LAD, RAD et OCR : un point d'entrée pour votre Ged (/blog/82-carolineb/229-lad-rad-ocr-capture-ged)

LAD, RAD, OCR ICR, IWR... autant d'acronymes barbares qui sèment la confusion dans l'esprit de ceux qui cherchent désespéremment à passer d'un traitement de leur document papier à un traitement numérique.

Pour éclaircir vos esprits, voici un article qui vous détaille les principales méthodes et technologies permettant de passer d'un support papier à un support numérique pouvant être traité par les logiciels bureautiques.

Introduction aux notions

Avant de parler des technologies commercialisables qui vous permettront d'arriver à vos fins, faisons un petit tour du côté de la recherche et des systèmes de reconnaissance de caractère, socle des technologies de LAD-RAD et OCR.

Pour les personnes intéressées par ces fondements technologiques suivez le lien vers un document de l'Institut des Ingénieurs électriciens et électroniciens (IEEE) du Canada "Systèmes de reconnaissance de caractères pour les non-experts (http://www.ieee.ca/canrev/canrev33/kharma_fr.pdf)".

Et maintenant un point sur les notions.

LAD : Lecture automatique de document

Le **LAD** est un ensemble des technologies (OCR, RAD, ICR...) utilisées pour lire, indexer et stocker les données contenues dans un support physique (papier, transparent...).

La LAD regroupe trois technologies indispensables à son fonctionnement :

- La RAD : reconnaissance automatique de documents
- L'OCR : reconnaissance optique des caractères avec OCR scanner
- L'IRC : reconnaissance intelligente de caractères

RAD (reconnaissance automatique du document)

Définition
Application qui permet de numériser un document à l'aide d'un scanner et d'un logiciel d'OCR en tenant compte du type de document et de ses contraintes (reconnaissance caractères numérique, alpha, codes barre).
Permet de trier automatiquement les documents à partir d'un modèle prédéfini et de l'utilisation d'OCR ou d'ICR.

Principe technologique
La technologie RAD consiste à reconnaitre le type du document à traiter. Ainsi après avoir analysé la mise en page du document comme l'emplacement d'image, d'encadré, le module RAD la compare à des modèles issus de sa base de données afin de déterminer s'il s'agit d'un devis, d'une facture, une commande ou tout autre document.
Selon la nature du document, l'utilisateur pourra vouloir y extraire différentes informations. Dès lors l'application LAD lance le module OCR (reconnaissance optique des caractères).

OCR (reconnaissance optique de caractères)

Définition
Matériel ou logiciel de conversion d'un document image (codé en mode image) en un document texte (codé en mode caractère), avec ou sans enrichissement typographique et conservation de la structure du document associé, exploitable par des programmes informatiques.
Les documents traités sont de diverses catégories d'écriture (dactylographiée, numérique, manuscrite, bâton et cursive) ; ils peuvent être multiécritures et hétérogènes (imprimés, dégradés, bruités). Plusieurs techniques de reconnaissance sont utilisées par les moteurs de reconnaissance optique de caractères, avec ou sans apprentissage.

Le principe technologique
Le principe d'une technologie OCR est de lire le document pour détecter les formes, puis les comparer à des bibliothèques de formes pour en faire correspondre un caractère. Si des erreurs surviennent lors de la reconnaissance d'un caractère, l'OCR compare alors le mot entier au contenu de son dictionnaire intégré pour en déduire l'équivalence la plus proche et ainsi corriger le caractère mal lu. Ainsi le texte pourra alors être segmenté selon l'information recherché.

Les principaux acteurs de l'OCR en 2009

ABBYY
http://france.abbyy.com/
IRIS
http://www.irislink.com
Nuance
http://www.nuance.fr/

N'oublions pas l'Open Source :

Tesseract ancien outil IBM mis récemment en Open Source par Google, le plus précis des OCR Open Source mais hélas uniquement en anglais
Tesseract sur Sourceforge (http://sourceforge.net/projects/tesseract-ocr/).
C'est d'ailleurs le grand problème des OCR Open Source : ils sont en anglais !

Les paramètres qui doivent influencer un choix de logiciel d'OCR

- Précision de la reconnaissance de caractère
- Précision de la reconstruction de mise en page
- Support de plusieurs langage
- Rapidité
- Interface utilisateur
- Caractéristique spéciales pour des projets de niches