SYSTEMATISCHE ERKENNUNG STRUKTURIERTER DATEN IN PDF-DATEIEN SOWIE ABSTRAKTION UND WEITERVERARBEITUNG ZUR DIGITALISIERUNG IM AUSSCHREIBUNGSPROZESS

BENEDIKT SCHMALER
(BACHELORSTUDIUM WIRTSCHAFTSINFORMATIK)

Betreuer: Prof. Dr. Wolfgang Mühlbauer, Prof. Dr. Florian Künzner

Die Ausschreibungsplattform ePlato (www.eplato.de) bildet einen Großteil des Ausschreibungsprozesses in der Baubranche ab. Der elektronische Austausch von Informationen innerhalb dieses Prozesses ist durch bestimmte Normen und Dateiformate geregelt, die vom Gemeinsamen Ausschuss Elektronik im Bauwesen erarbeitet und zur Verfügung gestellt werden. Diese Formate können jedoch nicht von allen Beteiligten im Prozess verarbeitet werden und so kommt es häufig vor, dass der Informationsaustausch über Portable Document Format (PDF)- Dateien abgewickelt wird. Da Formate wie PDF sowohl für die Planung als auch für den weiteren Prozessverlauf ungeeignet sind, resultiert hieraus ein hoher manueller und monetärer Aufwand bei den nachgelagerten Parteien im Prozess.

Das maschinelle Verarbeiten dieser PDF-Dateien und deren Integrationen in den digitalen Ausschreibungsprozess von ePlato sind Teil der vorliegenden Arbeit. Dabei wurden drei Technologien PDF-To-Text, Optical Character Recognition (OCR) und Microsoft Azure Formularerkennung prototypisch implementiert und deren Ergebnisse miteinander verglichen. Die Analyse von PDF-To-Text zeigte, dass diese Technologie zwar für das reine Auslesen von Texten sehr gut funktioniert, jedoch bei der Verarbeitung von gescannten Dokumenten keinerlei Daten mehr liefern kann. Dies ist zurückzuführen auf die Struktur von PDF-Dokumenten, welche bei gescannten Dokumenten rein aus Bildern besteht und diese Technologie rein auf den textlichen Inhalt einer PDF-Datei Zugriff hat. Aufgrund der Anforderung, dass auch gescannte Dokumente verarbeitet werden sollen, wurde die weitere Analyse der Technologie an dieser Stelle abgebrochen. Im weiteren Verlauf zeigte der Vergleich zwischen OCR und dem Azure Formularerkennungsdienst, dass mit beiden Technologien sehr gute Ergebnisse erzielt werden können. Beide Technologien wurden dann anhand der Kriterien Laufzeit und Editierdistanz miteinander verglichen. Hierbei zeigte sich, dass bei der OCR-Extraktion mit einer Auflösung von 300 DPI (Dots per Inch, dt. Punkte pro Zoll, ist eine Maßeinheit für Bildauflösung) die meisten Fehler auftraten. Diese Fehler konnten durch die Verdoppelung der Auflösung von 300 auf 600 DPI eliminiert werden. Jedoch führte diese Verdoppelung auch zu einer Verdoppelung der Laufzeit. In Bezug auf die Editierdistanz wurde ein sehr viel besseres Ergebnis im Vergleich zu den 300 DPI erzielt, es war aber immer noch schlechter als das bei der Azure Formularerkennung.

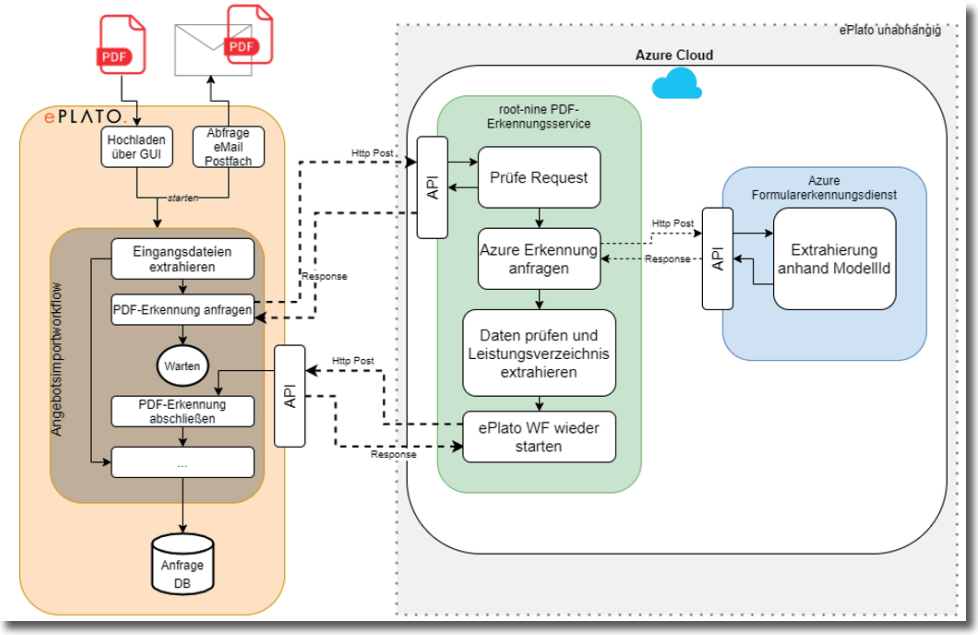

Abbildung 1: ePlato Microservicearchitektur. Vereinfachte eigene Darstellung.

Der Formularerkennungsdienst von Azure bietet zusätzlich zur performanteren Verarbeitung der Dateien weitere Funktionalitäten wie das automatische Erkennen von Metadaten durch gezieltes Anlernen am System. Als Konsequenz dieser Analyse wird der Azure Formularerkennungsdienst als finale Technologie ausgewählt und für die Integration in den digitalen Ausschreibungsprozess von ePlato vorbereitet sowie implementiert.

BESTE ABSCHLUSS-

ROSENHEIMER INFORMATIKPREIS WIF-BACHELOR

Durch eine zukunftssichere Architektur wurde der Formularerkennungsdienst an die Ausschreibungsplattform angebunden. Technisch ist ePlato eine cloudbasierte Anwendung mit einer Vielzahl an Microservices auf Basis der Azure Service Fabric in der Azure Cloud. Um sich nicht vollständig von der Technologie der Azure Formularerkennung abhängig zu machen, wurde in diesem Schritt ein neuer „root-nine PDF-Erkennungsservice“ als Kommunikationsschnittstelle zwischen ePlato und dem Dienst der Azure Formularerkennung entwickelt und integriert.

ePlato bietet bereits unterschiedliche Möglichkeiten, Ausschreibungsdokumente in die Plattform zu importieren und die Informationen daraus für den weiteren Verarbeitungsprozess innerhalb von ePlato zur Verfügung zu stellen. Das Einspielen der PDF-Dokumente wurde als weitere Möglichkeit zu den bereits bestehenden implementiert. Nach dem Einspielen eines PDF-Dokuments, wird dieses über eine REST-Schnittstelle an den neu entwickelten „root-nine PDF-Erkennungsservice“ gesendet. Dieser wiederum sendet die PDF-Dateien an den Azure Formularerkennungsdienst. Als Antwort liefert dieser Service dann alle von ihm erkannten Wörter mit Metadaten zurück. Diese Antwort wird anschließend mit einem für genau diesen Anwendungsfall entwickelten Algorithmus verarbeitet, um relevante Informationen des Leistungsverzeichnisses sowie des Angebots in einer strukturierten Form zur Verfügung zu stellen.

Das Ergebnis dieser Verarbeitung wird anschließend wieder per REST-Schnittstelle an ePlato zurückgesendet, um dort das Angebot in die Applikation zu importieren und den Bearbeitern zur Verfügung zu stellen. Die Architektur und deren Verbindungen sind vereinfacht dargestellt in Abbildung 1.

Abschließend wurde auf diesem neu implementierten Prozess ein größerer Testlauf mit 36 PDF-Dokumenten durchgeführt. Dabei wurden die importierten Leistungsverzeichnistabellen sowie deren Leistungspositionen ausgewertet und mit den Inhalten der PDF-Dateien verglichen. Hier konnte festgestellt werden, dass lediglich bei vier Dateien Differenzen zwischen erwarteter und tatsächlich erhaltener Anzahl an Leistungspositionen aufgetreten sind.

Dies zeigt, dass über die analysierten Technologien und die abschließende Integration in die Systemlandschaft der Ausschreibungsplattform ePlato auch PDF-Dateien wieder in den digitalen Prozess der Ausschreibungen integriert werden können. Durch die gewählte Architektur kann der Dienst kontinuierlich und unabhängig von Bereitstellungszyklen anderer am Prozess beteiligter Systeme angepasst und verbessert werden.

Abkürzungen

OCR	Optical Character Recognition
REST	Representational State Transfer
DPI	Dots per Inch