# SQUEEZE Leistungsumfang

Dieses Buch beschreibt den Leistungsumfang von Squeeze.


- Allgemeine Beschreibung
  - Funktion und Architektur
- Die Verarbeitung mit SQUEEZE
  - Dokumenteneingang
  - Bildaufbereitung
  - Texterkennung
  - Klassifikation
  - Erkennung/Extraktion
  - Validierung
  - Export
- Das SQUEEZE Invoice Template
  - Allgemein
  - Stammdaten

# Allgemeine Beschreibung

Eine kurze Einführung in das Squeeze System und dessen Funktionen

Allgemeine Beschreibung

# Funktion und Architektur

SQUEEZE ist eine Mandanten-fähige Software-Lösung für Dokumenten-Klassifikation und Inhaltsextraktion. Dieses Dokument beschreibt die technischen Leistungsmerkmale von SQUEEZE und beschreibt vorrangig die technische Architektur, Eigenschaften und Funktionen.

SQUEEZE wurde im Jahr 2016 designed und aufgebaut und wird seither stetig in einem wachsenden Team weiterentwickelt.

Es ist ein vollständig webbasiertes System zur Verarbeitung von elektronischen Dokumenten, mit wesentlichem Fokus auf die Erkennung und Auslesung von Eingangsdokumenten im B2B Sektor.

# Konzepte des Systemdesigns

Für das Systemdesign wurden folgende Konzepte verfolgt:

- Betriebssystem offen in Bezug auf Linux und Windows Server
- Datenbank-offen in Bezug auf mySQL/MariaDB und Microsoft SQL Server
- Konsequent webbasierter Administrations- und Anwendungszugang
- Vollständige REST-API auf Swagger.io / OpenAPI v3
- Schnelle und problemfreie Installation bei OnPrem-Projekten sowie Update-Fähigkeit
- Wartungsarmer Betrieb, interne Überwachung / Monitoring
- Hohe I in Dritt-Anwendungen
- Einfache Konfiguration über das Webinterface für Partner und Administratoren
- Hohe Stabilität in Bezug auf die Verfügbarkeit
- Hohe Verarbeitungsgeschwindigkeit und Erkennungsqualität

# Systemarchitektur

## SQUEEZE Server

SQUEEZE ist vorrangig eine HTTP-Applikation auf PHP-Basis, welche im Apache Webserver gehostet wird und die serverseitige Verarbeitung steuert. Weiterhin stellt SQUEEZE einen Webclient für die Administration und Endanwendernutzung zur Verfügung.

## SQUEEZE Worker

Worker verstehen sich als eigenständige, asynchron agierende Prozesse, welche Subprozesse (z.B. OCR-Vorgänge) des Systems veranlassen. Das Worker-Konzept bietet eine flexible Skalierung aller Systemarbeitsprozesse (Bildaufbereitung, OCR, Klassifikation, Extraktion) und sind ein wesentliches Element der SQUEEZE Arbeitsweise.

Innerhalb der Software können so viele Worker aktiviert werden, wie lizenziert sind. Die kleinste Standard-Lizenz umfasst 4 Worker. Die maximale Anzahl gleichzeitiger Worker-Prozesse orientiert sich an der Anzahl der Prozessorkerne, über die das Betriebssystem verfügt, respektive sollte der Anzahl der CPU Threads nicht übersteigen. Erfolgreich getestet wurde das System mit bis zu 64 Threads.

Die Worker auch auf anderen Servern zu verteilen, wird in der aktuellen Auslieferung nicht empfohlen auch wenn dies theoretisch möglich wäre. Die Schreibvorgänge der Datenbank können, je nach Ausstattung, an die Grenzen kommen. Hierzu wurden noch nicht ausreichende Tests durchgeführt.

# SQUEEZE Datenbank

Die Squeeze Datenbank enthält die Konfigurationen des Systems sowie die Nutzdaten der Kunden.

Bei der Datenbank handelt es sich um eine relationale Datenbank (siehe Systemvoraussetzungen)

# SQUEEZE Message Queue

Die SQUEEZE Message Queue nutzt RabbitMQ um die Aufgaben der Worker zu verwalten. Die Worker registrieren sich als Konsument bei der Message Queue und arbeiten die Messages ab, sobald sie Zeit dafür haben.

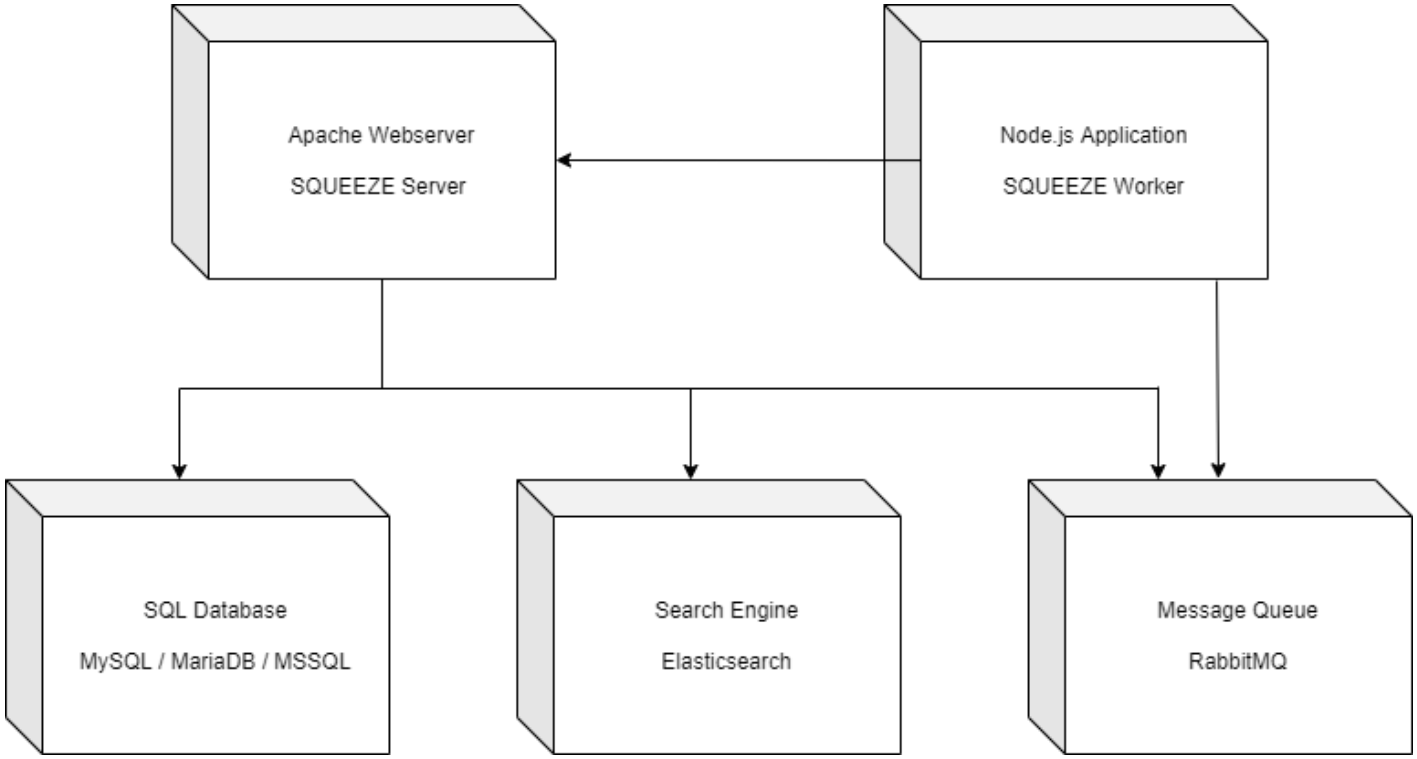
# SQUEEZE Repository

Das SQUEEZE Repository enthält alle eingelesenen Belege / Dokumente. Aktuell werden die Dokumente als Datei auf der Festplatte des SQUEEZE Servers gespeichert. Zukünftig wird es optional die Möglichkeit geben die Dokumente in einem S3 Bucket zu speichern, um auch auf dieser Ebene besser skalieren zu können.

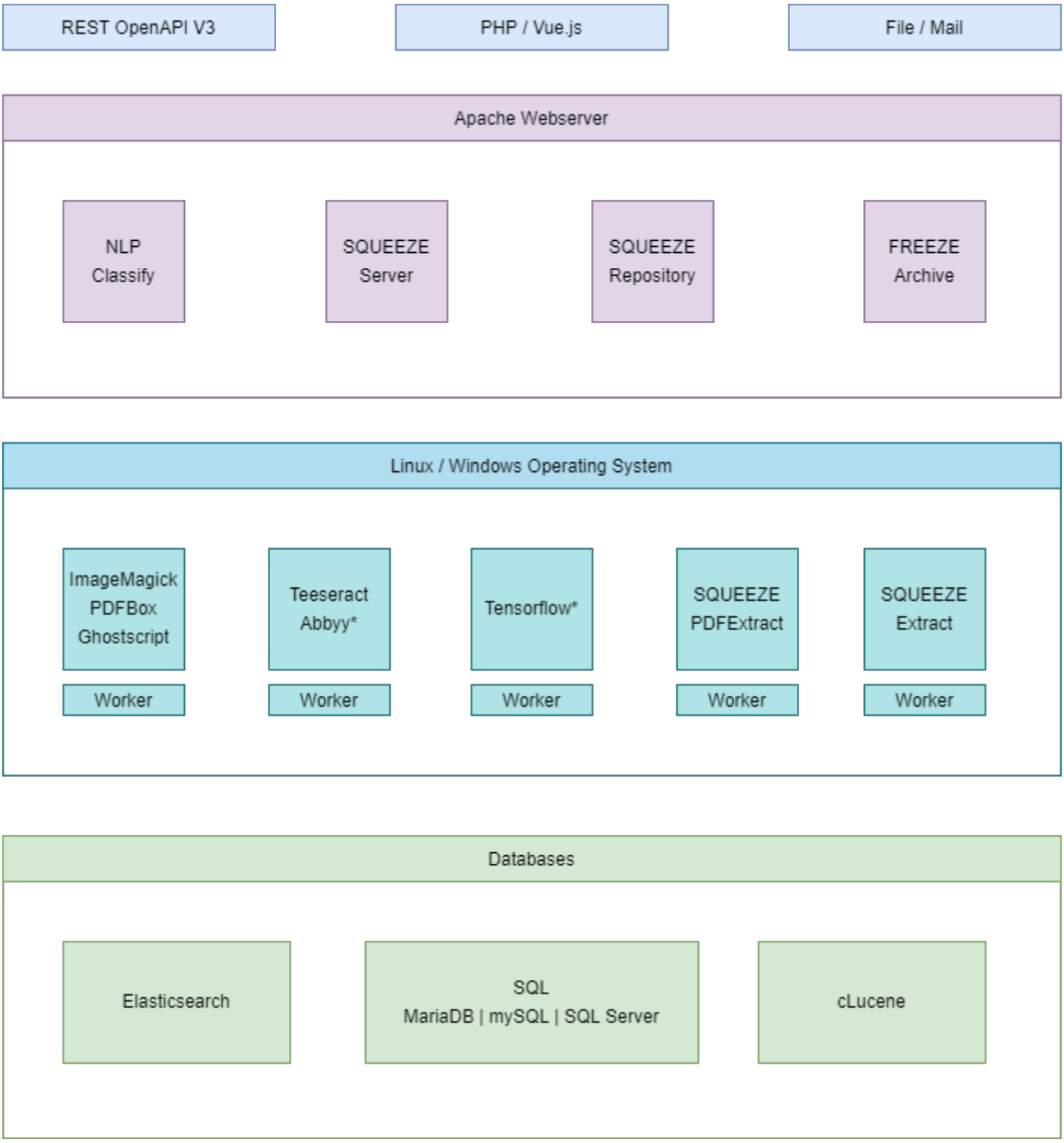
# SQUEEZE Volltext

SQUEEZE arbeitet mit Elasticsearch. Dies ist eine Volltextsuchengine und stellt phonetische, linguistische und unscharfe Suchmethoden zur Verfügung. Zukünftig dient es zusätzlich als Datenlieferant für Unternehmensstammdaten in erweiterten Konfigurationen (Mailroom Szenarien).

# SQUEEZE Systemabbildung




# SQUEEZE Stack


REST OpenAPI V3
PHP / Vue.js
File / Mail
Apache Webserver
NLP
Classify
SQUEEZE
Server
SQUEEZE
Repository
FREEZE
Archive
Linux / Windows Operating System
ImageMagick
PDFBox
Ghostscript
Worker
Teeseract
Abbyy*
Worker
Tensorflow*
Worker
SQUEEZE
PDFExtract
Worker
SQUEEZE
Extract
Worker
Databases
Elasticsearch
SQL
MariaDB | mySQL | SQL Server
cLucene

# Die Verarbeitung mit SQUEEZE

Beschreibung der Verarbeitungsschritte in SQUEEZE

# Dokumenteneingang

Für den Dokumenteneingang können neue Dokumente über verschiedene Wege in das SQUEEZE System gelangen. Im Standard werden die folgenden vier Wege unterstützt:

- Import durch eine Verzeichnisüberwachung
- Import aus Emails
- Import über die SQUEEZE eigene API
- Import über die Documents SOAP Server Emulation

## Verzeichnisüberwachung

SQUEEZE bietet die Möglichkeit 1-n Verzeichnisse zu überwachen, um neue Dokumente aus diesen Verzeichnissen zu importieren. Aktuell ist der Verzeichnisimport nicht an der Oberfläche zu konfigurieren. Hierzu wird ein Skript als geplanter Task eingerichtet, in dem die zu überwachenden Verzeichnisse hinterlegt werden müssen. Bei Neuinstallationen und auch bei der Einrichtung eines neuen Mandanten wird ein Beispielskript bereitgestellt, welches die Konfiguration im Skript vereinfachen soll, bis eine Oberfläche für die Konfiguration erstellt wurde.
Das Beispielskript befindet sich in folgendem Verzeichnis des SQUEEZE Systems:

`...\htdocs\repository\client.server.net\Jobs\PollDirectory.php`

In dieser Datei befindet sich ein Abschnitt in dem die Verzeichnisse und einige weitere Parameter angegeben werden müssen.
Hier ein Auszug aus der Datei mit den entsprechenden Konfigurationen:

```
$directory = array();
$directory['batchClass'] = '1';
$directory['documentClass'] = '1';
$directory['importPath'] = 'D:\\import\\pdf';
$directory['extension'] = 'pdf';
$directory['client'] = 'client.server.net';
$directory['port'] = '80';
$directories[] = $directory;

$directory = array();
$directory['batchClass'] = '1';
$directory['documentClass'] = '1';
```

```
$directory['importPath'] = 'D:\\import\\tif';
$directory['extension'] = 'tif';
$directory['client'] = 'client.server.net';
$directory['port'] = '80';
$directories[] = $directory;
```

In dieser Konfiguration sind zwei zu überwachende Verzeichnisse angegeben.
Folgend eine kurze Beschreibung der anzugebenden Parameter:

| Eigenschaft | Bedeutung |
|---|---|
| batchClass | ID der zu verwendenden Stapelklasse |
| documentClass | ID der zu verwendenden Dokumentenklasse.<br>Der Wert '0' führt zu einer Klassifikation des Dokumentes. |
| importPath | Pfad des zu überwachenden Verzeichnisses. Ein Backslash muss durch einen weiteren Backslash maskiert werden. |
| extension | Dateiendung die importiert werden soll. Derzeit werden nur die Dateiformate PDF und TIF/TIFF unterstützt |
| client | SQUEEZE Mandant für den die Dokumente importiert werden sollen |
| port | Port des SQUEEZE Servers |

Über den geplanten Task des Betriebssystems kann konfiguriert werden, in welchem Intervall das Verzeichnis überwacht/ geprüft werden soll.

> Bei erfolgreichem Import der Datei in SQUEEZE wird die Datei aus dem Verzeichnis gelöscht und nicht verschoben.

# Import von Emails

Für den Import von Emails stellt SQUEEZE zwei Mögliche Schnittstelle zur Verfügung:

- **EWS** Exchange Web Services
- **IMAP** Internet Message Access Protocol

Beide Schnittstellen/Protokolle bieten den identischen Funktionsumfang in Squeeze.

>

> Bei Microsoft Exchange Servern, die vom Kunden selbst betrieben werden, ist die IMAP Schnittstelle häufig deaktiviert, daher ist die EWS Schnittstelle zu präferieren.
> Bei Outlook365 (Exchange in der Cloud) sind beide Möglichkeiten aktiv.

Für die Konfiguration des Postfachs kann der SQUEEZE Webclient genutzt werden.
Die Emailpostfächer werden je Stapelklasse definiert und haben daher ein eigenes Register innerhalb einer Stapelklasse.

Details zur Einrichtung von Emailkonten finden Sie im SQUEEZE Administrationshandbuch.

# Import per SQUEEZE API

Einer der großen Vorteile der SQUEEZE API ist die direkte Rückmeldung beim Import. Im Falle eines Fehlers wird dieser Fehler direkt zurückgemeldet. Bei einer erfolgreichen Übergabe wird auch dies an den Client zurückgemeldet.
Die API Funktion für den Import neuer Dokumente kann über folgende URL angesprochen werden:

http://client.server.net/api/processDocument

Details zu diesem Aufruf und auch allen anderen API Funktionen können Sie in der Swagger Dokumentation einsehen.

# Import per Documents SOAP API Emulation

SQUEEZE ist in der Lage einige der Otris Document SOAP Funktionen zu emulieren. Unter anderem kann die `createFile` Funktion genutzt werden um Dokumente per SOAP an SQUEEZE übergeben zu können. Auch diese Schnittstelle ist bidirektional, d.h. auch bei der Nutzung dieser Schnittstelle wird der Client direkt über den Status des Imports informiert.

# Bildaufbereitung

Um die eingegangenen Dokumente verarbeiten und anzeigen zu können ist der Prozess der Bildaufbereitung notwendig.
Folgende Schritte sind dazu notwendig:

- **Konvertierung von TIF zu PDF**
  Da in der Regel PDFs archiviert werden sollen, ist eine Konvertierung der TIF Dokumente in das PDF Format erforderlich

- **Konvertierung von PDFs zu JPEG**
  Die Konvertierung der PDFs in JPEGs ist notwendig, da im SQUEEZE Viewer das Markieren der Fundstellen sonst nicht möglich ist

- **Trennen aller Seiten ein eine Datei je Seite**
  Das Trennen der Seiten in einzelne Dateien ist ebenfalls erforderlich um die Seiten im Viewer anzeigen zu können und eine schnelle Ladezeit zu gewährleisten. Ohne diese Trennung wären die Ladezeiten von Dokumenten mit mehreren hundert Seiten zu lang. Der Anwender müsste warten bis alle Seiten vollständig geladen wurden.

- **Barcodeerkennung**
  Bei Bedarf kann SQUEEZE nach aufgeklebten Barcodes oder Trennseiten suchen, um Dokumente nach Barcodes zu trennen.
  Hinweis: Im Standard wird von aufgeklebten Barcodes ausgegangen, daher werden die Seiten mit einem Barcodetrenner nicht gelöscht. Sollten also Trennseiten genutzt werden, bleiben die Trennseiten erhalten. Dieses Verhalten kann derzeit nur durch einen UserExit angepasst werden.

Für die Ausführung dieser Schritte muss nichts konfiguriert oder lizenziert werden. Alle diese Schritte werden automatisch ausgeführt.

# Texterkennung

Die SQUEEZE Texterkennung basiert auf der OCR Engine Tesseract. Tesseract wurde ausgewählt, da es unheimlich viele Konfigurationsmöglichkeiten bietet. SQUEEZE wurde so auf Tesseract abgestimmt, dass die bestmöglichen Ergebnisse für die meisten maschinell erstellten Dokumente erzielt werden.

## Texterkennung bei sehr großen Dokumenten

SQUEEZE ist in der Lage auch mit mehreren hundert Seiten umzugehen. In diversen Tests wurden Dokumente mit mehr als 700 Seiten verarbeitet und analysiert. Dabei sind jedoch lange Laufzeiten zu berücksichtigen. Bei einer konservativen Berechnung gehen wir von 15 Sekunden je Seite aus. Nimmt man also ein Dokument mit 700 Seiten, so ist mit einer Verarbeitungszeit von 700 Seiten x 15 Sekunden zu rechnen. Das entspricht einer Verarbeitungszeit von knapp 3 Stunden.

## Alternative OCR Engines

SQUEEZE unterstützt zusätzlich zu Tesseract weitere Lizenzpflichtige OCR Engines. Dazu gehören die **Abbyy** und **IRIS** OCR Engines. Sollte der Bedarf bestehen eine dieser alternativen Engines nutzen zu wollen, so ist dies natürlich möglich führt jedoch zu höheren Kosten, da die Engines kostenpflichtig sind.

## Texterkennung bei PDFs mit Volltext (native PDFs)

Da die Bereitstellung von Rechnungen per E-Mail zum Standard geworden ist, wird die Texterkennung mittels OCR immer seltener erforderlich. Die per E-Mail bereitgestellten Dokumente beinhalten häufig "native" PDFs, die den Volltext bereits enthalten. Dieser Text kann von SQUEEZE ebenfalls verarbeitet werden. In diesem Fall wird keine Texterkennung mittels OCR ausgeführt, der Text des PDFs wird stattdessen analysiert und ausgewertet. Dabei bleiben die geometrischen Informationen der Texte erhalten und können so auch im Viewer markiert und selektiert werden.

Es gibt eine Besonderheit die zu Berücksichtigen ist, bei gescannten Dokumenten, welche bereits einen Volltext enthalten, welche von einer anderen OCR Engine erstellt wurden. Bei diversen Tests wurde festgestellt, dass die genutzten OCR Engines nicht immer ein zufriedenstellendes Ergebnis liefern. Aus diesem Grund wird bei gescannten Dokumenten immer der Volltext verworfen und durch SQUEEZE und Tesseract neu erstellt. Nur so kann eine einheitliche Qualität der OCR gewährleistet werden.

# Klassifikation

Sofern nötig und lizenziert, ist SQUEEZE in der Lage Dokumente automatisiert zu klassifizieren. Eine manuelle Sortierung der Belege nach Dokumentenarten ist dadurch nicht mehr erforderlich.

In Projekten die nur der Rechnungsverarbeitung gelten, ist die Klassifizierung nicht erforderlich. Sollten jedoch verschiedene Dokumente/Dokumentenklassen verarbeitet werden, so kann diese Klassifizierung genutzt werden.

SQUEEZE analysiert hierzu den Volltext des Dokumentes vergleicht diesen mit bereits trainierten Dokumenten und ermittelt so die Dokumentenklasse. Die ermittelte Dokumentenklasse wird dann im Folgeschritt für die Erkennung genutzt. SQUEEZE extrahiert bei einer ermittelten Dokumentenklasse lediglich die Felder, die auch für diese Klasse relevant sind.

# Erkennung/Extraktion

Für die Erkennung/Extraktion von Dokumenten, ist es erforderlich, dass eine Dokumentenklasse für jedes Dokument zuvor bestimmt wurde.

Dies kann über die Importparameter erfolgen oder über die Klassifikation von SQUEEZE.
Die Bestimmung der Dokumentenklasse ist daher so wichtig, da die Dokumentenklasse den Feldkatalog bestimmt. Dieser wiederum legt fest, welche Werte auf dem Dokument mit welchem Verfahren (Lokatoren) gesucht werden sollen. Die ermittelten Werte werden entsprechend der Felddefinition formatiert.
Über den Feldkatalog wird außerdem bestimmt welche der Felder Pflichtfelder sind und ob der Wert eines Feldes durch einen Anwender bestätigt werden muss.

Da sich die Klassifikation des Systems auch mal irren kann, ist es natürlich während der Validierung der Dokumente möglich, die Dokumentenklasse anzupassen. Eine Änderung der Dokumentenklasse führt zur erneuten Erkennung des Dokuments mit dem Feldkatalog der neuen Dokumentenklasse.

# Validierung

Für die Validierung/Kontrolle der extrahierten Werte bietet SQUEEZE einen Webclient an. Dieser Webclient soll die Endanwender so gut wie möglich bei Ihrer Arbeit unterstützen.

SQUEEZE bietet folgende Funktionen zur Unterstützung des Anwenders:

- Markieren der Fundstellen auf dem Dokument im SQUEEZE eigenen Viewer
- Anzeige von möglichen Alternativen
- Formatierung der Werte entsprechend der Felddefinitionen
- Taschenrechner-Funktion auf Betragsfeldern
- Autovervollständigung bei manueller Eingabe von Werten
- Eingabehilfe bei Feldern mit Stammdatenbezug
- Übernahme von Texten des Dokuments in die Felder
- Trainingsfunktion zur Verbesserung des Leseergebnisses

# Export

Zum Abschluss der Verarbeitung nach der Validierung der Belege werden die Belege entsprechend der Konfiguration der Dokumentenklasse an die 1-n Exportschnittstellen übergeben. SQUEEZE bietet bereits im Standard einige Schnittstellen zu Folgesystemen an. Zu diesen Schnittstellen gehören:

- Otris Documents XML Export
- Otris Documents SOAP Export
- Otris EAS Archiv Export
- Navision SOAP Export
- Dynamics365 XML Export
- Dynamics365 OData Export
- EASY XML Server Export
- MaxPost XML Export
- SAP WMD xFlowInterface
- SAP REMADV IDOC Export
- d.velop JPL Export
- Diamant ER2 Export
- SLT.inplast
- Workday SOAP Connector
- Arriba Oracle Export

Die Liste der Schnittstellen wächst mit jedem neuen Projekt und wird regelmäßig erweitert.

SQUEEZE bietet die Möglichkeit mehrere Schnittstellen je Exportvorgang zu definieren, d. h. es ist z. B. möglich ein Dokument erst an den EASY XML Server zu übergeben, um im Anschluss daran die Ausgabe an Dynamics365 zu starten. Dabei ist zu berücksichtigen, dass nur im Falle eines erfolgreichen Exports die nächste Schnittstelle angesprochen wird. Scheitert der Export der ersten Schnittstelle wird der Exportvorgang abgebrochen und folgende Schnittstellen werden nicht gestartet. Eventuelle Fehlermeldungen werden direkt an den Anwender gemeldet, so dass dieser informiert ist, sollte ein Exportvorgang scheitern.

Sobald alle Exportschnittstellen erfolgreich abgearbeitet wurden, wir das Dokument aus der Warteschlange der zu validierenden Dokumente entfernt und für einen bestimmten Zeitraum (aktuell 14 Tage) im SQUEEZE System vorgehalten.

# Das SQUEEZE Invoice Template

# Allgemein

SQUEEZE wird mit einem vorkonfigurierten Invoice Template ausgeliefert. In den vergangenen Projekten wurde es immer weiterentwickelt und verfeinert. Mittlerweile ist der Standard soweit ausgearbeitet, dass nach der Installation des SQUEEZE Systems nur noch die Stammdaten zu füllen sind. Sobald diese im Invoice Template hinterlegt sind, steht der Verarbeitung der ersten Belege nichts entgegen. Installationen innerhalb einer weniger Stunden sollten daher nicht die Ausnahme sein.

Auf den folgenden Seiten stehen einige Informationen zum Invoice Template bereit.

# Stammdaten

Für die Erkennung der Rechnungsinfomationen werden einige kunden individuelle Stammdaten benötigt. Diese Stammdaten müssen zum Teil aus anderen führenden Systemen importiert werden. Hierfür stehen verschiedene Möglichkeiten zur Verfügung. Zu diesen Möglichkeiten gehören:

- Import von CSV Dateien
- Import von XLS Dateien
- Import von XLSX Dateien
- Übertragung der Stammdaten mit Hilfe der SQUEEZE API

Der Import der Dateien kann mit Hilfe eines geplanten Tasks automatisiert werden. Eine Downtime des Systems während der Stammdatenaufbereitung ist nicht erforderlich, so dass die Stammdatenübertragung auch während des Tages ausgeführt werden kann.

Mit dem SQUEEZE Invoice Template werden folgende Stammdatentabellen ausgeliefert:

| | **Name** | **Beschreibung** |
|---|---|---|
| 1 | companies | Liste aller Mandanten/Buchungskreise |
| 2 | creditors | Liste aller Lieferanten |
| 3 | eucountries | Liste aller EU Länder |
| 4 | currencies | Liste der zu suchenden Währungen |
| 5 | taxrates | Liste aller Steuersätze je Land |
| 6 | companysearch | Liste alle Ausdrücke zur Ermittlung der Mandanten/Buchungskreise |

Jede dieser Tabellen hat eine bestimmte Funktion während des Erkennungsprozesses oder aber während der Validierung. Die Funktion der Tabellen wird in den folgenden Punkten beschrieben.

## Companies

Die Liste der Mandanten/Buchungskreise dient als Eingabehilfe während der Validierung. Sie soll es dem Anwender so einfach wie möglich machen einen Mandanten/Buchungskreis auszuwählen, sofern dieser nicht während der Extraktion erkannt wurde.

# Creditors

Die Liste der Lieferanten wird sowohl während der Extraktion als auch während der Validierung genutzt. Während der Extraktion dient die Tabelle als Abgleich zu gefundenen Werten um z.B. den Lieferanten auf Basis der Umsatzsteueridentifikationsnummer oder der IBAN zu erkennen. Desweiteren werden folgende Werte genutzt um den Lieferanten zu erkennen.

- EMail-Adressen
- Webseiten
- Telefonnummern
- Faxnummern

Grundsätzlich sind natürlich auch andere Werte möglich die individuell erweitert werden können.

# EUcountries

Die Liste der EU Länder kann genutzt werden um Steuerkennzeichen ableiten zu können. Je nach ERP System kann diese Aufgabe auch vom ERP System übernommen werden.

# Currencies

Die Liste der Währungen wird genutzt um die Belegwährung ermitteln und ggf. umschlüsseln zu können. Auf Basis dieser Liste kann z.B. nach dem Währungssymbol gesucht werden, welches dann aber in den ISO Wert der Währung übersetzt wird.

# Taxrates

Die Liste der Steuersätze je Land wird für die Betragserkennung genutzt. Die Endbeträge werden i.d.R. rechnerisch ermittelt. Um dies tun zu können, muss der Steuersatz des jeweiligen Landes bekannt sein. Erst die Kombination aus Netto, Steuer und Brutto führt zu einer eindeutigen Erkennung der Endbeträge.

# Companysearch

Die Liste alle Ausdrücke zur Ermittlung der Mandanten/Buchungskreise kann wichtig werden, wenn der Rechnungsempfänger ermittelt und umgeschlüsselt werden muss. Mit Hilfe dieser Liste kann nach verschiedenen Ausdrücken auf dem Beleg gesucht werden, um so den Rechnungsempfänger zu ermitteln, dabei muss nicht immer der Empfängername genutzt werden. Es wäre ebenso möglich nach der UStId des Rechnungsempfängers zu suchen und daraus den Empfänger abzuleiten.