

Dexelia **RAD-LAD**



**Extraction automatisée
des informations**



Dexelia RAD-LAD

Dexelia RAD-LAD permet de automatiser la reconnaissance et la lecture des documents structurés et non structurés.

Il repose sur le module Dexelia TDI afin d'optimiser la qualité des documents.

Ce module permet de gérer les différents modes d'import et export des documents

Dexelia RAD-LAD permet de définir comment ces informations vont être extraites. Les stratégies d'évaluation utilisables sont plus ou moins complexes, allant du simple OCR sur une zone fixe à la recherche par mot clé utilisant une ou plusieurs expressions régulières.

Le module de Rad/Lad est polyvalent et évolutif :

Atelier de dématérialisation

La Rad/Lad est exécutée par Dexelia TDI Batch avant indexation. En cas de succès, il pourra être décidé de publier directement les résultats, sans passer par le vidéocodage.

Application décentralisée

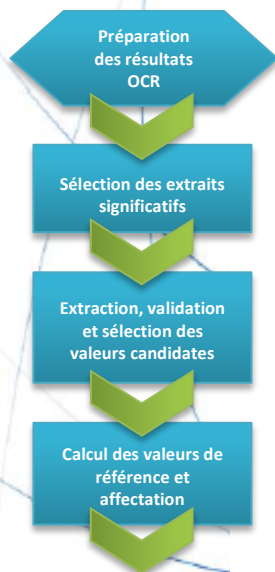
La RAD-LAD est exécutée au moment de la numérisation et l'opérateur en contrôle immédiatement les résultats.



Le fonctionnement

Les index peuvent être recherchés indépendamment du modèle de document (index générique) ou au contraire spécifiquement sur les pages satisfaisant certains critères (présence de mot clé à certains endroits, dimension, code à barre, logos, ...)

Plusieurs méthodes de recherche peuvent être définies pour un même index, la première méthode appliquée avec succès déterminera la valeur de l'index.



Les résultats OCR bruts sont traités pour fiabiliser la recherche et l'extraction (remplacement de caractères, suppression des accents ou des sauts de ligne, mise en forme, ...)

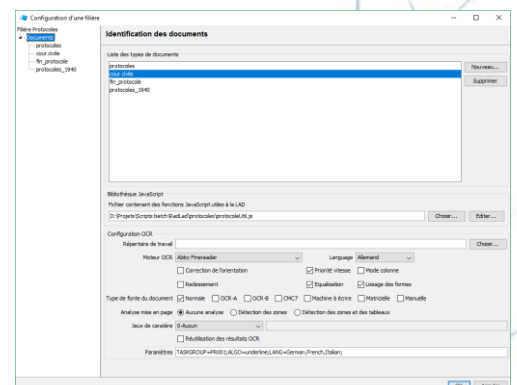
A partir des résultats OCR préparés, identification des extraits de texte significatifs : zones d'intérêts, structure du texte (paragraphes, tableaux, lignes, colonnes) et repérage de mot clés

Dans les extraits identifiés ci-dessus, extraction des valeurs candidates, filtrage des valeurs valides puis sélection de la ou des valeurs à conserver (nième, première, dernière, toutes,...)

A partir de la (des) valeur(s) sélectionnée(s) ci-dessus, calcul de la (des) valeur(s) finale(s) et affectation au champ d'indexation

L'écran principal permet de créer les modèles de documents et de réaliser leurs configurations RAD/LAD. Il permet également de définir les options OCR qui seront utilisées.

Il est possible de désigner les fonctions JavaScript qui seront utilisées dans les configurations pour valider ou formater les données





Le typage du document

Le typage des documents est réalisé à partir :

- D'une métadonnée, issue du PDF ou de la phase de numérisation (ex : CAB, OCR zoné ou traitement d'image pour repérer les logos, une forme particulière, des parties de texte souligné, ...), sur laquelle on peut appliquer une REGEX
- D'une liste de mots clés que le document doit contenir ou non (ex : Facture)
- Du nommage du document
- D'une fonction JavaScript qui va pouvoir parcourir les résultats OCR pour définir le type du document

Les + :

- Une interface simple pour définir toutes les règles de typage
- Des configurations préétablies pour différents types de documents standards (CNI, passeports, impôts, ...).

Dexelia RAD-LAD permet d'appliquer des options OCR différentes selon les types de documents ou selon des zones particulières de l'image (ex : bande MRZ d'une pièce d'identité, ligne CMC7 d'un chèque).



L'extraction des champs

Pour chaque type de document, les règles d'extraction de chaque champ sont définies. Il est possible d'activer une règle si une précédente a échoué.

Avant d'appliquer une règle d'extraction, il est possible de corriger les erreurs habituelles des OCR : par exemple sur un champ numérique de remplacer un S par un 5, I par 1, ..., de supprimer les accents et les tabulations, de ne pas tenir compte de la casse.

La recherche peut être limitée à une zone d'intérêt, par exemple l'entête de la facture, pour améliorer les performances.

L'extraction peut utiliser les résultats avancés de certains OCR comme la caractérisation des tableaux, lignes, colonnes et cellules.

La recherche d'un champ est réalisée par la définition d'expressions régulières qui permettent de déterminer le début et/ou la fin du champ. Si plusieurs candidats sont extraits, des fonctions permettent de retenir le plus pertinent.

Les + :

- Une extraction fiable qui corrige les erreurs les plus courantes des OCR
- Une recherche rapide.
- Des tests de non régression pour évaluer les améliorations de l'extraction lors d'un changement
- Pas de programmation : Un paramétrage simple des règles
- Un lien direct avec les écrans d'indexation



Caractéristiques

OCR	Utilise les principaux OCR du marché : TESSERATC en standard / ABBYY Finereader, Kofax Omnipage en options Intègre les fonctions ICR et OCR de ces produits Modules OMR et codes à barres 1D et 2D en standard Utilisation de toutes les fonctions disponibles de l'OCR, comme la reconnaissance des tableaux
Facilité d'intégration	Intégré à Dexelia TDI Batch : utilisé sur un serveur pour traiter de façon centralisée tous les documents issus d'ateliers de numérisation, de mails, de documents bureautique Intégré à Dexelia Indexation : utilisé dans une application décentralisée qui réalise la numérisation, l'extraction automatique, l'indexation et la publication des documents S'interface nativement aux formulaires d'indexation définis dans Dexelia Indexation. L'intégration dans des applications métier est possible grâce au SDK, auquel il faut ajouter un transfert de compétence pour maîtriser les fonctions et les paramètres disponibles.
Règles de typage et d'extraction	Basée sur la recherche de mots clé et les expressions régulières, les cas d'extraction les plus difficiles peuvent être traités en JavaScript. Une bibliothèque de documents standards déjà établie : CNI, passeport, impôt, carte grise, ...
Contrôle de validité des documents	Des contrôles de falsifications de certains documents sont intégrés (ex : ligne MRZ des CNI, passeports. Contrôles de cohérence au sein d'un même document, si une valeur peut être lue de différentes manières
Fonctions complémentaires	Evaluation des taux de réussite sur un jeu d'images de références. Tests de non régression
Configuration matérielle et logicielle	- Disponible sous Windows, OSX, Android, iOS et Linux - Processeur Intel Core i5 ou équivalent (2.5 GHz minimum) - 4 Go de RAM
Format de sortie	Les données sont exportées aux formats texte ou XML afin d'être intégrées dans une application tierce, GED, application métier, ERP. L'export peut également se faire dans un autre format grâce à un plugin Java. Lorsque l'application est interfacée à une base de données, on peut écrire les informations dans la base de données.

