Techniques et formats de conversion en mode texte

Certains types de documents gagnent à être diffusés en mode texte, afin de faciliter la recherche au sein des contenus textuels ainsi qu'une meilleure qualité de lecture. Cette valorisation implique plusieurs traitements : océrisation de l'image du document en vue d'un extraire le texte, création de tables de navigation dans le document numérique, et éventuellement conversion des documents sous forme de livres numériques. Tous ces traitements s'appuient sur des formats et des normes en matière de structuration et de diffusion du texte : ALTO, ePub, TEI.

L'océrisation

La technique d'OCR (*optical character recognition*) permet de **situer** et de **reconnaître les chaînes de caractères** dans une image et donc de faire la conversion des mots qui peuvent ensuite être utilisés pour faire une recherche plein texte. Cette conversion est assurée automatiquement par un logiciel et fait l'économie de la retranscription manuelle, beaucoup plus chère. Les mots et chaînes de caractères stockés dans un fichier texte peuvent être réutilisés pour une nouvelle mise en page, exploités dans une base de données, etc.

Exemple de segmentation d'une page de presse
Source : Gallica/BnF

Le principe est la reconnaissance des différentes zones de la page et des caractères contenus dans les zones textuelles. Les caractères sont identifiés à partir de formes mémorisées par le logiciel et de termes déjà connus car présents dans le dictionnaire utilisé par l'outil.

Ce traitement est composé de différentes étapes :
prétraitements : redressement de la page, suppression des bords ;
binarisation : conversion en noir et blanc ;
segmentation : extraction des zones de la page, découpage en lignes, mots, caractères ;
reconnaissance des caractères ;
reconnaissance des mots (avec des ressources linguistiques).

Qualité de l'OCR

Les techniques d'OCR sont en progrès constant pour répondre à la demande très forte, mais la qualité de reconnaissance dépend malgré tout d'un grand nombre de facteurs liés tant au document original qu'à la numérisation elle-même. Ainsi :

les images numériques doivent être suffisamment contrastées et redressées ;
les défauts d'impression (caractères trop empâtés, bavures, a fortiori transparence entre deux pages) diminuent la qualité de reconnaissance et de segmentation des mots ;

les ouvrages en colonnes et/ou illustrés, dans lesquels la lecture n'est pas linéaire sont plus complexes à traiter que les ouvrages à la présentation homogène ;
d'une manière générale, les polices très petites ou au contraire très grandes, et/ou à caractères espacés, sont difficilement traitables ;
les ouvrages en alphabets non latins sont également complexes à traiter, mais les progrès sont plus avancés que sur l'écriture manuscrite ancienne.

Contrôler l'OCR

La BnF propose une méthode de contrôle de la qualité de l'OCR, ainsi que les outils permettant de la mettre en œuvre.

Basée sur les enseignements du projet de recherche Europeana Newspapers, cette méthode vise à évaluer la qualité d'un OCR à travers ses deux principales dimensions :
l'analyse de la page et la reconnaissance de ses différentes composantes (ou « segmentation »)
la reconnaissance des contenus textuels de la page.

Consulter
Contrôle de la qualité OCR **[fichier .pdf – 2003 Ko – 01/06/15 – 63 p.]**

Projet de recherche FUI12 Ozalid

Dans le but d'améliorer la qualité de l'OCR, la BnF a participé au projet de recherche FUI12 Ozalid dont l'objectif était la conception et le développement de la plateforme de *crowdsourcing* CORRECT (correction et enrichissement collaboratifs de texte).

En savoir plus
Plateforme CORRECT