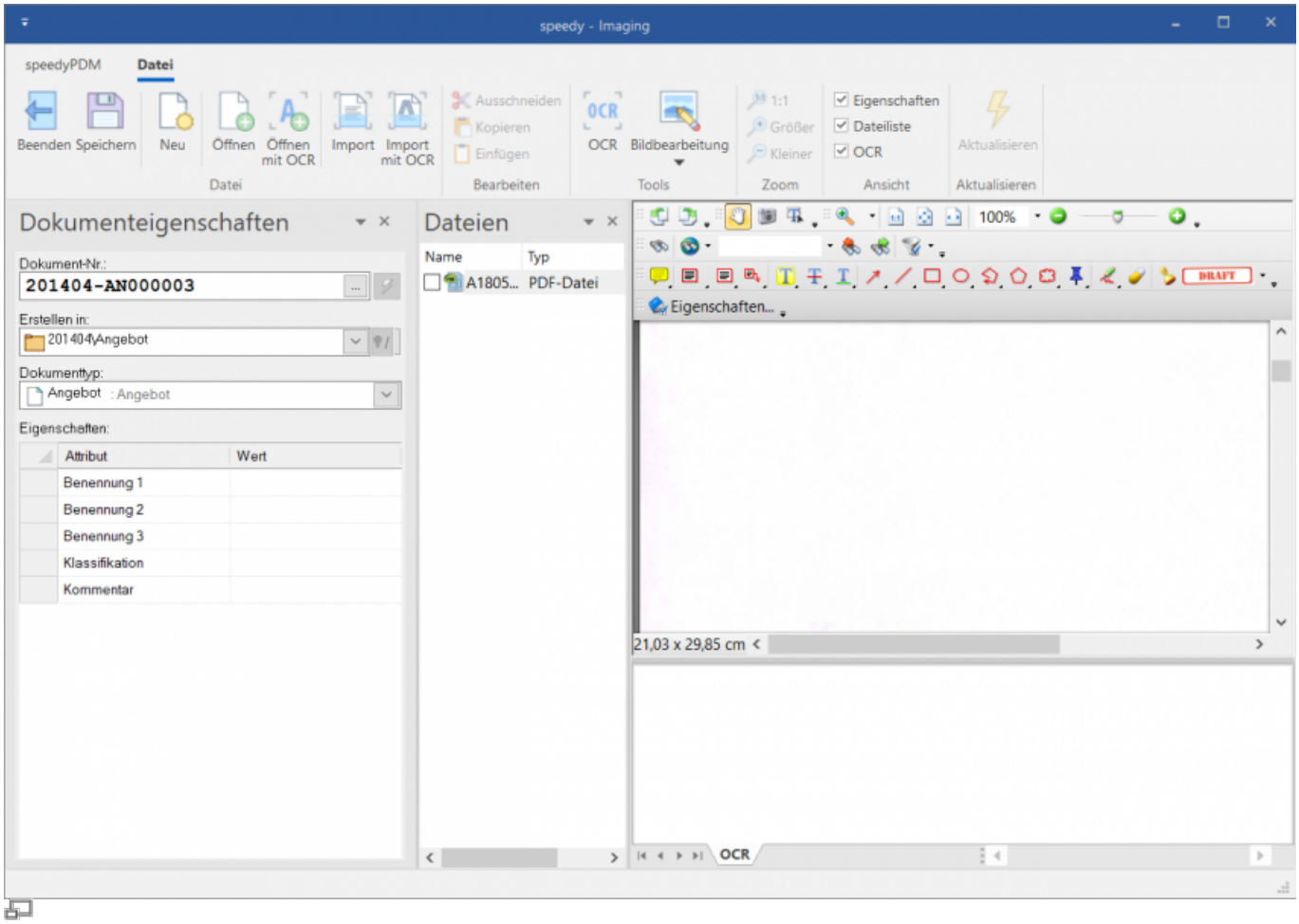
Imaging/OCR

Mit speedyIMAGING werden PDF, TIF oder andere Bilddateien importiert und verschlagwortet. Die Dateien können per Dateiauswahl ausgewählt werden oder direkt von einem Scanner geladen werden.

Durch eine optische Texterkennung (OCR - Optical Character Recognition) und anschließender Auswertung eines Regelwerks kann die Indizierung weitestgehend automatisiert werden.

Imaging

Die zu importierenden Dateien werden per Dateiauswahl eingesammelt oder direkt per Scanner eingelesen.



speedyIMAGING mit einer Liste zu importierender PDF Dateien

Die Dateien Liste zeigt alle gewählten Dateien an. Die markierte Datei wird zur einfachen Verschlagwortung im großen Vorschaufenster dargestellt.
Im Bereich „Dokumenteigenschaften“ werden die Dokumentinformationen für die jeweilige Datei definiert.
Mit dem Speichern Befehl wird die aktuelle Datei in speedyPDM abgelegt.

Das Eigenschaften Fenster passt sich je nach gewähltem Dokumenttyp an und stellt alle

Dokumenteigenschaften dar. Pflichtfelder werden hierbei farblich markiert.

OCR

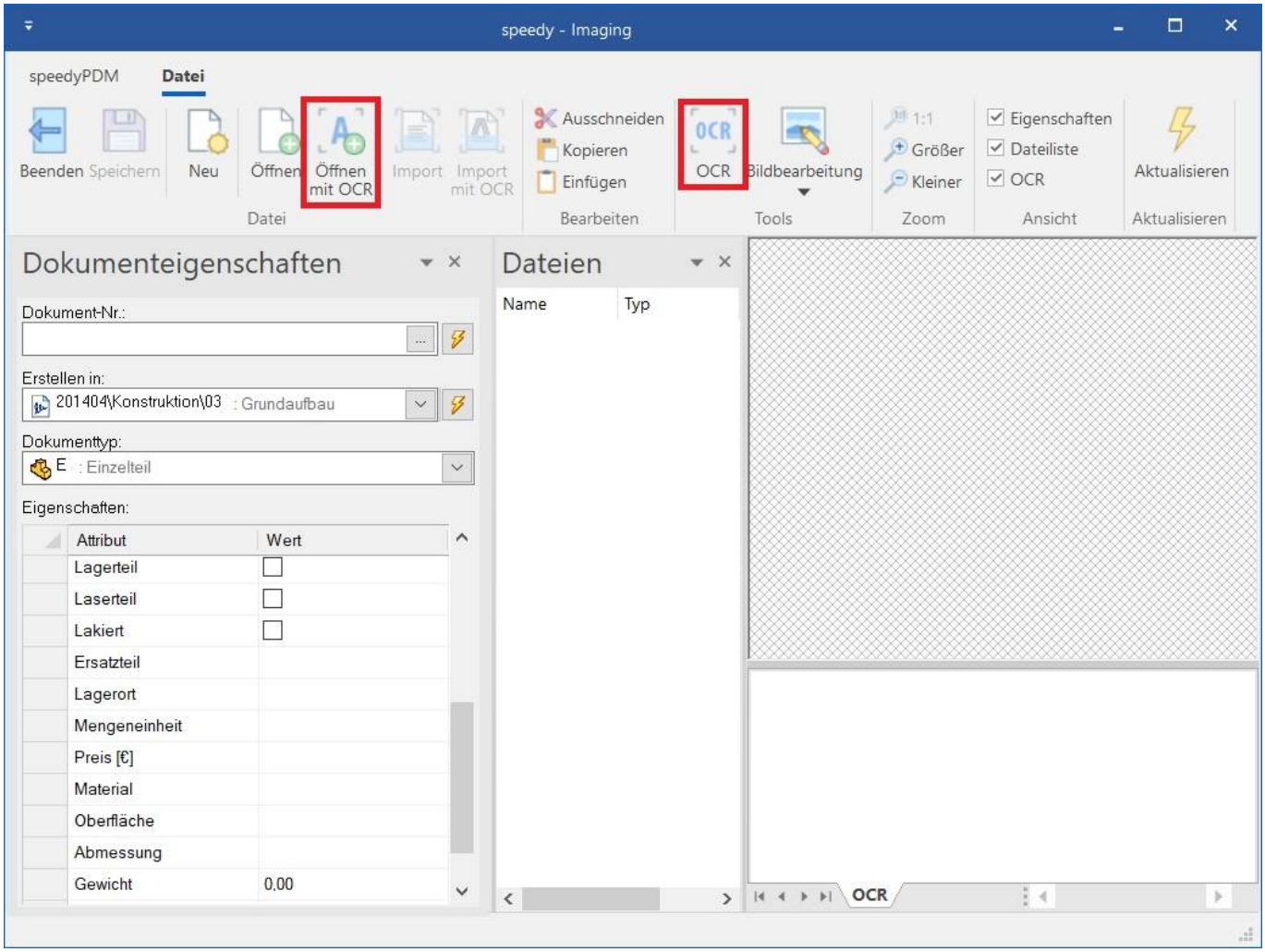
Mit Hilfe von OCR können eingescannte Dokumente, Bilddateien oder PDF Dateien digitalisiert werden und wieder in Text umgewandelt werden.
Durch ein Regelwerk können bestimmte Texte als Dokumenteigenschaften erkannt und zugeordnet werden. Damit ist es möglich Dokumente gleichen Aufbaus automatisch oder zumindest halbautomatisch zu erkennen und in speedyPDM abzulegen.
So können z.B. Eingangsrechnungen, Lieferscheine oder Prüfprotokolle automatisch in speedyPDM eingepflegt werden.

Ablauf der OCR Indizierung

Die Texterkennung in speedyPDM erfolgt auf Basis einer TIF Bilddatei. Egal welches Dateiformat die zu importierende Datei hat (PDF, JPG, BMP, PNG, ...) erfolgt eine Umwandlung in TIF. Dies geschieht automatisch im Hintergrund.
Mit Hilfe der optischen Zeichenerkennung wird nun die TIF Datei quasi in eine Textdatei umgewandelt. Nun werden die definierten Texterkennungs- und Zuordnungsregeln überprüft.

Halbautomatische Indizierung (Imaging)

Die halbautomatische Indizierung erfolgt im Programmfenster von speedyIMAGING.
Starten Sie speedyIMAGING und öffnen die zu indizierenden Dateien mit dem Befehl „Öffnen mit OCR“.



Imaging

Die Dateien durchlaufen nach der Auswahl die optische Texterkennung (OCR) und werden mit Hilfe des Regelwerks untersucht.

Alternativ können die Dateien auch zuerst mit dem Befehl „Öffnen“ in das Programmfenster geladen werden. Anschließend kann für jede Datei einzeln die Texterkennung erfolgen indem der Befehl „OCR“ gewählt wird.

Entspricht der erkannte Text einer Regel so werden die zugeordneten Werte im Eigenschaftenfenster dargestellt.
Je besser die entsprechende Regel definiert ist, umso mehr Dokumenteigenschaften werden erkannt und bereits vorausgefüllt.
Fehlende Eigenschaften können manuell ergänzt werden.

Vollautomatische Indizierung

Die vollautomatische Indizierung kommt dann zum Einsatz, wenn ausreichend gute Texterkennungsregeln definierbar sind und z.B. Dokumente automatisch nach dem Scannen abgearbeitet werden sollen.
Typischer Anwendungsfall sind hierbei z.B. Eingangs-Rechnungen oder Eingangs-Lieferscheine. Die Dokumente werden z.B. durch einen vordefinierten Scann-Prozess in einem Netzlaufwerk abgelegt. Der speedy-Spooler greift die Dateien auf und übergibt diese der OCR Erkennung (dwImportOCR.exe). Genügt die Datei einer definierten Regel wird sie automatisch in speedyPDM abgelegt.

Tritt ein Fehler bei der Erkennung auf oder passt keine der Regeln wird die Datei ein Fehler-Verzeichnis verschoben.

Regelwerk/Einrichtung

Um eine OCR Indizierung der Dokumente mit automatischer Verschlagwortung in speedy durchführen zu können muss ein Regelwerk vorhanden sein. Um dieses Regelwerk zu erstellen gibt es einen Regelwerk-Editor. Der Regelwerk-Editor befindet sich im speedy-Admin unter **Konfiguration→OCR-Regeln**. Im Editor ist es möglich eine Datei (z.B. eine Rechnung, ein Angebot,…) für die eine Regel erstellt werden soll zu öffnen. Klicken sie dazu den Öffnen-Button in der Ribbon-Leiste und wählen sie eine Datei aus. Nach dem sie die Datei gewählt haben startet der OCR-Indizierungsmechanismus und liest die Datei ein. Nachdem dieser fertig ist und ihre Datei indiziert hat sehen sie ihre Datei in der Vorschau[4] und den extrahierten Text im OCR-Textfenster[3]. Falls schon Regeln definiert wurden und diese zur ausgewählten Datei passen, werden die Regeln farbblich in der Regelstruktur[1] markiert. Somit kann der „Weg“ der Indizierung nachverfolgt werden4.

Regel

Über den Button **Neue Regel** kann eine neue Regel definiert werden. Die Regel kann im Eigenschaften-Fenster[2] bearbeitet und angepasst werden. Wurde die Regel fertig definiert kann über den **Analysieren** Button in der Ribbon-Leiste eine erneute Indizierung durchgeführt werden. Dadurch können Sie verifizieren ob die eben erstellte Regel auch die gewünschten Ergebnisse liefert.

Die Beschreibung einer Regel unterscheidet sich durch 3 Typen:
1. Fester Standardwert für speedy-Eigenschaft
Ein fester Standardwert für eine speedy-Eigenschaft kann Sinnvoll sein um z.B. eine Vorsortierung durchzuführen. Es kann damit zum Beispiel ein Dokumenttyp gesetzt werden wenn über eine Regex-Abfrage ein bestimmtes Muster gefunden wurde.
Einstellungen:

- Pattern → Das gewünschte Regex Pattern
- Eigenschaft → speedy Eigenschaftenname z.B. dm_doctype
- Standardwert → dazugehöriger Standartwert z.B. B für Baugruppe
- Flag→ Standardwert

2. Wert der über ein Regex Pattern gefunden wird
Dadurch kann ein Wert der über ein Regex Pattern gefunden wird in eine speedy-Eigenschaft geschrieben werden.
Einstellungen:

- Pattern → Das gewünschte Regex Pattern
- Match-index → 1 (1.Match des Regex Patterns)
- Eigenschaft → speedy Eigenschaftenname z.B. CREATE_DATE
- Flag → Pattern

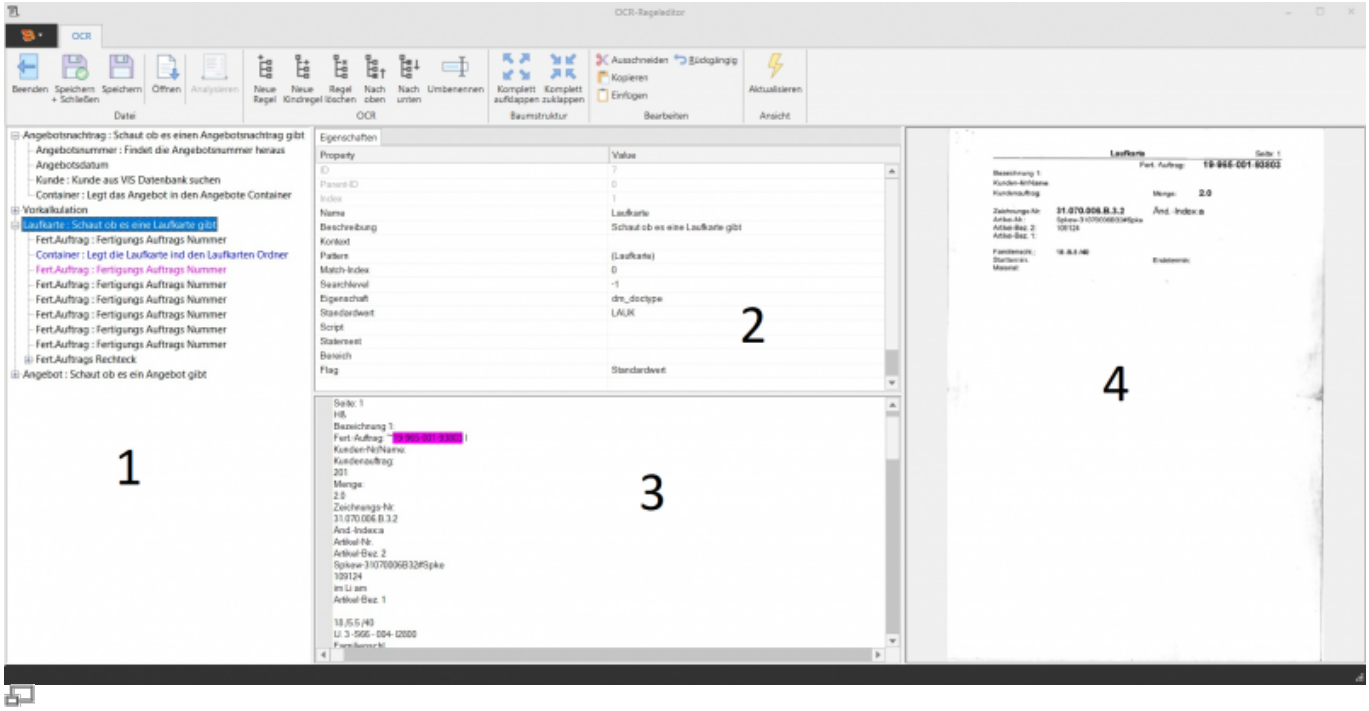
3. Wert der über eine SQL-Abfrage aus einer Datenbank kommt
Es gibt die Möglichkeit aus einer Datenbank entsprechende Werte auszulesen. Ein Beispiel wäre zum Beispiel das auslesen einer Kundennummer um dann über SQL einen Kundennamen heraus zu finden der in einer ERP-Datenbank steckt.

Einstellungen:

- Eigenschaft → speedy Eigenschaftenname z.B. kd_nr
- Statement → SQL Statement. In diesem Statement kann z.B. eine zuvor ermittelte Eigenschaft als Platzhalter in der Form <property> eingefügt werden. Beispiel: ...where ang_nr='<ang_nr>';
- Flag → Statement

- Um ein bestmögliches Ergebnis zu bekommen wird empfohlen immer den gleichen Scanner zu verwenden, da unterschiedliche Auflösungen, das Ergebnis verschlechtern können.
- Um Regex-Ausdrücke zu Testen und die Syntax zu erlernen kann die Website https://regex101.com zur Hilfe genommen werden. (Der erkannte OCR-Text befindet sich im OCR-Fenster[3].

Die einzelnen Fenster des Regel-Editors werden nun folgend erklärt:



OCR-Regeleditor

1 Regelstruktur

Auf der linken Seite im Dialog werden die bisher eingerichteten Regeln angezeigt. Diese Regeln werden in einer Struktur wie sie abgearbeitet werden dargestellt. Das heißt wenn eine Regel zutrifft werden die dazugehörigen „Kind Regeln“ abgearbeitet. Trifft die Regel nicht zu werden die dazugehörigen „Kind Regeln“ auch nicht abgearbeitet. Wenn über den Öffnen-Button eine Datei dazu geladen wurde und eine Regel gefunden wurde wird diese Regel farblich markiert. Damit kann man erkennen welche Regel zu diesem Dokument erkannt wurden.

2 Eigenschaften

Hier werden die dazugehörigen Eigenschaften zu der in der Regelstruktur[1] ausgewählten Regel angezeigt. Die Eigenschaften können dort eingebaut und eingestellt werden.
Hinweis: Wenn ein Bereich ausgewählt wird werden andere Eigenschaften ignoriert. Eigenschaften die in einem Bereich gesucht werden sollen müssen als „Kinder“ der Regel erstellt werden. Um den gewünschten Bereich einzustellen öffnet sich bei klicken der 3 Punkte in der Bereichs-Zeile ein Dialog. In diesem Dialog kann durch klicken und halten ein Viereck definiert werden dass dann den Bereich beschreibt.

3 OCR-Text

In diesem Fenster wird der erkannte OCR-Text eingetragen. Wenn eine Regel unter der Regelstruktur[1] erkannt wurden, diese dort Farblich markiert ist und über eine Pattern-Suche im OCR-Text gefunden wurde, wird der gefundene Text ebenfalls mit der gleichen Farbe markiert.

4 Vorschau

In diesem Fenster wird eine ausgewählte und indizierte Datei zur Vorschau angezeigt.

Konfigurationsparameter

Setting	Beschreibung
ocr.tesseract.exe	Gibt den Pfad zur Tesseract Exe an. Default: .\\tools\\tesseract\\tesseract.exe
ocr.tesseract.tessdata	tessdata Directory Default: .\\tools\\tesseract\\tessdata
ocr.force	Erzwingt die OCR Erkennung und ignoriert evt. vorhandenen Text inner halb von PDF Dateien.
ocr.multitiff	Gibt an ob nur die 1.Seite OCR Indiziert werden soll oder nicht. := 1: Alle Seiten werden Indiziert. := 0: Nur die erste Seite wird Indiziert (Default)
ocr.tesseract.language	Gibt die Sprache an mit der Indiziert werden soll. Default: deu
ocr.tesseract.psm1	Page Segmentation Mode, gibt an mit welcher Einstellung im ersten Durchgang der Tesseract Indizieren soll (tesseract hilfe). Default:0
ocr.tesseract.psm2	Page Segmentation Mode, gibt an mit welcher Einstellung im zweiten Durchgang der Tesseract Indizieren soll (tesseract hilfe). Default:12
ocr.tesseract.oem	Gibt den Ocr-Engine Mode an. Default:3
ocr.tesseract.configvar	Hier kann man die Variablen angeben die für die Konfiguration des Tesseract notwendig sind. Es können mehrere Variablen hintereinander angegeben werden. VAR=Value Default: keine Variablen gesetzt

Setting	Beschreibung
ocr.color.colored	Gibt an ob die Indizierung Farbig oder Schwarz weiß stattfindet. Default: 1 (Farbig)
ocr.deletefiles	Löscht die erzeugten Dateien im Temp Ordner. Default: 1 (Löschen)
ocr.zoom	Zoomfaktor der beim Umwandeln von PDF in TIFF verwendet wird um ein besseres OCR Ergebnis zu erzielen. Standardwert := 2.0

From:
https://wiki.speedy-pdm.de/ - **speedyPDM - Wiki**

Permanent link:
https://wiki.speedy-pdm.de/doku.php?id=speedy:30_modules:imaging_ocr

Last update: **2024/10/11 16:32**