Diffuser des documents numérisés

Mathieu Andro

Diffuser des documents numérisés : Internet Archive, text mining et crowdsourcing

Mathieu ANDRO
Inra, DIST

Résumé introductif
Cet exposé a pour ambition centrale de proposer une solution pragmatique afin que les institutions puissent diffuser sur le web le produit de leur numérisation. La solution Internet Archive qui est mise en avant a l'avantage d'être à la fois simple, de ne pas nécessiter de formation, d'être gratuite, d'être pérenne, d'offrir la meilleure visibilité sur le web et les meilleures fonctionnalités à leurs usagers. Mais, au-delà de la question de la diffusion, l'exposé propose également des applications possibles à la numérisation du patrimoine autour du text mining et du crowdsourcing et résume les principaux résultats d'une thèse en cours au sujet du crowdsourcing appliqué aux bibliothèques numériques.

Mots clés
Bibliothèques, numérisation, bibliothèques numériques, Internet Archive, text mining, crowdsourcing, crowdfunding

1. Comment développer une bibliothèque numérique gratuitement et simplement pour une visibilité, une pérennité et des fonctionnalités optimales ?

Il y a quelques années, nous avions mené une étude (Andro, 2012) consacrée à 10 logiciels dédiés au développement de bibliothèques numériques : YooLib (Polinum), Invenio (CERN), ORI-OAI (Universités), DSpace (DuraSpace), DigiTool (Ex Libris), Mnesys (Naoned), ContentDM (OCLC), Eprint (Université de Southampton), Greenstone (Université de Waikato) et Omeka (Université George Mason). A l'occasion de cette étude, nous avions déploré que la majeure partie des documents numérisés par les bibliothèques territoriales et universitaires en France ne soit pas mise en ligne et « dorme » sur des DVD et des disques durs dont la durée de vie demeure limitée. Les bibliothèques renoncent ainsi souvent à mettre en ligne le résultat de leurs campagnes de numérisation pour plusieurs raisons. La première d'entre elles, c'est que la bibliothèque numérique de la Bibliothèque nationale de France (BnF), Gallica, ne peut pas héberger de documents numérisés en dehors des

chaînes de la BnF ou ne disposant pas de notice bibliographique dans son catalogue. Si une institution parvient à développer sa propre bibliothèque numérique avec son propre entrepôt OAI de métadonnées, alors la BnF pourra en moissonner le contenu. Or, c'est précisément une difficulté pour de multiples institutions mal informées sur ce sujet. Par ailleurs, le développement d'une plate-forme autonome est très coûteux en infrastructures, et en ressources humaines et logicielles, pour un résultat très rarement satisfaisant du point de vue des fonctionnalités et de la visibilité sur le web. Dans ces conditions, nombreuses sont les bibliothèques qui renoncent finalement à diffuser sur le web le fruit de leurs campagnes de numérisation.

Et, en dehors d'un Gallica encore insuffisamment ouvert, il n'existe pas d'autre débouché pertinent en France, par exemple au niveau de l'Enseignement Supérieur ou même en Europe, Europeana n'étant aussi qu'un simple moissonneur de métadonnées, non un hébergeur de fichiers.

Dans ces conditions, la non diffusion ou la mauvaise diffusion sur le web des documents numérisés par les bibliothèques représente un gaspillage d'argent public important, d'autant que faute de mise en ligne (ou d'une mise en ligne sur un site mal référencé sur le web), le même imprimé risque fort d'être numérisé plusieurs fois.

La solution réside peut être dans le projet « Gallica marque blanche » qui consisterait à ce que Gallica puisse enfin héberger des fichiers produits par des bibliothèques. Mais, pour des raisons aussi techniques que probablement culturelles, ce projet d'ouverture à la participation de professionnels extérieurs à l'institution, est annoncé comme imminent depuis des années et n'a consisté, à ce jour, qu'à livrer Gallica en tant que logiciel supplémentaire par rapport à ceux qui existent déjà sur un marché saturé, non à proposer un hébergement direct, sur les serveurs de la BnF, des documents numérisés.

En attendant, la solution qui nous semble la plus efficace est le recours à des plateformes mutualisées comme Internet Archive, le Hathi Trust ou Flickr pour les images. C'est un moyen simple qui offre à la fois visibilité, qualité, pérennité et partage des coûts. Son seul inconvénient réside dans la perte d'autonomie de décision comme c'est le cas dans tout cadre collectif. Dans une autre étude (Andro, 2015, 1), nous avons montré qu'un livre numérisé et diffusé sur Internet Archive était consulté en moyenne plus de 5 fois par mois quand il ne l'était que quelques dixièmes de fois sur une bibliothèque numérique autonome. En effet, une bibliothèque numérique contenant des millions de documents aura inévitablement plus de liens sur le web pointant vers son nom de domaine qu'une bibliothèque numérique de quelques centaines de document. Son PageRank sera donc meilleur. Elle sera mieux référencée dans la liste des réponses à une requête Google et ses documents

seront donc plus consultés. Au contraire, la diffusion d'un livre numérisé sur une petite bibliothèque numérique risque fort de s'apparenter finalement à une non diffusion si celle-ci demeure dans le web invisible ou est mal référencée par les moteurs de recherche, ce qui est, hélas, bien souvent le cas.
Les objectifs des plateformes mutualisées et des bibliothèques numériques autonomes sont diamétralement opposés. Dans le dernier cas, le choix est fait de créer une représentation numérique de l'institution et de ses collections sur le web souvent afin de satisfaire la soif de reconnaissance d'un directeur d'institution ou d'un élu dans une logique de communication institutionnelle. Dans le cas d'une bibliothèque numérique mutualisée, on cherche surtout à satisfaire les besoins des usagers et à aller là où ils sont déjà, c'est-à-dire, en première page d'une requête Google sur un titre de livre au lieu de chercher à les attirer sur un site institutionnel. Du point de vue communicationnel, cette stratégie s'avère d'ailleurs aussi terriblement efficace puisque le nom de la bibliothèque sera finalement vu par un nombre bien plus important d'internautes sur une grande plateforme mondiale que sur un site institutionnel aussi coûteux qu'invisible.
A l'issue de nos études et de nos expériences, nous considérons donc que Internet Archive est la meilleure solution de diffusion pour les bibliothèques territoriales et universitaires. Internet Archive est une organisation internationale non gouvernementale et à but non lucratif sur le modèle de Wikipédia. Elle bénéficie, grâce à du mécénat public et privé, de plus de 15 millions de dollars par an, de 200 personnes dont 50 programmeurs, bibliothécaires et administrateurs, de plusieurs data centers aux USA et de plusieurs serveurs miroir dans le monde. Internet Archive propose aussi l'archivage du web avec sa célèbre Wayback machine, des métadonnées de livres (OpenLibrary), des vidéos, des sons et des jeux vidéos. Concernant les livres numérisés, au moment où nous écrivons ces lignes, ce sont près de 10 millions d'entre eux qui sont déjà diffusés notamment aux formats pour liseuses EPUB et MOBI. En France, nous l'avons utilisé à la Bibliothèque Sainte-Geneviève et à l'Institut National de la Recherche Agronomique. A la suite de ces institutions, Sciences Po Paris participe également, et totalement gratuitement, à Internet Archive. Et d'autres grandes institutions devraient suivre prochainement. Les collections de toutes ces institutions disposent sur Internet Archive d'entrepôts OAI PMH moissonnables par Gallica et Isidore. Cela leur permet d'être visible sur la bibliothèque numérique nationale, mais aussi de pouvoir bénéficier d'éventuels financements de la BnF, l'existence d'un tel entrepôt étant une condition requise.

Last but not least, le versement de fichiers sur Internet Archive est aussi intuitif que de diffuser une photographie sur Flickr ou une vidéo sur YouTube.

2. Au delà de la numérisation : text mining et crowdsourcing

Mais, la numérisation sera bientôt de l'histoire ancienne et il devient déjà de plus en plus difficile d'identifier des imprimés méritant encore d'être numérisés après le passage de Google Books qui a déjà numérisé plus de 30 millions de livres. Dans ces conditions, la question ne sera bientôt plus de savoir où diffuser le produit de sa numérisation mais plutôt de savoir ce qu'il faut en faire et quelles applications en trouver.

En premier lieu, les données numérisées doivent être conservées au-delà des changements de logiciels, de formats et de supports, afin qu'elles demeurent accessibles aux générations futures. Outre les solutions internationales déjà mentionnées, des solutions d'archivage pérenne sont proposées au niveau national par la Bibliothèque nationale de France avec son Système de Préservation et d'Archivage Réparti (SPAR) ou par le Centre Informatique National de l'Enseignement Supérieur (CINES).

Ensuite la question des fonctionnalités de lecture doit également être posée. La quasi-totalité des bibliothèques numériques autonomes développées par les collectivités et les universités imposent la lecture des documents numérisés sur un écran de PC. Seules les plateformes collectives comme Internet Archive proposent systématiquement des fichiers aux formats EPUB et MOBI afin de pouvoir être lus sur tous types de liseuses.

Concernant le futur de la numérisation, il est également probable que, lorsque les œuvres de la deuxième moitié du 20è siècle pourront enfin être numérisées, ont identifie qu'un certain nombre d'entre elles sont le résultat d'un plagiat d'œuvres plus anciennes. L'identification du plagiat rendu possible par la numérisation des textes engendrera probablement une révision de notre perception de l'histoire de la littérature mais aussi des sciences. Et il n'est pas impossible que certaines statues aujourd'hui idolâtrées soient un jour déboulonnées.

Mais les possibilités offertes par la numérisation des textes résident aussi dans la possibilité de fouiller au sein de ces textes, grâce aux technologies de text mining, et d'en extraire des données afin de produire de nouvelles analyses et de nouvelles connaissances. Ainsi, au lieu de rechercher un mot dans un document, on peut rechercher des dictionnaires de mots, des thesauri, des ontologies dans des corpus de documents et des bibliothèques entières. Dans le domaine des humanités numériques, on évoque parfois la

culturomique comme une science permettant de produire de telles analyses. On peut, par exemple, déjà observer l'évolution temporelle des verbes irréguliers anglais dans la littérature numérisée par Google Books rendue accessible via Google Ngram Viewer. On peut aussi remarquer la disparition du nom de Léon Trotsky dans la littérature soviétique entre la mort de Lénine en 1924 et les débuts de la déstalinisation par Khrouchtchev au XXè congrès du Parti Communiste de l'Union Soviétique…

Les bibliothèques numériques offrent enfin des possibilités de collaborations avec les usagers, en faisant appel à du travail bénévole d'internautes (crowdsourcing) pour corriger de l'OCR, indexer ou encore pour financer la numérisation de livres (crowdfunding). On distingue ainsi plusieurs formes de crowdsourcing au sein de la taxonomie que nous proposons (Andro, 2016) :

- le **crowdsourcing bénévole** classique
- le recours au travail des internautes sous la forme de jeux, qualifié aussi de **gamification** (Andro, 2015, 4)
- le recours au travail involontaire et inconscient des internautes qualifié aussi de **crowdsourcing implicite**. Par exemple, Google fait corriger l'OCR brute des textes qu'il a numérisés grâce aux internautes qui doivent ressaisir des mots déformés afin de prouver qu'ils ne sont pas des robots malveillants et pouvoir créer des comptes sur des sites web (reCAPTCHA). Ce faisant, ils corrigent 200 millions de mots par jour et évitent à Google de dépenser environ 146 millions d'euros par an d'après nos calculs (Andro, 2015, 3)
- le recours au travail rémunéré des internautes (ou **crowdsourcing rémunéré**) et qui permet ainsi à un laboratoire de recherche, une start-up ou même un individu de recruter en quelques clics, grâce à des plateformes comme l'Amazon Mechanical Turk Marketplace, des milliers d'internautes pour effectuer des micro-tâches. Ces travailleurs peuvent ainsi travailler où ils le souhaitent, quand ils le souhaitent, pour qui ils le souhaitent, d'être tantôt employeurs tantôt employés. Ces travailleurs peuvent aussi parfois d'être sous-payés en dehors de tout cadre juridique et social.
- Le recours aux ressources financières des internautes qualifié aussi de **crowdfunding**. Grâce à la numérisation à la demande, les bibliothèques peuvent ainsi proposer aux internautes de financer la numérisation de tel ou tel livre qu'elles conservent. En partageant leurs politiques de sélection avec leurs publics, les bibliothèques offrent ainsi des services de numérisation sans avoir à en supporter

> le coût, elles renforcent leurs programmes de numérisation, elles peuvent concentrer leurs efforts financiers sur les documents qui n'intéressent guère les privés mais présentent un intérêt scientifique patrimonial et historique. De leur coté, les usagers peuvent satisfaire leurs propres besoins de numérisation, des fondations peuvent accroître leur visibilité et, pourquoi pas, des investisseurs espérer un retour sur investissement sous la forme de trafic web (Andro, 2014, 1). La numérisation à la demande est généralement couplée avec l'impression à la demande qui permet à des lecteurs de ressusciter des livres numérisés au format imprimé à distance ou sur place avec l'Espresso Book Machine par exemple. (Andro, 2015, 2).

Conclusion

Les amateurs n'ont pas été formés avec les normes, les formats, et les modèles établis du métier de bibliothécaire. Ces amateurs ne cherchent donc pas à les reproduire. Une collaboration avec eux peut donc, non seulement être nécessaire pour réaliser des objectifs impossibles à imaginer sans leur aide ou réduire les coûts, mais aussi, être source de ruptures innovantes.
Mais, pour les institutions culturelles, cette collaboration peut aussi être de cheval de Troie de leur ubérisation, c'est-à-dire d'une forte remise en question du travail, de la médiation et de l'autorité de leurs professionnels. (Andro, 2014, 2)

Bibliographie

Mathieu Andro, Emmanuelle Asselin, Marc Maisonneuve, *Bibliothèques numériques : logiciels et plateformes*, ADBS, 2012.
Mathieu Andro, Pauline Rivière, Anaïs Dupuy-Olivier, Filippo Gropallo, Denis Maingreaud, « Numalire, une expérimentation de numérisation à la demande du patrimoine conservé par les bibliothèques sous la forme de financements participatifs (crowdfunding) », *Bulletin des Bibliothèques de France*, contribution du 2 octobre 2014.
Mathieu Andro, Imad Saleh, « Bibliothèques numériques et crowdsourcing : une synthèse de la littérature académique et professionnelle internationale sur le sujet », *Livre post-numérique : historique, mutations et perspectives*, Actes du 17e colloque international sur le document électronique (CiDE.17), Khaldoun Zreik, Ghislaine Azemard, Stéphane Chaudiron, Gaétan Darquie, 2014, 152 p.

Mathieu Andro, « Bibliothèques numériques : fréquentation et prospective », *Archimag guide pratique*, n° 52, 2015, p. 22-25

Mathieu Andro, Sophie Klopp, « L'impression à la demande et les bibliothèques », *Bulletin des Bibliothèques de France*, contribution du 13 février 2015.

Mathieu Andro, Imad Saleh, « La correction participative de l'OCR par crowdsourcing au profit des bibliothèques numériques », *Bulletin des Bibliothèques de France*, Contribution du 16 juin 2015.

Mathieu Andro, Imad Saleh. « Bibliothèques numériques et gamification : panorama et état de l'art », *I2D - Information, données & documents*, vol. 52, n°4, 2015, p. 70-79.

Mathieu Andro, Imad Saleh, « Le crowdsourcing appliqué aux bibliothèques numériques », *Bibliothèque(s)*, n° 83, mars 2016, p. 23-25.