Pour des collections numériques durables

La lecture optique des documents numérisés

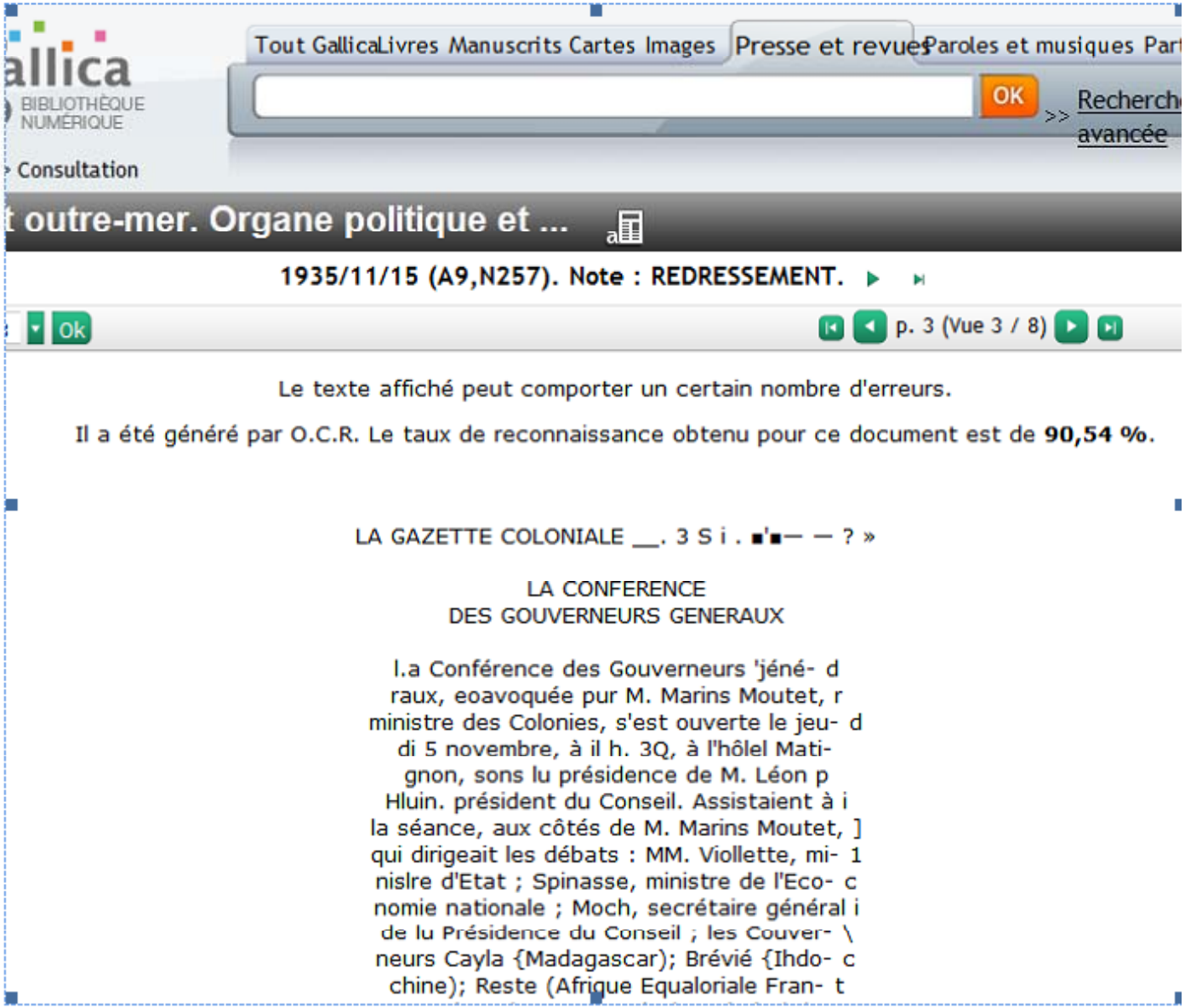
Alain Boucher
Directeur de la numérisation
Bibliothèque et Archives nationales du Québec

Stage pratique sur la numérisation
Dakar, 24 janvier – 1er février 2011

- Rendre possible la recherche dans le texte intégral des documents numérisés est naturellement un objectif visé par la plupart des bibliothèques numériques.
  - Cette possibilité de recherche permet d'obtenir des résultats pour des demandes très spécifiques
  - Elle complète (ou peut même remplacer) avantageusement les moyens d'indexation courants (métadonnées descriptives) quand ces éléments sont sommaires ou font défaut
- La technologie de la lecture optique (OCR) progresse sans cesse, mais elle a des limites qu'il faut connaître.

- La lecture optique donne de bons résultats avec
  - La plupart des imprimés contemporains (livres, revues, thèses, etc.)
  - Les documents dont la typographie est régulière (standard)
- Elle donne de mauvais résultats ou pas de résultats du tout avec :
  - Les manuscrits
  - Les imprimés anciens dont l'impression est irrégulière et qui emploient une typographie très différente de celle qui a cours de nos jours (forme de certaines lettres, ligatures, etc.)
  - Les documents imprimés en petits caractères (moins de 8 points), comme la plupart des journaux anciens ou les dictionnaires

- Qu’est-ce qui détermine un taux satisfaisant ? Le pourcentage de résultats corrects, qu’on fixe généralement à 98% (2 fautes pour 100 caractères, ce qui représente environ une erreur tous les 8 ou 10 mots)
- Les résultats observés par BAnQ pour un journal contemporain (plusieurs dizaines de milliers de pages):
  - 95 à 100 % 57%
  - 90 à 94% 26%
  - 85 à 89% 10%
  - 80 à 84% 4%
  - Moins de 80 % 3%



Le texte affiché peut comporter un certain nombre d'erreurs.

Il a été généré par O.C.R. Le taux de reconnaissance obtenu pour ce document est de **90,54 %**.

LA GAZETTE COLONIALE __. 3 S i . ■'■— — ? »

LA CONFERENCE
DES GOUVERNEURS GENERAUX

l.a Conférence des Gouverneurs 'jéné- d
raux, eoavoquée pur M. Marins Moutet, r
ministre des Colonies, s'est ouverte le jeu- d
di 5 novembre, à il h. 3Q, à l'hôlel Mati-
gnon, sons lu présidence de M. Léon p
Hluin. président du Conseil. Assistaient à i
la séance, aux côtés de M. Marins Moutet,]
qui dirigeait les débats : MM. Viollette, mi- 1
nislre d'Etat ; Spinasse, ministre de l'Eco- c
nomie nationale ; Moch, secrétaire général i
de lu Présidence du Conseil ; les Couver- \
neurs Cayla {Madagascar); Brévié {Ihdo- c
chine); Reste (Afrique Equaloriale Fran- t

- Il faut noter que la lecture optique reconnaît des caractères et des mots sans leur attribuer un sens.
- La phrase *La noix de coco est dure à casser* peut être reconnue comme *La noix de caca est dune à caser.* Le logiciel considérera qu'il s'agit d'un succès total (100% d'exactitude), puisque ce sont tous des mots légitimes.

On peut confier à des prestataires la réalisation de la lecture optique, allant jusqu'à l'analyse plus ou moins automatique du contenu pour structurer les documents (titres, articles, chapitres, sections, etc.)

Ex. Olive Software (États-Unis)

(utilisé par la Bibliothèque nationale suisse pour les journaux)

CCS (Allemagne)

(utilisé par BAnQ pour certains documents de référence)

- Cependant, le plus souvent, on s'en tient à la lecture optique « brute », sans corrections ultérieures.
- Plusieurs logiciels sont disponibles à cette fin. Pour n'en citer que trois:
  - *Abbyy FineReader* (Russie)
  - *OmniPage* (États-Unis)
  - Fonction lecture optique d'*Acrobat* (États-Unis)

- La solution la plus simple et la moins coûteuse: *Acrobat*.
  - Partant des fichiers d'archivage en format TIFF, on assemble le document en PDF mode image avec *Acrobat*
  - On en réalise ensuite la lecture optique avec le module Lecture optique d'*Acrobat*
  - Le processus peut s'automatiser pour de grands volumes avec la version professionnelle d'Acrobat

- Il est toujours prudent de s’assurer de la qualité du résultat:
  - On sélectionne dans *Acrobat* la totalité du texte reconnu
  - On l’examine dans un logiciel de traitement de textes ou un éditeur de textes.
  - Acrobat peut signaler les mots « suspects », mais le processus de correction est laborieux, donc coûteux.

UNIVERSITE CHEIKH ANTA DIOP DE DAKAR

☆★☆★☆★

FACULTE DE MEDECINE, DE PHARMACIE ET D'ODONTO-STOMATOLOGIE

☆★☆★☆★

ANNEE 2001 N°56

LE SYNDROME DE MORRIS :
À PROPOS DE QUATRE OBSERVATIONS COLLIGÉES À L'HÔPITAL ARISTIDE LE DANTEC DE DAKAR

THESE

POUR OBTENIR LE GRADE DE DOCTEUR EN MÉDECINE
(DIPLÔME D'ETAT)

PRÉSENTÉE ET SOUTENUE PUBLIQUEMENT

LE 10 AOÛT 2001

PAR
Marème DIOP *Epouse* FALL

Née le 20 Janvier 1969 à Bruxelles (Belgique)

UNIVERSITE CHEIKH ANTA DIOP DE DAKAR < (' : ' ~ ' [~
&ik &*&*
FACULTE DE MEDECINE, DE PHARMACIE ET D'ODONTO-STOMATOLOGIE
&lk&*&*
ANNEE 2001
J

LE SYNDROME DE MORRIS :
À PROPOS DE QUATRE OBSERVATIONS COLLIGÉES À L'HÔPITAL ARISTIDE LE DANTEC DE DAKAR

JURY

THESE
POUR OBTENIR LE GRADE DE DOCTEUR EN MEDECINE
(DIPLÔME D'ETAT)
PRESENTEE ET SOUTENUE PUBLIQUEMENT
LE 10 AOÛT 2001
PAR

· Marème DIOP *Epouse* FALL
Née le 20 Janvier 1969 à Bruxelles (Belgique)

PRÉSIDENT: M. Baye Assane DIAGNE, Professeur
MEMBRES : M. Mamadou BA,
M. Serigne Maguèye GUEYE,
Mme Haby SIGNATE·SY,

- Une fois un document enregistré en PDF recherchable, on peut bien sûr y effectuer des recherches, mais aussi l'écouter avec les versions récentes d'*Acrobat Reader.*
- Il s'agit d'une voix électronique qui ne donne pas des résultats parfaits, mais cette possibilité est très intéressante pour les personnes qui ont des difficultés de lecture.