Actualité de la numérisation

Mathieu Andro

Actualité de la numérisation

Mathieu Andro
Bibliothèque Sainte-Geneviève

Après avoir travaillé pour les bibliothèques du Muséum national d'histoire naturelle puis été directeur de la bibliothèque de l'École nationale vétérinaire de Toulouse, **Mathieu Andro** *est actuellement chef de projet numérisation à la bibliothèque Sainte-Geneviève. Il est porteur, dans le cadre du PRES Sorbonne Paris-Cité, d'un projet de bibliothèque numérique mutualisée offrant des services innovants (correction participative de l'OCR, TEI, numérisation à la demande, impression à la demande).*

Les enjeux de la numérisation

Nombre de bibliothèques, mais aussi de prestataires privés comme Google, sont désormais engagés dans de vastes campagnes de numérisation de documents physiques, imprimés ou autres. La mise en œuvre de ces opérations prend en compte les nombreux avantages que peut présenter la numérisation, ainsi :

- une diffusion plus large, théoriquement en tout lieu, et un accès facilité, à distance, sans contrainte d'accès ni de temps ;
- la préservation et la conservation des documents originaux, notamment les documents fragiles et précieux, qui n'ont dès lors plus à être manipulés pour être communiqués aux chercheurs, sauf dans des cas exceptionnels ;
- l'exploitation scientifique démultipliée par la consultation simultanée de vastes corpus de textes numérisés, exploitation souvent facilitée par la prise en compte de structurations très efficaces des formats informatiques des documents numérisés, comme la « Text Encoding Initiative » (TEI)[1].

D'évidence, de telles opérations favorisent la recherche et l'innovation en rendant accessibles et exploitables des masses importantes d'informations. Elles permettent aussi, aux bibliothèques comme aux prestataires, de proposer à leurs usagers des services d'impression à la demande (« *print on demand* »), voire de fourniture de livres électroniques (e-books) et de bénéficier ainsi d'un retour sur investissements.

Enfin, il est envisageable que la numérisation permette, à l'avenir, d'identifier automatiquement les plagiats dans l'histoire de la littérature. Ceci pourrait modifier sensiblement la perception que nous en avons !

Que numériser ?

Au niveau mondial, la mise en œuvre d'un gigantesque projet de numérisation de plusieurs millions de livres, Google Books[2], a déjà permis de traiter d'importants fonds qu'il serait donc peu opportun de numériser à nouveau. En France, la Bibliothèque nationale de France a quant à elle mis en place « sa » bibliothèque numérique, Gallica[3], déjà forte de plus de 280 000 livres et de dizaines de milliers d'autres documents (fascicules de revues, photographies, cartes, etc.).

Dès lors, une bibliothèque qui choisirait malgré tout de s'engager dans la numérisation d'une partie de ses fonds devrait sélectionner ceux-ci à partir de critères de rareté, en privilégiant par exemple les manuscrits, les documents comportant des dédicaces, les « Unica », c'est-à-dire les documents qu'elle est seule à posséder, ou ceux qu'elle estime disponibles sous une forme numérisée appauvrie, etc. Dans le même souci de spécificité des collections numérisées, il est aussi possible de s'intéresser à des « niches thématiques » s'appuyant sur les fonds spécialisés de l'établissement.

Il faut aussi combattre l'idée que la numérisation ne serait qu'une pâle reproduction du papier. Au contraire, la numérisation permet une valorisation éditoriale, notamment par le biais de la TEI déjà citée, ainsi qu'en témoignent, par exemple, les Bibliothèques virtuelles humanistes[4] mises en œuvre par le Centre d'études supérieures de la Renaissance de Tours. Les techniques de *text mining* permettent désormais des exploitations très sophistiquées des corpus de textes numérisés. Par exemple, Google a mis en place un outil de bibliométrie,

1. www.tei-c.org/index.xml
2. http://books.google.fr
3. http://gallica.bnf.fr
4. www.bvh.univ-tours.fr

Google Ngram Viewer[5], qui permet de mesurer l'occurrence des mots dans l'ensemble du corpus qu'il a numérisé. On peut ainsi dater l'apparition ou la disparition de mots ou constater des évolutions d'écriture, ou encore mesurer, dans le temps, le taux de citation d'auteurs, d'évènements, de dates...

Techniques et acteurs

Sans qu'il soit question d'examiner ici les complexes questions techniques liées à la numérisation des fonds, trois principaux fondamentaux doivent toujours être pris en compte dans la mise en œuvre d'une campagne de numérisation :

- Distinguer les fichiers de conservation et les fichiers de diffusion, qui obéissent à des contraintes de numérisation, de stockage, etc., différentes. Les premiers sont destinés à un archivage pérenne tandis que les seconds, qui contiennent un volume moindre de données, seront diffusés sur le web.
- Privilégier la numérisation par le biais de la « reconnaissance optique de caractères » (OCR), garante des meilleures possibilités de recherche et d'exploitation des textes. Le fichier image obtenu à l'issue de la numérisation « simple » ne le permet pas. Il est donc nécessaire d'utiliser un logiciel qui va identifier l'image de chaque caractère et générer ainsi un fichier texte à partir du fichier image.
- S'assurer de l'interopérabilité des fichiers obtenus, c'est-à-dire de la possibilité pour d'autres bibliothèques, d'autres prestataires, de pouvoir exploiter les métadonnées créées pour accompagner les documents numérisés, en respectant le protocole de moissonnage OAI-PMH[6].

Pour ce qui est des acteurs de la numérisation, en plus de Google Books, qui annonce avoir numérisé près de 15 millions d'ouvrages, et de Gallica, déjà évoqués, on peut citer Europeana[7], auquel la Bibliothèque nationale et Gallica participent activement, et qui « moissonne » actuellement plus de 15 millions de documents. Deux autres « acteurs » américains peuvent être cités, dont les ambitions et les modes de fonctionnement sont très différents de ceux de Google Books : le Hathi Trust[8], qui revendique plus de 6 millions de documents numérisés, et regroupe plus de 50 partenaires institutionnels (bibliothèques et grands organismes de recherche essentiellement) ; Internet Archive[9], qui, comme son nom l'indique, se propose d'archiver l'internet, comprend déjà plus de 2 millions de ressources... et 150 milliards de pages archivées !

Comment diffuser ?

Aussi étonnant que cela puisse paraître, la majeure part des documents numérisés par les bibliothèques ne sont pas disponibles en ligne, et « dorment » sur des DVD ou des disques durs dont la durée de vie est d'ailleurs très limitée. Les raisons de cette non-diffusion sont multiples : ainsi, il n'existe pas de plateforme de diffusion pour les documents produits par les bibliothèques relevant de la recherche et de l'enseignement supérieur. Les deux principaux acteurs du marché ne sont pas forcément disponibles pour de nouveaux hébergements, ni Google, qui ne recherche plus de nouveaux partenaires en France, ni Gallica, qui n'a pas vocation à héberger tous les types de collections numériques.

Par contre, Gallica pourra moissonner les documents mis en ligne par d'autres bibliothèques à condition que ces bibliothèques soient parvenues à développer leurs propres plateformes et à mettre en place leurs propres entrepôts OAI de métadonnées, faisant apparaître dans Gallica les notices des documents numérisés par ces bibliothèques, accompagnées d'un lien. Mais, si ces conditions, coûteuses en personnel et en budget, ne sont pas remplies par les bibliothèques, Gallica ne pourra pas leur offrir un débouché ; du moins, tant que les projets de tiers archivage et de Gallica marque blanche n'auront pas été mis en œuvre par la BnF.

La solution de développer en propre une plateforme de diffusion doit autant que possible être évitée, car elle donne rarement des résultats satisfaisants : la visibilité des documents ainsi diffusés reste faible, les possibilités de recherche et d'exploitation sont souvent frustes, etc.

La solution à privilégier est donc la mutualisation, qui permet un partage des coûts et une meilleure visibilité, d'autant plus grande que le nombre de documents diffusés est important, car l'algorithme « PageRank » de Google, qui permet de classer les résultats d'une requête dans le moteur de recherche par ordre de pertinence, prend largement en compte le nombre de liens qui pointent vers les sites. Plus le nombre de documents proposés sera important, plus le nombre de liens pointant vers le site le sera et plus son PageRank et donc sa visibilité seront forts ; de plus, un projet collectif est, plus qu'un projet individuel, une garantie de qualité et de pérennité. Enfin, la mutualisation permet malgré tout de respecter l'identité de chacun, par le biais d'un portail spécifique permettant d'avoir un nom de domaine particulier, un graphisme et un logo indépendants, des statistiques de consultation propres.

D'autres exploitations des documents numérisés

La numérisation des documents permet d'envisager de fournir aux usagers des services inédits dans le « monde » des documents physiques.

Correction participative de l'OCR

Quelles que soient les performances des logiciels utilisés, les opérations de numérisation en mode OCR comportent toujours un pourcentage d'erreurs souvent inacceptable

5. http://ngrams.googlelabs.com
6. Open Archives Initiative Protocol for Metadata Harvesting : www.openarchives.org/OAI/openarchivesprotocol.html
7. www.europeana.eu/portal
8. www.hathitrust.org
9. www.archive.org

quand il s'agit de recherche en texte intégral, d'indexation par les moteurs de recherche ou de lecture sur des tablettes numériques. Un certain nombre d'opérations de correction participative sont mises en œuvre, impliquant l'aide de l'ensemble des usagers des bibliothèques numériques concernées. On peut citer le partenariat entre la Bibliothèque nationale de France et Wikisources [10], celui entre Google Books et Captcha [11], mais aussi l'Australian Newspapers Digitisation Program [12]. Dans le premier cas, ce sont les internautes qui corrigent bénévolement les textes océrisés. Dans le second cas, Google profite du travail de saisie de milliards d'internautes qui doivent re-saisir un mot déformé pour prouver qu'ils ne sont pas de malveillants robots afin de pouvoir créer des comptes sur divers sites web. Sans le savoir, ils travaillent ainsi à la correction de l'OCR pour Google Books.

Ebooks on demand

La Bibliothèque interuniversitaire de médecine participe déjà au réseau européen « Ebooks on demand » (EOD)[13], qui offre une plateforme permettant aux internautes de commander en ligne la numérisation de tel ou tel livre de leur catalogue. Ce réseau pourrait également, dans le cadre d'une délégation de service public, offrir la possibilité de numériser à la demande des documents, en proposant éventuellement aux internautes ou à des mécènes de financer cette numérisation. Dans ce cadre, c'est le prestataire qui facturerait directement à l'internaute le coût du service. La bibliothèque, quant à elle, offrirait un nouveau service à ses usagers (un service de reproduction numérique professionnel externalisé) sans avoir à en supporter le coût. Au contraire, la politique d'acquisition de la bibliothèque numérique serait ainsi partagée avec le grand public qui contribuerait à la compléter. Nous pourrions ainsi voir apparaître la mention « *ce livre a été numérisé grâce au soutien de Madame X, de l'UMR CNRS Y, ou de la fondation Z* ». Enfin, en signalant aux autres établissements les documents qui, pour une raison ou pour une autre, n'ont pas encore pu être numérisés mais qui le seront un jour, la bibliothèque favoriserait l'harmonisation des politiques de numérisation de manière plus efficace que par le simple échange de listes ou par un signalement trop général de grands corpus.

Impression à la demande/ Print on demand

L'impression à la demande permet, à partir de fichiers numérisés, d'imprimer sur papier des documents en très petite quantité, voire à un seul exemplaire, sans que les coûts soient pour autant excessifs, et avec une qualité de restitution très proche des documents papier « originaux ». Le « print on demand » permet ainsi aux éditeurs de s'affranchir de la (coûteuse) gestion de stocks, et de produire « en flux tendu », c'est-à-dire en adaptant au plus juste l'offre et la demande. Il existe désormais de nombreux prestataires dans ce domaine, comme Amazon BookSurge [14], Jouve [15], Librissimo [16], UniBook [17], etc. La bibliothèque ayant conduit un programme de numérisation peut ainsi offrir un service supplémentaire à ses lecteurs tout en bénéficiant d'une petite marge sur la vente de ces imprimés et en permettant à des livres épuisés depuis longtemps de ressusciter sur support papier. La bibliothèque de l'université du Michigan [18] elle-même propose désormais ce type de prestation. Il existe même, à l'achat, des machines spécialement dédiées à cet usage, qui peuvent être installées en libre accès dans les bibliothèques, comme l'Espresso Book Machine [19], à partir de laquelle il est possible d'obtenir en moins de 5 minutes une version imprimée et brochée de tout document issu de Google Books ou d'Internet Archive.

La veille stratégique du projet de numérisation

Afin de conduire un projet innovant de numérisation, la mise en place d'un dispositif de veille peut s'avérer fort utile. En effet, au lieu d'aller chercher périodiquement de l'information en saisissant les mêmes requêtes dans les mêmes sources d'informations, il peut être préférable de faire remonter automatiquement toute nouvelle information correspondant à ces paramètres. Ainsi, on sera prévenu très rapidement d'éventuelles menaces (nouveaux projets susceptibles de remettre en question la pertinence de son propre projet ou encore *e-reputation*), mais surtout de détecter des opportunités stratégiques (innovations, partenariats, appels à projets).

Des outils comme Google Reader ou Netvibes permettent d'agréger de nombreux flux RSS et peuvent même être synchronisés avec Twitter. Enfin, rien ne remplace le renseignement humain, qu'on peut trouver auprès des tutelles, des entreprises, d'autres institutions. ●

Juillet 2011

Orientations bibliowebographiques

Les informations et documents signalés au cours de l'exposé ont été partagés sur le site :

www.bibliotheque-numerique.fr

Ce site propose entre autres une étude sur les solutions de diffusion, des tests sur diverses plateformes, un tableau de bord, des outils de veille et des cahiers des charges sur les différents aspects d'un projet de numérisation : numérisation proprement dite, plateforme de diffusion, impression à la demande, numérisation à la demande.

10. http://fr.wikisource.org

11. www.captcha.net

12. www.nla.gov.au/ndp

13. www.books2ebooks.eu

14. www.booksurge.com

15. www.jouve.fr

16. www.librissimo.fr

17. www.unibook.com/fr

18. www.lib.umich.edu/digital-library-production-service-dlps

19. www.ondemandbooks.com