La numérisation des caractères non latins et ses contraintes

La numérisation de masse : un changement d'échelle, un nouveau contexte.

Dans le cadre de sa contribution à l'essor de la bibliothèque numérique européenne (BNuE) /Europeana, la Bibliothèque nationale de France s'est engagée en mars 2007, et ce avec le soutien financier du CNL, à numériser en masse ses collections patrimoniales.

La mise en œuvre de cet ambitieux chantier de dématérialisation (100 000 doc/an sur une durée de 3 ans) a introduit un changement d'échelle considérable dans la politique de numérisation du patrimoine national écrit français, mettant ainsi en lumière un nouveau contexte, tant :

sur le plan politique : la multiplication des initiatives institutionnelles, régionales, nationales, européennes ou internationales ont permis le développement de partenariats pour l'élaboration de chantiers de coopération de numérisation d'envergure (numérisation et valorisation concertées en Histoire coloniale, Guerre de 14-18, Histoire sociale et ouvrière avec la BnF/BDIC/Ministères des Affaires étrangères et de la Défense/La documentation française/Assemblée nationale/Sénat/Bibliothèque administrative de la Ville de Paris/Archives nationales d'Outremer/Musée du quai Branly/ Chambres de commerce et d'industrie/ BM et BU).

sur le plan technique : la numérisation se réalise maintenant en majorité en mode mixte : à la numérisation en mode image -simple photographie du document original- s'associe celle en mode texte, permettant aisément une recherche en plein texte, du mot à la chaîne de caractères grâce à la technique avancée mais encore non aboutie de la reconnaissance optique de caractères, l'OCR.

sur le plan méthodologique : la numérisation de masse a nécessité d'effectuer des sélections par vastes ensembles documentaires correspondant aux fonds thématiques historiques de la Bibliothèque nationale de France organisés par lettrage suivant la classification de Nicolas Clément. Si ce choix s'est imposé, dans un désir d'exhaustivité ce fut sans compter sur les limites imposées par les contraintes documentaires, juridiques, techniques et physiques existantes dans un marché de numérisation consacré à 90% aux publications françaises.

Ce passage à une logique de numérisation de masse, fit que la bibliothèque numérique de la BnF, Gallica, connut un accroissement rapide de ses collections, associé à un développement et une refonte structurelle non négligeable, faisant d'elle aujourd'hui une bibliothèque numérique de référence à l'échelle européenne comme mondiale.

Gallica c'est plus de 2 millions de documents ! en ligne

Valérie Louison-Oudot (Bibliothèque nationale de France)
valerie.louison@bnf.fr

Pour rappel :

- Une bibliothèque patrimoniale et encyclopédique offre un accès gratuit à des :
- Le fonds Gallica
  - Collections libres de droits ou dont les droits ont été négociés avec les ayant-droits
  - Collections des partenaires publics
  - Collections de documents sous droits, proposés par les partenaires commerciaux de la BnF

Slide

1992 : voit la naissance de la bibliothèque numérique Gallica avec la constitution d'un fonds initial de référence de 20 000 documents numérisés accessibles en ligne.

- Mars 1992 : début de la sélection documentaire pour la numérisation
- 1997 : mise en service de Gallica 1 (documents du XIX^{e} siècle)
- 2000 : mise en ligne de la version actuelle de Gallica (documents du domaine public)
- 2001 : 1er dossier « Voyages en France »
- 2005 : Charte documentaire Gallica
- 2007 : Numérisation à grande échelle
- 2007 : Lancement de Gallica 2/offre de services plus interactifs et actualisation des modalités de recherche et de navigation au sein des documents numériques.
- Mars 2008 : Gallica et l'édition contemporaine : Expérimentation de l'offre payante avec les éditeurs (E-Distributeurs)
- Mars 2009 : Le Blog Gallica
- 10 Février 2010 : Gallica atteint le cap symbolique du million de documents en ligne (dont près de 400.000 en mode texte) et adopte une nouvelle interface graphique

Actuellement la BnF achève son second marché de numérisation de masse réalisé par la société JOUVE et rédige le CCTP (Cahier des clauses techniques particulières) et CCAP (Cahier de clauses administratives particulières) de son prochain marché de numérisation en nombre qui devrait débuter en janvier 2014.

La Chaîne de numérisation dans un Marché de numérisation de masse.

Je vais vous retracer les principales étapes d'une chaîne de numérisation, qui transforme un livre physique en un livre numérique, dans le cadre d'un marché de numérisation en nombre.

Slide

Les ouvrages sont sélectionnés intellectuellement, physiquement indexés et traités :

- La sélection physique du document papier est un contrôle de numérisabilité

- L'indexation/exemplarisation (création d'exemplaires numériques, saisie de données d'identification et de métadonnées)

- numérisation dans une large mesure automatisée, numérisation manuelle pour les ouvrages fragiles et OCRisation (reconnaissance optique de caractères), qui est la transformation du "format image" – simple photographie – en "mode texte" – qui permet de réaliser des recherches par mots ou par chaîne de caractères dans le document.

- La saisie des tables de matières (TDM) en TEI = Text Encoding Initiative (permet de décrire la structuration du texte, son imbrication, son découpage, sa description/ Lou Burnard)

- vérification du résultat par le "contrôle qualité"

- mise en ligne des ouvrages sur Gallica et dans le Catalogue général de la BnF

- archivage dans Spar (le système de préservation d'archivage réparti), serveur de stockage des données mis au point par la BnF

Politique documentaire et sélection intellectuelle

Dans le cadre du premier marché de Numérisation de masse réalisé par la société SAFIG, le département Philosophie, Histoire et Science de l'Homme a retenu pour ses sélections documentaires les fonds regroupés sous les lettres cataloguées :

O : Histoire d'Espagne et du Portugal

O2 : Histoire d'Asie

O3 : Histoire d'Afrique

Chacune de ces lettres est subdivisée en sous-section géographie et/ou thématique

Expl :

O2n Chine

O2o Japon

O2s Biographies asiatiques

O2w Publications officielles

Les contraintes documentaires, physiques et techniques propres au marché de numérisation en nombre, réduisent considérablement le nombre de documents éligibles pour un traitement en numérisation. Ainsi même s'il fut effectué un peignage systématique des fonds, il en résulte une sélection physique **non exhaustive**, pour la constitution des lots d'envoi à la numérisation.

Chiffre du département Philosophie, Histoire et Science de l'Homme : Slide

Contraintes physiques :

Slide

Actuellement pour effectuer nos sélections documentaires et physiques d'imprimés en magasin, les sélectionneurs disposent d'un outil d'aide à la sélection nommé ADCAT-15.
Ce logiciel web développé en interne est un miroir du Catalogue général de la Bibliothèque nationale.
Il nous permet de lister un ensemble de cotes, un corpus ou l'intégralité d'un lettrage. Lors d'une séance de sélection physique chaque ouvrage est minutieusement décrit dans ce logiciel.
Pour chaque ouvrage nous devons renseigner quatre critères :

-documentaire
-droit (libre de droit)
-autre numérisation (pas de numérisation antérieure ou si existante de mauvaise qualité)
-physique
L'ensemble de ces contraintes sont préalablement définies lors de la rédaction du cahier des charges du marché de numérisation et plus précisément dans le Cahier de Clauses Techniques Particulières (CCTP).

Slide

Ainsi nos documents doivent être en majorité en langue française publiés en France (critère exclusif du marché SAFIG) ou à l'étranger (élargissement des critères / marché JOUVE), mais aussi en langue étrangère dans la mesure où l'ouvrage est publié en France ou qu'il est constitué à 80% en langue française et contenant au maximum 20 % de caractères non latins.
L'ensemble de ces contraintes a eu pour effet d'invalider un grand nombre de documents pouvant prétendre à un traitement de numérisation pour leur intérêt documentaire.

Ainsi des textes de référence n'ont pu être numérisés en 2008 pour cause de langue et de typographie. Un texte en bilingue est refusé pour envoi, pour le motif d'un taux trop élevé de caractères non latins.

Slide expl. Refus

FOL- O2N- 1396 / Refusé Langue, pays.

Fol-O2n-1396
Millot, Stanislas
- Dictionnaire des formes cursives des caractères chinois. – Paris, 1909.

Type : **texte imprimé, monographie**

Auteur(s) : **Millot, Stanislas (pseud. Aldébaran)**

Titre(s) : **Dictionnaire des formes cursives des caractères chinois, par Stanislas Millot,... [Texte imprimé]**

Publication : **Paris : E. Leroux, 1909**

Description matérielle : **In-fol., 202 p.**
Notice n° : **FRBNF30949179**

4- O2O- 368 / Refusé Langue, pays. L'ère Meiji (1853-1912)

4-O2o-368 (n°15, 1901 par exemple)
Résumé statistique de l'empire du Japon : Texte bilingue japonais-français. – Tōkyō, 1887-1939.

Type : texte imprimé, publication en série
Auteur(s) : Japon. Statistics bureau
Titre de référence : Dai Nihon teikoku tōkei tekiyo

Titre(s) : Dai Nihon teikoku tōkei tekiyo [Texte imprimé] / Naikaku tōkeikyoku hensan
Numérotation : Dai 1-kai (Meiji 20-nen [1887])-dai 53-kai (1939)
Publication : Tōkyō : Naikaku tōkeykyoku, 1887-1939
Description matérielle : 53 vol. : cartes ; 26 cm

Note(s) : Texte bilingue japonais-français
L'intitulé français de la collectivité varie
Autre(s) forme(s) du titre :
- Titre(s) parallèle(s) : Résumé statistique de l'Empire du Japon / Cabinet impérial, Section de la statistique générale
Indice(s) Dewey : 315.2 (20e éd.)
Notice n° : FRBNF32853433

Si l'on va plus avant dans le déroulé de la chaîne de numérisation une fois que nos documents sont sélectionnés intellectuellement, physiquement, ceux-ci subissent un traitement de dématérialisation, procédé réalisé par un prestataire extérieur.

Contrainte technique :

Les multiples plans de numérisation entrepris par les institutions ou collectivités locales françaises sur le patrimoine écrit depuis maintenant 15 ans, nous révèlent aussi bien les rapides évolutions techniques qu'a pu connaître le secteur du numérique, l'existence des différentes strates, de modèles de numérisation voire leur coexistence sur les serveurs Internet des bibliothèques numériques françaises.

Slide

4-O2-545 (7)

Titre : Recueil d'itinéraires et de voyages dans l'Asie centrale et l'Extrême-Orient...

Éditeur : E. Leroux (Paris)
Date d'édition : 1878
Sujet : Asie centrale
Type : monographie imprimée
Langue : Français
Format : 1 vol. (III-380 p.) : carte ; Gr. in-8
Format : application/pdf
Droits : domaine public
Identifiant : ark:/12148/bpt6k5441003m
Source : Bibliothèque nationale de France, département Philosophie, histoire, sciences de l'homme, 4-O2-545 (7)
Relation : http://catalogue.bnf.fr/ark:/12148/cb31313919v
Description : Collection : Publications de l'École des langues orientales vivantes ; 7
Description : Comprend : Journal d'une mission en Corée / traduit du chinois par F. Scherzer ; Mémoires d'un voyageur chinois sur l'empire d'Annam / par Tsai-tin-lang ; traduit par L. Leger ; Itinéraires de l'Asie centrale / par Alexandre-Pavlovitch Khorochkine ; traduit par L. Leger ; Itinéraire de la vallée du moyen Zerefchan / par Leopold Theodorovich Radlov ; traduit par L. Leger ; Itinéraires de Pechaver à Kaboul, de Kaboul à Qandahar et de Qandahar à Hérat] / par Mohamed Abdoul Kerim Mounchy ; [Extraits de l'ouvrage Tarikh-é-Aḥmad, traduits par C. Schefer]
Provenance : bnf.fr
Date de mise en ligne : 18/11/2008

Si je me réfère à la Bibliothèque numérique Gallica l'on peut constater ce phénomène d'empilage et de coexistence de modèles de numérisation, lié aux plans successifs de dématérialisation entrepris en interne ou en collaboration depuis 1992.

Ainsi un internaute pourra consulter des DN :

- en mode image seul :
  les pages du livre ont été scannées et les images de ces pages sont accessibles en ligne mais aucune recherche en plein texte ne peut être effectuée.

- en mode texte seul :
  tout en prenant garde que soit visible par l'internaute à l'ouverture d'un DN un texte restitué à l'identique soit un texte ayant conservé la structure de l'original papier (sa mise en page, paragraphe, sa casse, et non un texte brut (dit aussi plein texte –texte écrit au kilomètre). Par ce procédé une recherche par mots ou chaîne de caractères est possible et ce grâce au travail d'océrisation fait en amont sur le fichier numérique du document

- en mode mixte ou texte et image sont couplés, ce qui représente aujourd'hui la norme.

La numérisation transforme un document papier en données informatiques. Ce procédé de dématérialisation d'un document consiste dans un premier temps à l'obtention d'un fichier image de ce document papier et dans un second temps d'un fichier texte de ce même document. Cette opération n'est pas sans contrainte.

Slide

OCR (Optical Character Recognition) : la reconnaissance optique des caractères non latins et Océrisation de 20% de caractères non latins.

Pour obtenir un fichier texte d'un document tout prestataire utilise les procédés techniques de la reconnaissance optique de caractère dit OCR= **Optical Character Recognition** et ce via son logiciel d'océrisation intervenant sur les fichiers images.

La technique d'OCR 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 (technologie propre à chaque prestataire). Cependant de toute la chaîne de numérisation c'est cette étape de conversion qui s'avère des plus ardues.

Le principe de l'OCRisation est la reconnaissance de caractères à partir de formes mémorisées par le logiciel et de termes déjà connus car présents dans le dictionnaire utilisé par l'outil. 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.

Chaque espace et chaque chaîne de caractères (appelé "string") est donc identifié précisément et doit être restitué dans le sens principal de lecture.

La qualité de l'OCR dépend du document original et de la qualité de la numérisation.

De fait chaque prestataire à ses propres contraintes techniques et sémantiques :
par exemple Jouve comme Safig lors du premier marché de masse, ne peuvent océriser dans un ouvrage majoritairement de langue latine pas plus de 20% de caractères non latins parce qu' un grand nombre de langues vernaculaires ont des caractères cursifs à l'encrage trop épais ou des caractères aux formes atypiques pour notre hémisphère, ou use de polices de caractères trop empâtées, ne pouvant être reconnues par leur logiciel d'océrisation et nécessite une retranscription manuelle donc beaucoup plus coûteuse en moyen humain et financier.

Comme l'évoque si justement Jacques André dans son article intitulé : « Numérisation et codage des caractères de livres anciens », « [ ], *avant de faire de la reconnaissance de caractères, encore faut-il avoir une certaine connaissance de ceux-ci*. »

Cette étape d'océrisation qui nous semble si naturelle pour nous, sachant toutefois qu'un être humain ne connaît pas toutes les langues existantes, est donc un procédé des plus complexes pour un ordinateur. S'il n'y avait pas de logiciel d'océrisation qui assure l'automatisation du procédé, il faudrait des petites mains qui passeraient des journées sur Word pour retranscrire le contenu des livres nous ramenant à l'époque des copistes du moyen-âge. L'OCR est encore actuellement un sujet de recherche en plein essor. Cette technique fait intervenir des domaines tels que l'intelligence artificielle, la reconnaissance de formes et le traitement d'image.
Par conséquent chaque pays à tendance à se focaliser sur la reconnaissance de son alphabet, même Google qui fait référence en matière d'océrisation de documents a besoin de l'intervention humaine pour améliorer ses taux de reconnaissance de caractères grâce au recaptcha.

Un **CAPTCHA** est une forme de **test de Turing** permettant de différencier de manière automatisée un **utilisateur humain** d'un **ordinateur**.

C'est un test de défi-réponse utilisé dans le domaine de l'informatique, ayant pour but de s'assurer qu'une réponse n'est pas générée par un ordinateur. L'acronyme « CAPTCHA » est basé sur le mot *capture*, et vient de l'anglais *completely automated public Turing test to tell computers and humans apart*.

Parce que le test est réalisé par un ordinateur, en opposition avec les **tests de Turing** standard réalisés par des humains, un captcha est souvent décrit comme un test de Turing inversé.

En numérisation de livres : le **reCAPTCHA** propose deux mots dont le premier est connu et sert de CAPTCHA et dont le second est incertain voire inconnu car issu de la numérisation d'un livre.

reCAPTCHA est un système mettant à profit les capacités de reconnaissance des **utilisateurs humains** mobilisées par les tests **Captcha**, pour améliorer par la même occasion le processus de **numérisation de livres**, là où échouent les systèmes de reconnaissance optique de caractères (OCR). Le système a été mis au point par des chercheurs de l'Université Carnegie-Mellon et appartient à Google depuis fin 2009.

Bilan : Google numérise des livres en usant gratuitement du contrôle qualité fait à l'insu des internautes.

exemple de captcha quotidien et non désiré:

Si l'on numérise un livre en ouverture réduite (soit -90°), les mots situés le long de la marge intérieure sont soumis, d'autant plus si le livre est épais et la marge intérieure insuffisante, au bombage du livre entrainant sur les caractères un effet déformant visible à la numérisation.

Slide

En sortie de chaîne de numérisation l'on obtient un document numérique (DN).

Document numérique : répertoire produit et transmis par le prestataire et correspondant à un exemplaire numérique. Le document numérique contient :

- un fichier de métadonnées (refNum)
- un répertoire d'images (TIFF ou JPEG 2000)
- un répertoire d'OCRisation (fichier Alto)
- un fichier table des matières (TEI)

Le cadre du futur projet du portail France-Japon permettra à notre département de repeigner nos cotes dans un désir de complétude et de mise en valeur de documents de référence majoritairement en bilingue pour la période l'ère Meiji (1853-1912) comme de la période antérieure.