Backlog RM - Fonctionnalité #18327

OCR sur PDF Image

20/09/2021 14:53 - Cyril VAZQUEZ

Statut:	Rejeté	**Début:**	20/09/2021
Priorité:	3-Mineur	**Echéance:**	
Assigné à:	Cyril VAZQUEZ		
Catégorie:			
Version cible:	Inscription Backlog		
Tags RM:			

Description

En tant que SA, je veux extraire le texte des documents au format PDF issus de la numérisation afin de les indexer

Détail

La fonction actuellement implémentée permet de brancher des outils d'extraction en fonction du format de ressource, identifiée par la PUID.

L'extraction pour les PDF utilise Apache Tika et fonctionne bien pour les PDF issus de documents bureautique ou formats balisés.

Problème:
Pour les PDF issus de la numérisation sans étape d'OCR: le PDF image seule ne contient pas de texte.
Il n'y a aucun moyen de différencier a priori les PDF+texte et les PDF image seule.

Proposition:
- extraire le texte avec TIKA
- si le texte est vide, extraire par OCR

Processus:
- extraire les images du PDF avec un premier outil (PDFLib)
- passer l'OCR sur les images
- générer un nouveau PDF+texte

Attention à la version du PDF qui nécessitera sans doute des logiciels payants.
Attention à la génération d'un nouveau PDF à partir d'images dont on n'a pas la position initiale !

Historique

#1 - 20/09/2021 16:24 - Emmanuel DILLARD

- Statut changé de A qualifier à R&D - A étudier

#2 - 20/09/2021 16:25 - Emmanuel DILLARD

- Version cible mis à 281

#3 - 04/10/2021 09:02 - Cyril VAZQUEZ

- Priorité changé de 1-Majeur à 3-Mineur

#4 - 04/10/2021 09:03 - Cyril VAZQUEZ

- Statut changé de R&D - A étudier à Rejeté

#5 - 20/04/2022 16:28 - Cyril VAZQUEZ

- Version cible changé de 281 à Inscription Backlog