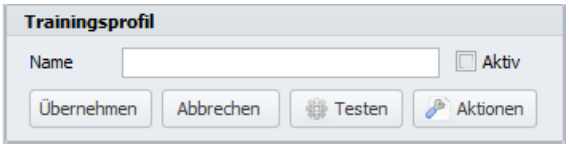
Schulung

Unabhängig von dieser Dokumentation ist eine Schulung durch einen fachkundigen Supporter oder bereits geschulten Mitarbeiter im eigenen Unternehmen anzuraten. Das Belegtraining mag auf den ersten Blick spielerisch einfach wirken, sorgt ohne entsprechendes Hintergrundwissen aber oft zu unsicheren Regeln und damit verbunden späteren Fehlerkennungen und Folgeproblemen (z.B. zeitintensive Nacharbeit/-korrekturen) im Alltagsgeschäft.

Deshalb sprechen Sie Ihren DMS-Betreuer an, ob eine Schulung oder Workshop zum Belegtraining angeboten werden kann.

Profileigenschaften

In diesem Abschnitt werden die Eigenschaften eines Profils erklärt, ohne dass auf praxisbezogene Fälle eingegangen wird.



Name
Der Name ist beliebig und kann frei gewählt werden, muss aber über alle Profile hinweg eindeutig sein. Der Name eines Profils taucht später in der Stapelverarbeitung in Klammern hinter dem Dokument auf. Hier eigenen sich Namen von Lieferanten oder die Bezeichnung von Belegtypen.

Aktiv
Sobald ein Profil mit Haken bei „Aktiv“ gespeichert wird, steht es in der normalen Stapelverarbeitung (auch für andere SCAN-Arbeitsstationen nach deren Neustart) zur Verfügung. Solange der Haken nicht gesetzt ist, wird das Profil außerhalb des Trainingsmodus ignoriert. So kann man in Ruhe (auch mit Unterbrechungen) ein Trainingsprofil erst freigeben, wenn es komplett getestet wurde.

Übernehmen
Speichert das Profil auf dem Server.

Abbrechen
Verwirft die gemachten Änderungen am Profil.

Testen
Testet das geladene Profil gegen das derzeit im Trainingsmodus befindliche Dokument.

Aktionen
Sammelschaltfläche, die nach einem Klick ein Kontextmenü mit weiteren eher selten genutzten Aktionen öffnet:

- Profil kopieren
  Legt eine Kopie des Profils auf dem Server an.
- Profil löschen
  Löscht das Profil unwiderruflich vom Server.
- Importieren…
  Importiert fremde Profile aus einem Verzeichnis heraus.
- Exportieren…
  Exportiert eigene Profile in ein Verzeichnis.
- Eindeutigkeit des Dokuments prüfen
  Prüft, welche Trainingsprofile auf das derzeit im Trainingsmodus befindliche Dokument passen. Idealerweise sollte immer nur ein Profil gültig sein.

Regeleigenschaften

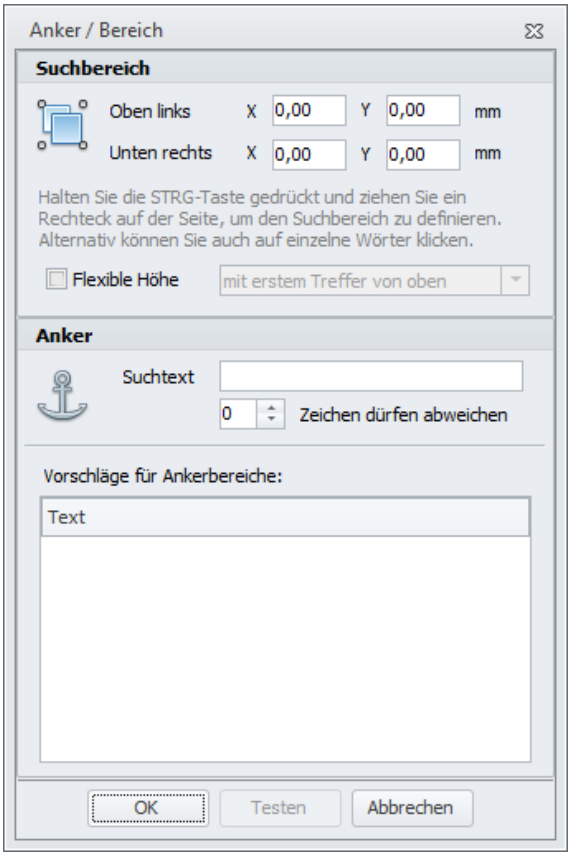
In diesem Abschnitt werden die Eigenschaften einer Regel oberflächlich erklärt, ohne dass auf praxisbezogene Fälle eingegangen wird.

Allgemein

Name
Der Name ist beliebig und kann frei gewählt werden. Über den Name wird später in Formelvalidierungen auf die Regeln zugegriffen. Es sollten schlagkräftige und sinnvolle Namen wie *Belegdatum* oder *Endbetrag* usw. verwendet werden.

Pflicht
Sobald eine Regel als Pflicht markiert wird, muss sie später gültig sein, damit die Belegerkennung das Profil akzeptiert.

Datentyp
Bei dem Datentyp steht *Text*, *Numerisch* und *Datum* zur Verfügung. Diese Einstellung wirkt sich später auf Validierungen aus, wenn man mit wertbezogenen Vergleichen (größer/kleiner) oder Formeln arbeitet.



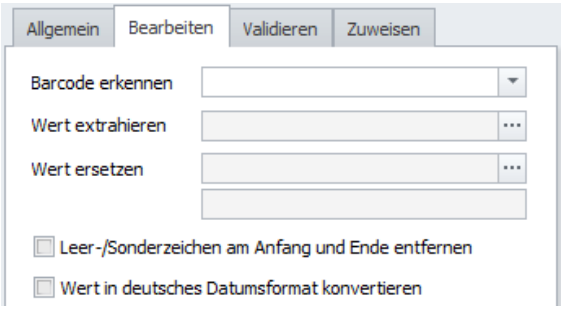
Anker
Der Anker ist ein Orientierungspunkt auf dem Dokument, von dem aus der Bereich relativ Texte extrahieren kann. Er setzt sich aus einem Suchbereich und einem Suchtext zusammen.

Flexible Höhe ermöglicht es, einen Anker auf allen Seiten des Dokuments zu suchen, entweder beginnend auf der erste Seite nach unten oder anders herum. Ohne diese Einstellung gelten Anker nur für die erste Seite. Diese Funktion ist z.B. für die Ermittlung von Endsummen wichtig.

n Zeichen dürfen abweichen ist eine Einstellung, die eine einfache Art der Ähnlichkeitssuche (Fuzzy) erlaubt. Damit lassen sich Ungenauigkeiten bei der Texterkennung bzgl. bestimmter Buchstaben kompensieren und somit die Erkennungsquote verbessern.

Bereich
Ein Rechteck auf dem Dokument, der den zu extrahierenden Textbereich definiert. Hier werden später nicht nur exakt die Zeichen aus dem Rechteck berücksichtigt, sondern die Schnittmenge mit den OCR-Paragrafen ist relevant (dazu später mehr).

Bearbeiten



Barcode erkennen
Sobald ein Barcode-Typ ausgewählt wird, schaut die Erkennung nach, ob sich im Regelbereich ein passender Barcode befindet und stellt den Barcodeinhalt als Regelwert zur Verfügung.

Wert extrahieren
Über die Schaltfläche […] öffnet sich ein separater Dialog, in dem man über reguläre Ausdrücke nur bestimmte Teile des erkannten Textes als Regelwert übernehmen kann.

Wert ersetzen
Über die Schaltfläche […] öffnet sich ein separater Dialog, in dem man über reguläre Ausdrücke bestimmte Teile des Textes mit anderen Textteilen ersetzen kann. Dieser Bereich kann zusätzlich genutzt werden, um den Aufbau bzw. die Reihenfolge von Textteilen zu verändern (z.B. für Datumswerte interessant).

Leer-/Sonderzeichen am Anfang und Ende entfernen
Durch die Texterkennung kann es vorkommen, dass Leerzeichen sowie Steuerzeichen (Tab, Zeilenumbruch, Zeilenvorschub) vor und hinter dem eigentlichen Text stehen. In den meisten Fällen sind diese Informationen unerwünscht, weshalb diese Option im Standard aktiviert ist.

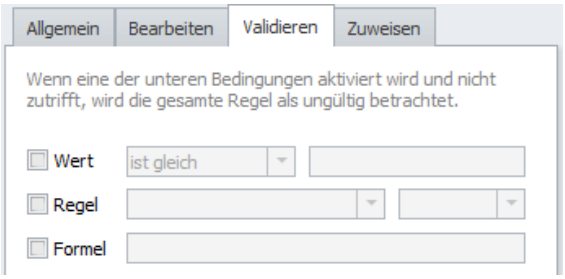
Wert in deutsches Datumsformat konvertieren
Manche Belege weisen keine genormten Schreibweisen von Datumsangaben auf. So sind verschiedene Schreibweisen im Umlauf:

03. Januar 2015
3. Jan 2015
3. January 2015

Da man für die weitere Verarbeitung in Validierungen und DMS-Indizes immer ein deutsches Datumsformat benötigt, steht man vor einem Problem, was auch durch „Wert ersetzen“ nicht mehr zu bewältigen ist. Setzt man hier einen Haken, versucht die Erkennung mit mehr als einem Dutzend verschiedenen Ansätze, aus dem Text ein Datum zu konvertieren - angefangen bei deutschen und englischen Langschreibweisen hin zu unterschiedlichen Trennern und Aufbauten.

Validieren

Alle hier aktivierten Bedingungen wirken sich direkt auf die Gültigkeit der Regel aus. Ist eine der Bedingungen nicht erfüllt, ist die Regel automatisch ungültig.

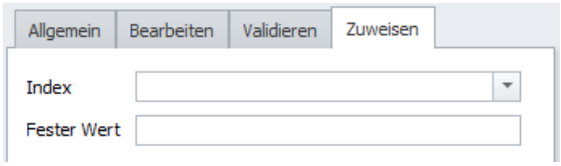


Wert
Sobald ein Barcode-Typ ausgewählt wird, schaut die Erkennung nach, ob sich im Regelbereich ein passender Barcode befindet und stellt den Barcodeinhalt als Regelwert zur Verfügung.

Regel
Hier kann die aktuelle Regel in Abhängigkeit zu einer anderen Regel im Regelwerk gesetzt werden. Das kann dazu genutzt werden, alternative Regeln zu definieren.

Formel
Es kann eine Formel definiert werden, die mit einem boolischen Ergebnis über die Regelgültigkeit entscheidet. Formeln eignen sich für mathematische Abhängigkeiten von 2-n Regelwerten. Bekanntestes Beispiel ist die rechnerische Prüfung von Endsummen (Netto + Mwst = Brutto).

Zuweisen



Index
Der Regelwert kann einem bestimmten Index zugewiesen werden. Es stehen nur die Felder der jeweils gewählten Maske zur Verfügung.

Fester Wert
Sofern eine Regel gültig ist, kann unabhängig vom vorher erkannten Text ein fester Wert gesetzt werden. So kann eine Umsatzsteuer-Identnr. ausgelesen, aber in den Index später die Lieferantennummer eingetragen werden.

Regeldefinitionen für Fortgeschrittene

In diesem Abschnitt werden Sie in die erweiterten Konfigurationsmöglichkeiten der Regeln eingeführt.

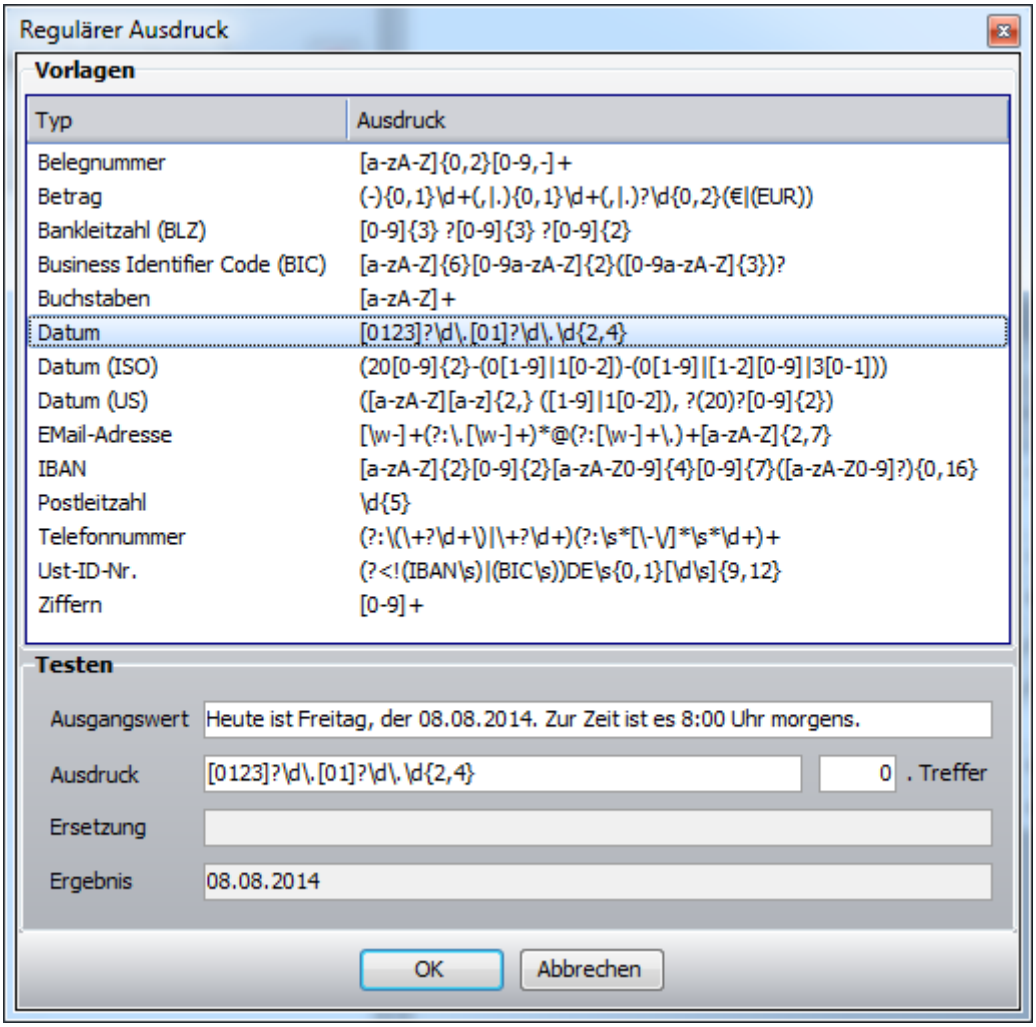
Komplexere Erkennung von Werten

Während das Setzen von Ankern und Bereichen für manche Belegstrukturen bereits ausreicht, stößt man schnell auf Problemfälle, wo das einfache Umranden oder Anklicken von Textzeilen nicht mehr ausreicht.

Extraktion mit regulären Ausdrücken

In vielen Fällen werden Sie damit konfrontiert werden, dass zu viel unerwünschter Text bei einer Regel erkannt wird. Das liegt entweder daran, dass die OCR bestimmte Texte als zusammenhängenden Paragrafen erkennt oder der gewünschte Wert mitten in einem Satz steht. Immer wenn es darum geht, bestimmte Informationen aus einem Text heraus zu holen, kommt die Einstellung „Wert extrahieren“ in der Regel zum Tragen.

Mit einem Klick auf […] im Feld „Wert extrahieren“ öffnet sich der Editor für den regulären Ausdruck. Der Dialog unterstützt sie mit Vorlagen und einem Testbereich, der sich während des Tippens automatisch aktualisiert:



Das Beispiel im Bild zeigt, wie man ein Datum mitten aus einem erkannten Text herauslöst. Der Umfang von regulären Ausdrücken ist enorm, deren Potenzial ebenfalls. Eine Erläuterung des Funktionsumfangs würde diese Dokumentation sprengen. Weitere Informationen zu regulären Ausdrücken finden Sie daher auf Wikipedia unter dem Link

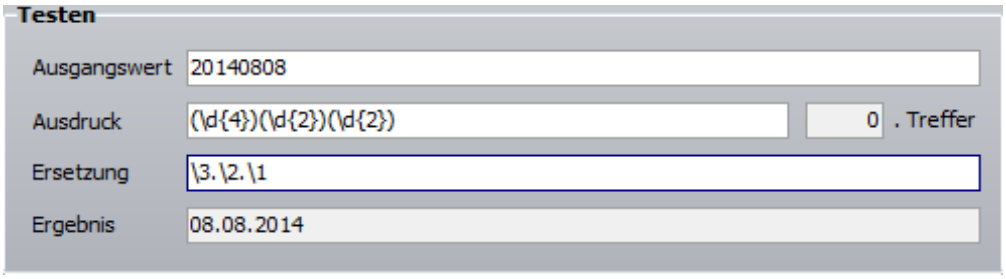
http://de.wikipedia.org/wiki/Regul%C3%A4rer_Ausdruck

Ersetzungen mit regulären Ausdrücken

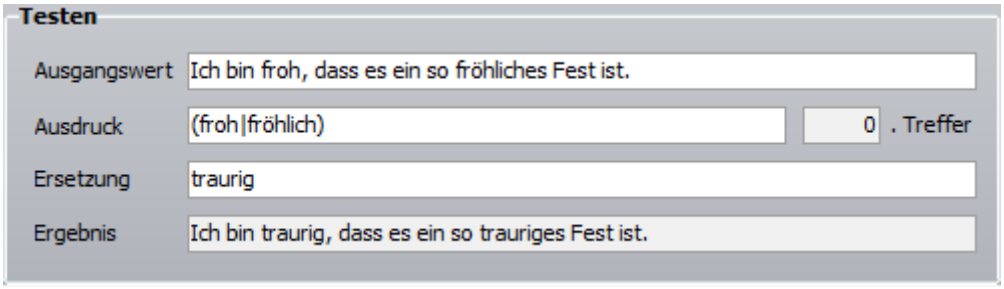
Neben dem Herauslösen von Informationen kann man diese nachträglich noch verändern. Das Thema ist ebenso ergiebig wie der reguläre Ausdruck an sich. Daher bleibt neben den nachfolgenden Beispielen der Verweis auf eine Suche im Internet nach weiteren Informationen.

Euro-Symbol entfernen/ersetzen:

Datumsformat normieren:



Textersetzung:



Validierung

Die Gültigkeit eines Profils leitet sich aus der Gültigkeit des Regelwerks ab, das wiederum abhängig von den einzelnen Regeln ist. Abgesehen vom eigentlichen Regelwert, den man auslesen, extrahieren, ersetzen und fest definieren kann, spielt die Gültigkeit einer Regel bei umfangreicheren Trainingsprofilen einen große Rolle.

Gültigkeit einer Regel

Die Erkennung kennt zwei Zustände:

1. Die Regel ist gültig
2. Die Regel ist ungültig.

Die drei wichtigsten Auswirkungen der Gültigkeit sind

- das nur eine gültige Regel seinen Wert in einen Index überträgt.
- das eine ungültige Pflichtregel das gesamte Profil verwirft.
- das Abhängigkeiten zwischen Regeln entsprechend gewertet werden.

Eine Regel ist per Definition erst einmal ungültig. Das ändert sich, wenn die primäre und eine der vier sekundären aber optionalen Bedingungen erfüllt sind:

- Primär: Der Regelwert muss ein Zeichen beinhalten.
- Sekundär (optional): Der Regelwert muss dem Datentyp entsprechen.
- Sekundär (optional): Der Regelwert muss dem Vergleich entsprechen.
- Sekundär (optional): Die abhängige Regel muss die hinterlegte Gültigkeit haben.
- Sekundär (optional): Die Formel muss wahr sein.

Regel ohne Wert

Beinhaltet der Regelwert kein Zeichen, ist die Regel automatisch ungültig, unabhängig weiterer Bedingungen. Sollte eine Regel ohne Auslesen eines Wertes vom Dokument notwendig sein, kann man einfach den Bereich für die gesamte erste Seite des Dokuments definieren (geht am schnellsten über einen Mausklick auf das Bereichssymbol oben links im Bereichsdialog).

Datentyp

- Text
  Erlaubt alle Regelwerte und ist vorbelegt.

- Numerisch
  Setzt zwingend voraus, dass nach der Bearbeitung des Regelwertes eine Ganzzahl oder Fließkommazahl in deutscher Normierung vorliegt.

- Datum
  Setzt zwingend voraus, dass nach der Bearbeitung des Regelwertes ein Datum in deutscher Normierung vorliegt.

Wertvergleich

- ist gleich
  Der Regelwert muss dem Vergleichswert entsprechen, wobei Groß- und Kleinschreibung ignoriert wird.

- ist ungleich
  Der Regelwert muss vom Vergleichswert abweichen, wobei Groß- und Kleinschreibung ignoriert wird.

- enthält
  Der Vergleichswert muss im Regelwert vorkommen, wobei Groß- und Kleinschreibung ignoriert wird.

- beginnt mit
  Der Regelwert muss mit dem Vergleichswert beginnen, wobei Groß- und Kleinschreibung ignoriert wird.

- endet auf
  Der Regelwert muss mit dem Vergleichswert enden, wobei Groß- und Kleinschreibung ignoriert wird.

- ist größer als
  Der Regelwert muss größer dem Vergleichswert sein, wobei Groß- und Kleinschreibung ignoriert wird.
  Für einen numerischen Vergleich von Beträgen muss als Datentyp „Numerisch“ hinterlegt sein.
  Für einen zeitbasierenden Vergleich von Datumsangaben muss als Datentyp „Datum“ hinterlegt sein.

- ist kleiner als
  Der Regelwert muss kleiner dem Vergleichswert sein, wobei Groß- und Kleinschreibung ignoriert wird.
  Für einen numerischen Vergleich von Beträgen muss als Datentyp „Numerisch“ hinterlegt sein.
  Für einen zeitbasierenden Vergleich von Datumsangaben muss als Datentyp „Datum“ hinterlegt sein.

- entspricht
  Der Regelwert muss auf den regulären Ausdruck passen.

Regelabhängigkeit

- Andere Regel ist gültig
  Die aktuelle Regel ist gültig, wenn die andere Regel ebenfalls gültig ist.

- Andere Regel ist ungültig
  Die aktuelle Regel ist gültig, wenn die andere Regel ungültig ist.

Sonderfall:

Wenn Pflichtregel B abhängig davon ist, dass Pflichtregel A ungültig ist, greift eine Sonderregelung, da per Definition Regel B immer ungültig wäre. In der Folge würde das Profil niemals von der Belegerkennung herangezogen werden. In dieser Konstellation vertreten sich beide Regeln gegenseitig. Es gilt für die verschiedenen Fälle folgendes:

1. Fall: Regel A ist gültig.
   Das Profil ist weiterhin gültig. Regel A entscheidet. Regel B ist ungültig, wird aber ignoriert.

2. Fall: Regel A ist ungültig, Regel B ist ungültig.
   Das Profil ist ungültig, da keine der beiden Pflichtregeln gültig ist.

3. Fall: Regel A ist ungültig, Regel B ist gültig.
   Das Profil ist weiterhin gültig. Regel B entscheidet. Regel A ist ungültig, wird aber ignoriert.

Formel

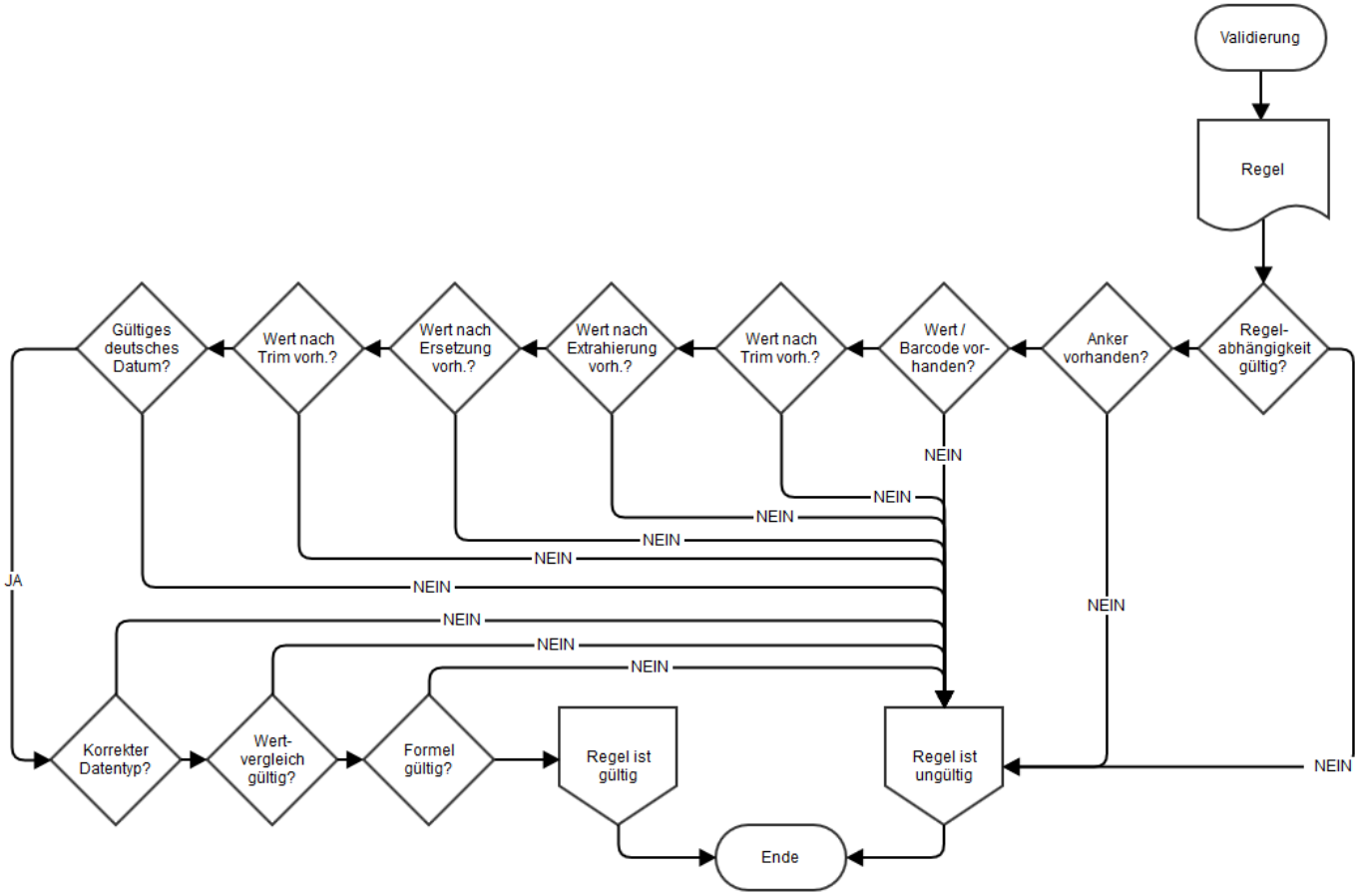
Bei der Formelvalidierung handelt es sich um die Möglichkeit, eine nahezu beliebig komplexe Gleichung mit mathematischen Grundrechenarten inkl. Berücksichtigung der Klammerrechnung zu hinterlegen. Neben der Nutzung von konstanten Werten kann man durch Platzhalter auf die Werte anderer Regeln zugreifen:

- (3 + 2) * 5 = 25
- (3 + 2) * 5 > 20
- 10*5 = 25 + 25
- %RegelA% + %RegelB% = %RegelC%
- %Netto% + %MwSt% = %Brutto%

Bei dem Zugriff auf Werte anderer Regeln müssen diese bereits validiert worden sein, bevor die Formel verarbeitet wird (sprich im Regelwerk des Profils weiter oben stehen). Man kann in der Formel auch auf sich selbst zugreifen.

Reihenfolge der logischen Prüfung für die Gültigkeit einer Regel

Das folgende Diagramm zeigt die interne Reihenfolge der Prüfungen, wobei ein Schritt entfällt (und automatisch mit „Ja“ beantwortet werden kann) wenn dieser optional ist und in der Regel nicht aktiviert wurde.

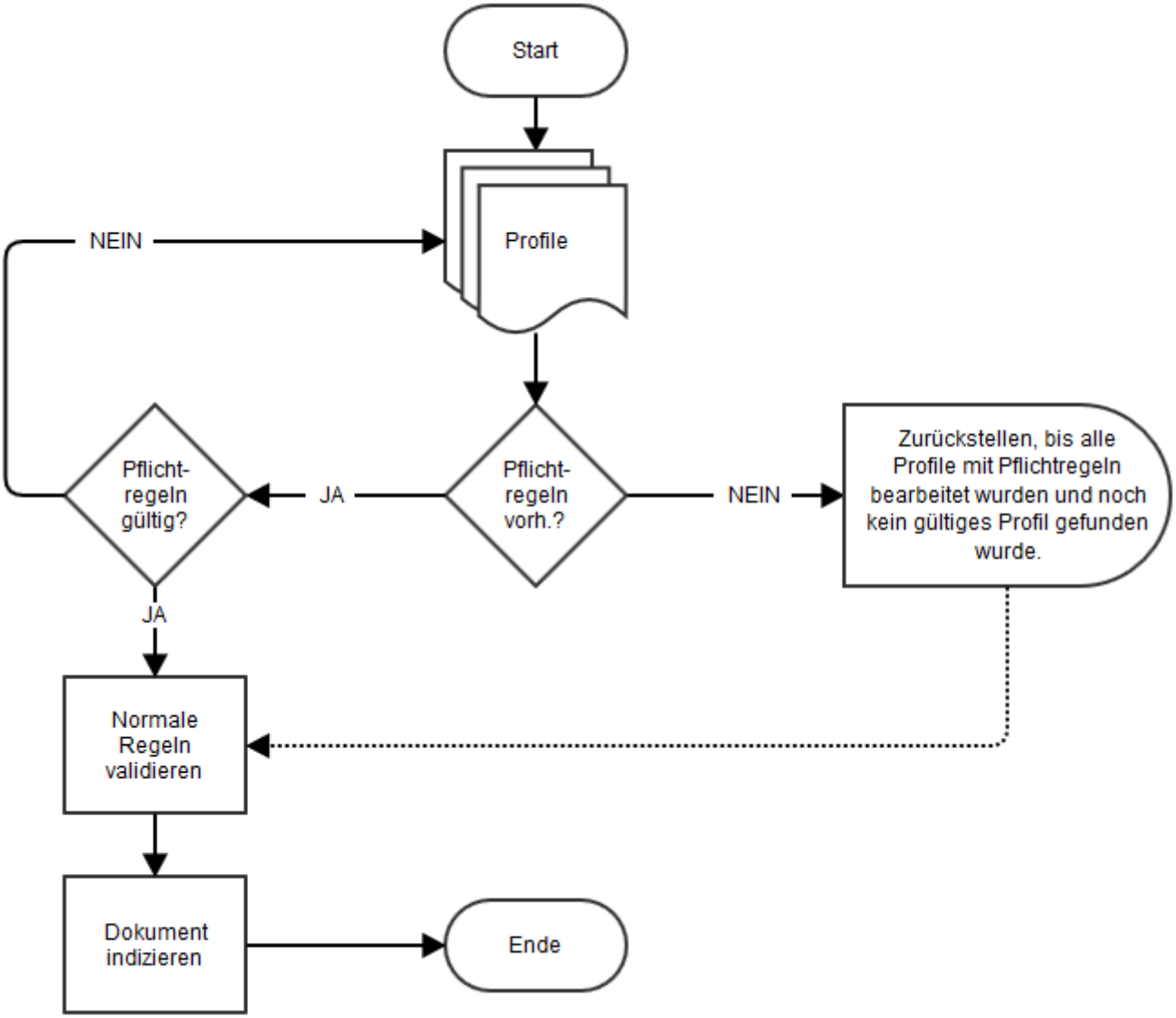


Gültigkeit eines Profils

Die Erkennung kennt zwei Zustände:

1. Das Profil ist gültig
2. Das Profil ist ungültig.

Bei der Analyse von Dokumenten findet folgende Abarbeitung der Profile statt:

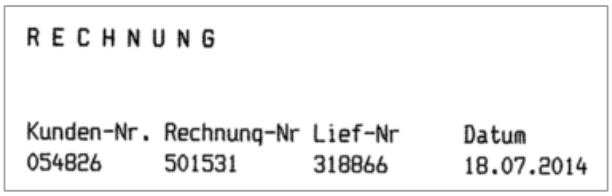


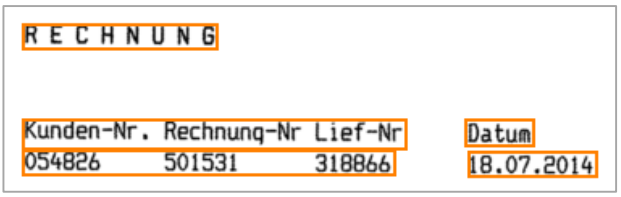
Praxisnahe Beispiele

Dieser Abschnitt soll in erster Linie als Anregung für fortgeschrittene Anwender dienen. Hier werden anhand einiger alltagsüblicher Beispiele gezeigt, wie man bestimmte Herausforderungen lösen kann.

1. Zahlenreihe

Gegeben ist folgender Belegabschnitt mit der Anforderung, die Rechnungsnummer auszulesen:





Analyse / Vorüberlegungen:

a) Bei dem Begriff „R E C H N U N G“ kann nicht mit Sicherheit gesagt werden, dass die Texterkennung tatsächlich immer sauber überall Leerzeichen einfügt. Es kann genauso vorkommen, dass die OCR die Leerzeichen ignoriert und „RECHNUNG“ als Ergebnis liefert.

b) Aufgrund der Belegqualität wird der Buchstabe G in dem Begriff „Rechnung-Nr“ hin und wieder als Q interpretiert, da der Bogen des kleine g auf einigen Kopien nicht sauber erkennbar ist.

c) Die Belegerkennung fasst die Daten in einem Paragraf zusammen, so dass beim Auslesen der Daten in der Regel der gesamte Zahlenblock steht: „054826 501531 318866“

Lösungsvorschlag:

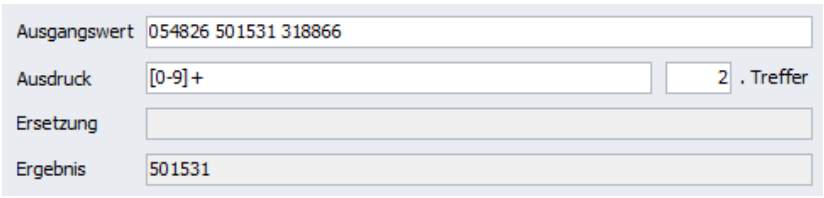
Unter der Annahme, dass die Rechnungsnummer numerisch ist und bleibt, kann man folgende Regel definieren:

a) Datentyp auf „numerisch“ stellen.

b) Anker mit Suchbegriff „Rechnung-Nr“ in dem entsprechenden Suchbereich definieren und erlauben, dass maximal 2 Zeichen abweichen dürfen. Das kleine G und der Bindestrich sind in diesem Fall als problematisch einzustufen.

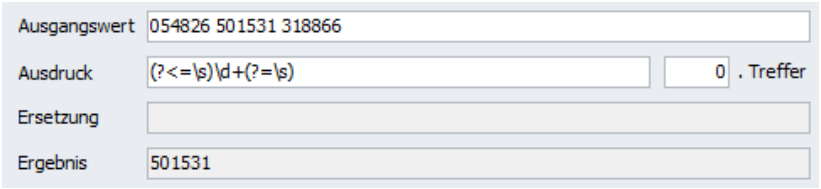
c) <u>Alternative A:</u>

Man definiert unter „Wert extrahieren“, dass man eine Zahl erwartet und den 2. Treffer innerhalb der Daten möchte:



<u>Alternative B:</u>

Man definiert unter „Wert extrahieren“, dass man eine Zahl erwartet und das vor und hinter der Zahl ein Leerzeichen stehen muss:



2. Flexible Zahlenreihe + rechnerische Prüfung

Folgender Beleg ist vorgegeben:

Zahlbar bis:			skontofähiger Betrag	Netto	MwSt-%	MwSt	Endbetrag EUR
Ausgleich per Lastschrift			699,72	588,00	19,00	111,72	**699,72**
18.07.2014	3,00%	Skonto =	20,99				

Die Anforderung ist, dass wir Netto, MwSt und Endbetrag auslesen.
Die Paragrafenaufteilung ist wie folgt zu sehen:

Zahlbar bis:			skontofähiger Betrag	Netto	MwSt-%	MwSt	Endbetrag EUR
Ausgleich per Lastschrift			699,72	588,00	19,00	111,72	**699,72**
18.07.2014	3,00%	Skonto =	20,99				

Analyse / Vorüberlegungen:

a) Die Struktur der Daten ist ungünstig, da der Abstand des skontofähigen Betrags beinahe genauso groß ist wie die anderen Werte untereinander. Hier muss davon ausgegangen werden, dass die Texterkennung auf anderen Belegen durchaus dazu tendieren könnte, aus allen Beträgen einen gemeinsamen Paragrafen zu bilden. Das kann dann passieren, wenn der Nettobetrag deutlich größer wird, z.B. 20.000€ beträgt und somit der Abstand beider Zahlenblöcke verringert wird. Dasselbe Problem besteht bei MwSt und Endbetrag.

b) Wir müssen zudem davon ausgehen, dass es andere Belege des Lieferanten gibt, wo kein skontofähiger Betrag vorhanden ist und dementsprechend auch nicht ausgewiesen wird (sprich komplett auf dem Beleg fehlen kann).

c) Um das ganze Konstrukt später abzusichern, sollte eine rechnerische Prüfung gemäß der Formel

Netto + MwSt = Brutto

durchgeführt werden.

Lösungsvorschlag:

Noch in Arbeit…