**Analyse der PDF-Dokumentengruppe**

Die Analyse der PDF-Dokumente übernimmt das Java-Programm, welches bereits Thema der Belegarbeit vom 10. Februar 2020 war. Das Programm überprüft mehrere gleichartige PDF-Dateien. Gleichartig bedeutet, dass die Inhalte der Dokumente immer gleich strukturiert sind und sich zusammengehörige Informationen immer an gleicher Stelle im Dokument befinden. Solche Dateien werden häufig automatisiert generiert, beispielsweise als standardisierte Formulare. Die zusammengehörigen Informationen an gleicher Stelle werden Attribute genannt, falls sie eine bestimmte Relevanz für das einzelne Dokument besitzen. Um solche Attribute zu finden, vergleicht das Java-Programm alle Zeichenketten mit nahezu übereinstimmenden Koordinaten im PDF-Dokument. Unterscheiden sich diese Zeichenketten, werden sie vorläufig Attributen zugewiesen. So werden z.B. Rechnungsnummern in einem bestimmten Formularfeld gefunden, die sich in verschiedenen Rechnungsdokumenten unterscheiden. Der Nutzer kann diese Attribute übernehmen oder entfernen, ihnen Namen zuordnen und die Attributwerte mithilfe von regulären Ausdrücken weiter filtern. In einer Übersichtsseite werden dann PDF-Dokumente ihren Attributwerten zugeordnet und sortiert, sodass z.B. schnell auf das Dokument mit einer bestimmten Rechnungsnummer zugegriffen werden kann.

**Bearbeiten von Dokumenten**

Da bei der Recherche nach Software-Modulen zur browserbasierten Bearbeitung von PDF-Dateien keine brauchbaren Lösungen mit freier Lizenz gefunden wurden und die Entwicklung einer eigenen Lösung zu aufwendig und komplex wäre, wird eine lokale Bearbeitung von PDFs vorgesehen. Um eine PDF-Datei zu bearbeiten, kann diese in der Projektverwaltung ausgewählt werden. Der Server gibt die Datei dann zum Download frei und erzeugt einen Link, mit welchem die PDF automatisch mit Microsoft Edge geöffnet wird. Microsoft Edge wurde als Standardprogramm für die PDF-Bearbeitung ausgewählt, weil es eine sehr gute Unterstützung für Stifteingaben bietet und auf den meisten Microsoft-Geräten vorhanden sein sollte. Die PDF kann dann im Standard-Downloadverzeichnis des Clients gespeichert werden. Nach der Bearbeitung kann sie dort einfach gespeichert werden und wird vom Synchronisationsprogramm auf den Server hochgeladen, wie unter [Synchronisationsprogramm] beschrieben.

**Synchronisationsprogramm**

Da bei der Recherche nach Software-Modulen zur browserbasierten Bearbeitung von PDF-Dateien keine brauchbaren Lösungen mit freier Lizenz gefunden wurden und die Entwicklung einer eigenen Lösung zu aufwendig und komplex wäre, wird eine lokale Bearbeitung von PDFs vorgesehen. Aus diesem Grund müssen die Dateien nach der Bearbeitung wieder auf den Server hochgeladen werden. Um diese Aufgabe zu automatisieren, wurde ein Programm mit Java erstellt, welches während der Bearbeitung **im Hintergrund läuft**. Es muss vor der Bearbeitung manuell gestartet werden. In diesem Programm können Nutzer und Downloadverzeichnis für die PDF-Dokumente eingestellt werden. Wurde eine Datei bearbeitet, muss der Nutzer den entsprechenden Eintrag im Webservice auf erledigt setzen. Der Webservice trägt das dann in die Datenbank ein. Das Hintergrundprogramm überprüft die Datenbank auf Einträge, in der dieses Feld gesetzt ist. Die Datenbankverbindung wird über den Java-MySQL-Datenbanktreiber in der Version ... realisiert. Ist der Datenbankeintrag gesetzt, überprüft das Hintergrundprogramm, ob die Datei sich im zuvor eingestellten Download-Verzeichnis befindet. Ist das der Fall, wird die veränderte Datei hochgeladen. Der Upload erfolgt über einen HTTP-Stream. Das ermöglicht es dem Nutzer, die Datei nach dem Download zu bearbeiten und einfach abzuspeichern. Das Hintergrundprogramm übernimmt dann die Synchronisation mit dem Server.