Stage Maroc

Méta données et OCR

Dominique Maillet
Bibliothèque nationale de France

21 au 25 mai 2012

LA CHAINE DE NUMERISATION : LES 11 ETAPES

- 1) La sélection des fonds : QUOI
- 2) Définition les modalités de mise à disposition de ces fonds: POURQUOI
- COMMENT :
- 3) La gestion des droits d'auteur
- 4) Catalogage des documents à numériser
- 5) La préparation des documents à numériser
- 6) La numérisation des documents
- **7) La post-production (création des métadonnées)**
- **8) Le passage en OCR pour les documents imprimés**
- 9) Le contrôle qualité des documents numérisés.
- 10) La sauvegarde des documents numérisés sur les serveurs informatiques
- 11) La mise en consultation des documents numérisés.

Les Meta données

- Ce sont des informations qui sont produites à l'occasion de la prise de vues et qui vont permettre de compléter l'image numériques
- Meta données techniques
- Meta données descriptives
- Meta données informatives

Les meta données techniques

Elles permettent d'identifier et de gérer le fichier numérique.

Qui a fait la numérisation, sur quel appareil, à quelle date, quelle résolution a été utilisée, quel espace couleurs, etc

Les Méta données descriptives

Saisie d’informations descriptives (pagination, légendes) qui seront liées au catalogue et qui vont faciliter la navigation dans le document sur ordinateur

Métadonnées descriptives

- Eléments de pagination/ Foliotation: inclut les éléments de couverture et les types de pages (index, tdm, dessins, etc)

Type page, n° page

Métadonnées descriptives

- Instructions de numérisation

paperolle dépliée, collette levée, verso, marionnette face, profil, maquette plan large, etc

- Légende: texte inscrit sur le document à numériser qui sera saisie pendant l'étape de numérisation et qui permettra de compléter le catalogue.

Métadonnées informatives

- Commentaire à destination de l'usager

Permet d'informer le lecteur d'une particularité du document traité (page manquante, ouvrage maculé, etc)

- Commentaire de production

Permet à l'opérateur d'indiquer une difficulté de numérisation (reliure très serrée: caractères tronqués, feuillet fragile: mise à plat impossible, etc)

La reconnaissance optique de caractères (ROC)

Le passage en OCR (OCR: Optical character Recognition.)

p o u r l e s d o c u m e n t s i m p r i m é s

S'étend si possible à tout document susceptible d'être océrisé.

Le passage en OCR pour les documents imprimés

Procédé informatique qui exploite un fichier numérisé en TIFF compressé noir et blanc ou un PDF pour en faire une page comme si elle avait été écrite avec un logiciel de traitement de texte ou dactylographié. Ce fichier texte est ensuite exploitable par des outils de manipulation de texte

Ce qui permet ensuite de faire une indexation de la page avec tous les mots du texte et donc de trouver plus facilement un document qu'avec sa notice et de faire des recherches plus poussées.

OCR les logiciels

Parmi les meilleurs et les plus performants on citera:

Fine reader d'Abby (150 – 270 €)

OMNIPAGE

Readiris

Tesseract

Acrobat professionnel d'Adobe (285 €)

OCR

Processus

La **numérisation** et la **reconnaissance optique peuvent faire l'objet de 2 étapes consécutives ou séparées.**

1) Numérisation

2) Application du logiciel OCR

Numérisation et OCR

- Format de sortie de la reconnaissance de texte
  - Codage
  - Texte
  - Position des éléments dans la page

OCR

Le fonctionnement des systèmes d'OCR performants est peu connu, car protégé par le secret industriel. Hormis les techniques qui peuvent varier, tous les logiciels analysent et traitent les documents dans cet ordre :

1. Pré-analyse de l'image
   Améliorer la qualité de l'image
2. Analyse de page
   Identifier les zones de texte, les tableaux et les images
3. Reconnaissance des caractères
   Comparer avec des formes connues en fonction de différentes techniques
4. Post-traitement
   Utiliser des méthodes linguistiques et contextuelles pour réduire le nombre d'erreurs
5. Génération du format de sortie

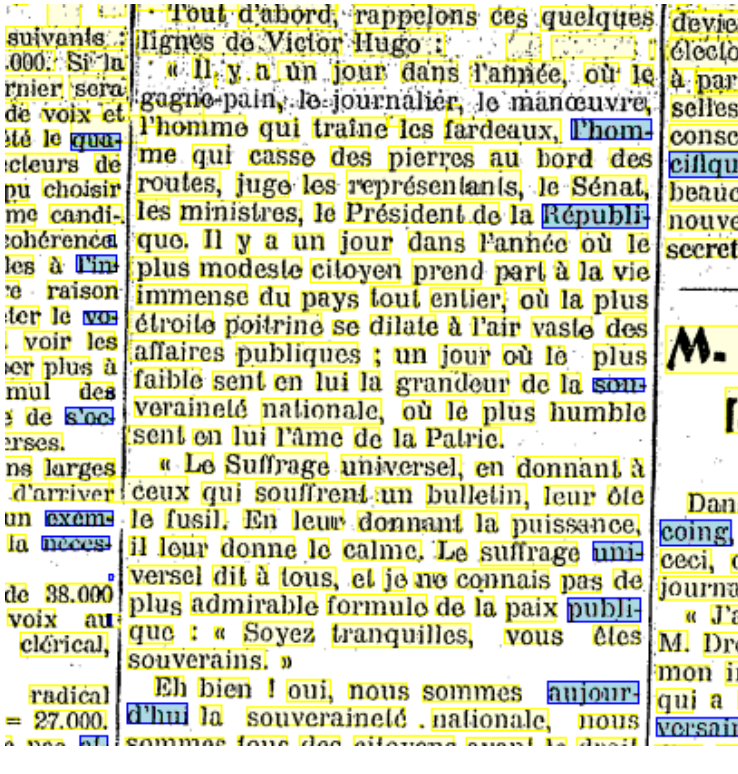
OCR

- Les logiciels effectuent une segmentation informatique de la page afin d'identifier les parties comportant des illustrations, des tableaux et du texte.
- Ensuite chaque zone est elle-même segmentée par ligne mots et caractères pour le texte et par ligne colonne pour les tableaux.
- Les résultats de la reconnaissance peuvent ensuite être corrigés améliorés modifiés avant d'être exportés dans un format texte brut, un fichier bureautique, un PDF ou un format xml qui permet de préserver l'ensemble de l'information reconnue dans un format standard (Alto)

- Les algorithmes de reconnaissance associent la position de l'image par rapport au caractère reconnu, sa police, sa taille, sa casse, le mot auquel il appartient le taux de confiance de la reconnaissance, etc
- Ce qui permet :
- Soit d'afficher l image d'origine, le contenu reconnu étant caché à l'utilisateur mais permettant des recherches
- Soit d'afficher le fichier reconnu et restructuré dans une présentation identique à l'original
- Le fichier PDF est plus un fichier de publication et de diffusion. Le fichier XML est plus complet

Niveau mot

- Attributs
  - ID
  - Hauteur, largeur, position x et y
  - Fiabilité (WC = word confidence) : [0;1]
  - Présence dans le dictonnaire (WD = word dictionnary) {true, false}
  - Style
  - Hyphène


Tout d'abord, rappelons ces quelques lignes de Victor Hugo :

« Il y a un jour dans l'année, où le gagne-pain, le journalier, le manœuvre, l'homme qui traîne les fardeaux, l'homme qui casse des pierres au bord des routes, juge les représentants, le Sénat, les ministres, le Président de la Républiquo. Il y a un jour dans l'année où le plus modeste citoyen prend part à la vie immense du pays tout entier, où la plus étroite poitrine se dilate à l'air vaste des affaires publiques ; un jour où le plus faible sent en lui la grandeur de la souveraineté nationale, où le plus humble sent en lui l'âme de la Patrie.

« Le Suffrage universel, en donnant à ceux qui souffrent un bulletin, leur ôte le fusil. En leur donnant la puissance, il leur donne le calme. Le suffrage universel dit à tous, et je ne connais pas de plus admirable formule de la paix publique : « Soyez tranquilles, vous êtes souverains. »

Eh bien ! oui, nous sommes aujourd'hui la souveraineté nationale, nous

OCR Erreurs fréquentes

- Erreurs fréquentes de mauvaise qualité d'un ocr :
- Erreur de segmentation : des zones contenant du texte sont ignorées, ou à l'inverse l'OCR invente des caractères à partir de taches.
- Erreur d'ordre des blocs qui ne sont pas présentés dans l'ordre naturel de lecture.
- Erreur de reconnaissance : un e est lu c, un B est lu 8, etc

OCR Causes courantes d'erreur

- Qualité Numérisation insuffisante : les moteurs OCR préconisent une numérisation à 300 dpi pour les polices supérieures à 10 et à, 400 ou 600 dpi pour les polices inférieures.
- Polices fantaisistes
- Orthographe non standard, écritures à ligatures, noms propres, etc
- texte proche d'éléments non textuels (lignes ou des graphiques)
- texte dans des tableurs, des tableaux ou des formulaires
- lettres espacées ou lettres qui « bavent » ou touchent d'autres lettres
- texte souligné
- texte sur papier de couleur

OCR

- L’OCR doit être paramétré avec soin : langue, police, jeux de caractères accentués, avec ligature, etc)
- Fonction d’apprentissage
- Utilisation des dictionnaires

Qualité : objectifs

- Niveau de qualité dépend de l'usage
- OCR masqué pour indexation à usage grand public :80%
- OCR masqué pour indexation à usage recherche : 95%
- OCR pour faire des ebook : minimum 98% le mieux étant pour de l'éditorial 99,98%

Qualité: Contrôle

1. Vérification orthographique (logiciel d'OCR)
   Tous les logiciels d'OCR disposent d'un vérificateur orthographique intégré qui mettent en valeur les lettres suspectes.

2. Vérification globale
   À la fin du processus, vérifier les pages et les parties du documents (titres de chapitres, paragraphes) pour détecter d'éventuels oublis.

3. Vérification orthographique (logiciel de traitement de textes)
   Utiliser des dictionnaires plus sophistiqués (ex. Word) pour trouver et corriger des erreurs supplémentaires.

4. Vérification par un relecteur
   Relire le document au complet.

OCR

Formats de fichiers **(disponibles avec FineReader)**

- Document Microsoft Word (.doc et .docx)
- Format de texte enrichi (.rtf)
- Texte OpenDocument (.odt)
- Document Adobe Acrobat et PDF/A (.pdf)
- Document HTML (.htm)
- Microsoft Office Excel (.csv, .xls et .xlsx)
- Document texte (*.txt)

...

Numériser au mieux pour que l'OCR soit de la meilleure qualité possible

- Pas de pages bombées.
- Pas de pages trapézoïdales
- Pas de texte du feuillet opposé
- Limiter la transparence de la page
- Pas de travers de l'écrit supérieur à 1,3°
- Un bon contraste
- Une bonne netteté
- Pas de bruit (Une résolution trop importante peut générer du bruit)

MERCI POUR VOTRE ATTENTION