

Sous la direction de
BERNADETTE DUFRÊNE
MADJID IHADJADENE
DENIS BRUCKMANN

Avec la collaboration de
BENJAMIN BARBIER

Numérisation du patrimoine

*Quelles médiations ? Quels accès ?
Quelles cultures ?*

Préface de Bruno Racine

I

La maîtrise des gisements patrimoniaux : moteurs de recherche et responsabilité des humains

BRUNO ORY-LAVOLLÉE

Étudiant, il m'est arrivé d'aller travailler dans la verdure, dans la bibliothèque des jésuites aux Fontaines, près de Chantilly. Le vieux jésuite bibliothécaire ne laissait pas le lecteur accéder au fichier. Il fallait lui dire ce qu'on cherchait, il vous disait : « il y a tel, tel et tels livres sur ce sujet », et il vous les faisait apporter. À lui seul fichier et moteur de recherche, il maîtrisait ainsi les lectures de la jeunesse.

Les moteurs de recherche ont remplacé le vieux jésuite, et peut-être existe-t-il d'autres analogies entre le rôle de Google aujourd'hui et celui de la compagnie de Jésus dans la formation des esprits à travers les siècles.

Le titre de cette journée, « Numérisation du patrimoine des bibliothèques et moteurs de recherche », recouvrait en fait deux questions : d'une part celle, d'actualité, des relations des bibliothèques avec Google, le premier des moteurs de recherche ; d'autre part celle, plus générale, de la place des moteurs de recherche dans leur stratégie de numérisation. Au préalable, je souhaite rappeler quels enjeux sont en cause.

I. LES ENJEUX DES MOTEURS DE RECHERCHE POUR LA CULTURE

Plusieurs interventions ont montré à quel point, dans un univers où les livres et toutes sortes de documents sont numérisés et en ligne, les moteurs de recherche vont déterminer comment sera diffusé le savoir, quels contenus seront placés en avant ou en retrait, et quelles valeurs ces choix privilégieront.

De plus, comme l'a expliqué Roger Chartier, ils transforment notre manière d'apprendre en substituant à la lecture longue de livres ou de manuels une juxtaposition d'éléments d'origines diverses, évolution que certains voient avec horreur, d'autres avec sérénité.

De ce fait, l'enjeu des moteurs de recherche n'est pas seulement économique, technique, mais aussi culturel et donc politique. La façon dont chaque individu élabore son savoir et sa culture est un enjeu de civilisation et de liberté. Or, de la même façon qu'il n'existe pas de journal objectif, tout moteur présente des biais, inclut des choix et des éliminations qui privilégient certains contenus au détriment des autres. La domination d'un seul moteur, de surcroît mondiale, présente un risque : celui d'occulter une partie de la réalité, sinon d'endoctriner ; cela relève donc de la liberté.

Rappelons que ce n'est pas le seul enjeu de liberté lié aux moteurs. Le respect de la vie privée et le droit à l'oubli sont aussi essentiels, et le sujet va devenir toujours plus brûlant, puisque le volume de données où s'effectue la recherche va continuer de s'accroître, tandis que les moteurs gagneront en puissance et sauront chercher parmi des photos, des sons, des vidéos...

Enfin, les moteurs de recherche représentent un enjeu essentiel pour la diffusion de la culture, mission traditionnelle des bibliothèques. À deux reprises, les rapports que j'ai remis aux ministres de la culture successifs ont insisté sur ce point.

En 2002, celui sur « la numérisation du patrimoine, enjeu de la politique culturelle » constatait que le ministère de la culture et ses établissements avaient déjà numérisé beaucoup d'œuvres et de documents de toute nature, mais sans se soucier assez

de leur diffusion. Plusieurs propositions visaient à l'améliorer, la principale étant de créer le portail www.culture.fr, ce qui a été fait. Mais le rapport insistait aussi sur le référencement, car sans visibilité sur les moteurs de recherche, les contenus ont toute probabilité de rester ignorés, ainsi que sur l'importance de la réutilisation des œuvres numérisées par des éditeurs de sites ou de contenus multimédias.

Cette réutilisation est le thème du deuxième rapport, remis à la fin de 2008. Le terme peut sembler technique, voire rébarbatif, mais l'enjeu est vital. Sur internet, chaque acteur ou presque émet des contenus reprenant ceux qu'il a reçus. Parce qu'elle correspond à la nature même des réseaux, la réutilisation par des tiers – individus, organismes culturels, mais aussi médias et entreprises – du patrimoine numérisé est à mes yeux le principal moyen de le diffuser, de le partager. Par exemple, comme vient de le faire une grande photothèque allemande, rendre visible son fonds sur Wikimedia, branche multimédia de Wikipedia, le 7^e site le plus visité du monde, lui promet visibilité et diffusion. La recommandation principale à tous les acteurs culturels est dès lors très simple : signez le maximum de licences de réutilisation des fonds numérisés que vous détenez, avec des sites internet, des médias en ligne, des blogs, des éditeurs de produits audiovisuels et multimédias ou de services de toute nature, en donnant la priorité aux acteurs disposant d'une grande audience. Et c'est ici que nous arrivons à la première des deux grandes questions que soulève le thème de la journée.

II. QUE DOIVENT FAIRE LES DÉTENTEURS DE FONDS NUMÉRISÉS FACE À GOOGLE ?

L'accord évoqué tout à l'heure entre la bibliothèque municipale de Lyon et Google, ou celui qui a été envisagé avec la BnF, relèvent de cette problématique de la réutilisation : un patrimoine numérique détenu par une institution publique est mis en ligne par un autre acteur.

La première réaction face à ces accords est qu'ils participent à l'objectif de large diffusion de notre patrimoine numérisé,

ce qui justifierait de les encourager. Dans notre rapport, nous invitons les institutions détentrices à négocier avec le réutilisateur la présence sur chaque document d'une étiquette numérique qui identifie l'institution d'origine, et qui doit fonctionner comme un lien vers son site, pour attirer vers elle du trafic, puis des visites.

Le problème vient des exclusivités que demande Google : à Lyon, elles portent sur l'usage commercial des fichiers numériques ; à la BnF, elles devaient concerner les métadonnées. La journée, notamment l'intervention de M. Robert Darnton, a mis en évidence la conséquence potentielle de ces exclusivités : la détention par Google du plus grand fonds numérisé du monde, se chiffrant en millions d'ouvrages, avec un monopole sur son exploitation, ou en tout cas une position dominante.

De plus, l'exclusivité donne à Google le moyen de choisir ce qu'il diffuse ou non, et donc un pouvoir potentiel sur l'accès au savoir ; elle ouvre la possibilité de pratiquer, comme peut le faire un monopole, des tarifs qui excluraient certains de l'accès à ce patrimoine numérisé. Ce qui est à l'opposé de la tradition de la lecture publique.

Un autre enjeu de ces accords dépasse le seul cadre et les seuls intérêts bilatéraux de leurs signataires : l'avenir de la distribution du livre. Le livre électronique est là, et dans deux ou trois ans, il pourrait représenter la moitié de la lecture et du marché du livre. Peut-être ai-je tort, que cela prendra dix ans. Mais la sagesse est de s'y préparer, au cas où j'aurais raison.

En détenant le plus gros fonds de livres numérisés, avec l'exclusivité de leur exploitation commerciale, ou avec l'exclusivité des métadonnées qui aident à trouver ce qu'on cherche, et en possédant le moteur de recherche le plus utilisé, Google a toutes chances de proposer les réponses les plus riches à une requête bibliographique, et pourra presque toujours les assortir de tout ou partie du contenu des livres et documents repérés. Il sera ensuite en position privilégiée pour inciter l'internaute à les acquérir sur son site ou chez ses partenaires. Google aura donc tous les atouts en main pour acquérir une position dominante dans la distribution du livre numérique, et ce au niveau mondial.

Dans la plupart des filières économiques, dominer la distribution permet de capter l'essentiel des profits de la filière. Google sera en mesure d'imposer sa loi à l'édition. L'enjeu des accords entre les bibliothèques et Google est dès lors la survie ou la mort de toute une part du secteur de l'édition, à commencer par les plus petits ou les plus indépendants, qui sont souvent les plus fragiles.

J'aime représenter cette situation en filant la métaphore minière : imaginons des gisements, dont on doit extraire le minerai, pour ensuite trier le métal de la gangue, et enfin le distribuer aux acheteurs. Par ces accords de réutilisation, Google construit un empire qui comprend : la détention de la quasi-totalité ou de la grande majorité des gisements, à savoir les millions d'ouvrages qu'il a numérisés ; l'exclusivité du droit de les exploiter commercialement, ou tout au moins celle de l'outil indispensable à cette exploitation que sont les métadonnées ; une position dominante dans le raffinage, c'est-à-dire dans le tri des contenus recherchés par les internautes, grâce à son moteur de recherche représentant la majorité et parfois la quasi-totalité des requêtes ; et enfin un réseau de distribution, qui vendra les fichiers numériques aux internautes intéressés, avec le projet *Google Books*.

C'est donc la Standard Oil of California qui se reconstruit sous nos yeux. Tout détenteur de fonds se retrouve dans la peau d'indigènes avec une plume sur la tête et du pétrole sous les pieds, mais pas les moyens de construire les derricks. Dans cette situation, les recommandations à émettre me semblent couler de source : l'Indien doit garder la maîtrise de son gisement, donc effectuer lui-même l'investissement de numérisation ; s'il ne le peut pas, il doit refuser les exclusivités, y compris sur les métadonnées ; il doit veiller à ne pas placer l'exploitation du gisement entre les mains d'une seule compagnie, et donc conduire une stratégie qui fera émerger une pluralité de réutilisateurs commerciaux de ses fonds ; il doit conduire des politiques visant à éviter la constitution de monopoles de la distribution ; d'une façon générale, il doit chercher à éviter les situations de monopole à tous les étages de la filière.

III. QUELLE PLACE DONNER AUX MOTEURS DE RECHERCHE DANS LA GESTION DES FONDS NUMÉRISÉS ?

Au-delà du cas Google, la journée a mis en évidence une deuxième grande question : l'attention que doivent porter les détenteurs de fonds numérisés, institutions aussi bien qu'entreprises, à leurs rapports avec les moteurs de recherche, principal vecteur d'accès à ces fonds.

À cet égard, la qualité des métadonnées et celle de l'indexation se révèlent un chantier tout aussi vital que la numérisation, car ces métadonnées déterminent la capacité des moteurs de recherche à extraire ce qu'on voudrait diffuser au public. Objectif essentiel, la visibilité par les moteurs de recherche passe aussi par la traduction, qui garantit d'apparaître quelle que soit la langue dans laquelle est effectuée la requête, ainsi que par le géoréférencement, car beaucoup de requêtes et de services se rattachent à un lieu.

En matière d'indexation, nos bibliothèques possèdent l'atout de leur savoir-faire et de leurs réalisations en matière de catalogage. Selon Patrick Bazin, aujourd'hui directeur de la BPI, ces chantiers, qui sont une des voies de la maîtrise du gisement, représentent de gros investissements, qui peuvent être équivalents, sinon supérieurs, à ceux de la numérisation. L'indexation jouera un rôle déterminant dans la diffusion de la connaissance, dans l'accès au savoir ; elle sera un atout essentiel dans la compétition entre les fonds issus des différentes cultures pour occuper l'espace numérique.

J'évoquais les biais de l'accès à la connaissance résultant de l'usage répété d'un moteur de recherche dominant, d'un même algorithme. On ne peut souhaiter que la concurrence et la pluralité des moteurs de recherche, et nous avons entendu la présentation des particularités du moteur Voila par David Botella. Mais il reste du chemin, car il est un fait que le plus souvent, c'est Google qui donne les meilleurs résultats.

Une autre voie a été évoquée au cours de la journée : les moteurs de recherche, ce n'est pas seulement ceux qui effectuent

une recherche sur l'ensemble du web. Il en existe dans les entreprises, sur les sites, pour des corpus particuliers, pour certains services en ligne. Créer soi-même, sur un sous-ensemble pertinent, des outils de recherche plus raffinés que les grands bulldozers qui labourent la totalité du web, c'est une piste importante. Face à 30 millions de livres numérisés, pourquoi y aurait-il un seul niveau d'agrégation et de recherche ? On peut imaginer, pour des sous-ensembles cohérents, des moteurs plus spécialisés, plus précis. Ils pourraient aider ceux qui ont une plume sur la tête à garder la main sur leurs gisements.

D'une manière générale, cette nécessaire pluralité des moteurs, contribuant à un accès ouvert, multiculturel et démocratique à l'internet, est un thème auquel la recherche, la politique et le débat sur les questions de numérisation devraient consacrer plus d'attention.

IV. CONCLUSION

Deux interrogations formulées au cours de la journée me permettent de conclure.

La première est : qu'est-ce qui est le plus important (en matière de diffusion, de pouvoir dans la filière, de perspectives de développement...) : détenir les fonds numérisés ou les moteurs de recherche ? Jacques Perriault nous a rappelé qu'en matière d'usages, les internautes décident, et donc qu'on ne peut rien prédire à coup sûr. Dès lors, il faut se garder d'effectuer trop tôt des choix irréversibles : il faut donner une attention égale à la numérisation d'une part, à l'indexation et à la visibilité par les moteurs de recherche d'autre part. Aussi, alors que le « grand emprunt » va financer d'importants investissements en numérisation, on doit souhaiter qu'une partie soit affectée à l'indexation, et notamment à des innovations dans ce domaine.

La seconde interrogation est celle de Bruno Racine en introduction de cette journée : qu'est-ce qu'une bibliothèque numérique ? Question abbyssale, mais à l'issue de cette journée, nous savons au moins ceci : cela inclut forcément un dispositif de recherche et d'accès aux contenus disponibles sur le réseau.

Ainsi émergent deux enjeux majeurs : la grande bibliothèque numérique mondiale qui se construit connaîtra-t-elle un moteur de recherche unique ou dominant, ou une pluralité des moteurs ? Et l'accès à ses contenus dépendra-t-il uniquement de moteurs, autrement dit de robots, ou les humains y auront-ils une part ?

Les moteurs font des progrès. Ils parviendront peu à peu à ne pas proposer *Mme Butterfly* après une requête sur les papillons, à citer *Les Misérables* en réponse à une requête sur les égouts de Paris. Parviendront-ils un jour à citer *Art* de Jasmina Reza si on a fait une recherche sur l'amitié ? Il y a un stade où l'extraction du sens d'un livre ou d'un ensemble de textes relève de l'être humain, car elle implique une interprétation, qui est l'essence de l'acte de lecture. Voici que les robots nous en ravissent une partie, il faut se battre pour fixer la limite de leur intervention.

Le contrôle du vieux jésuite sur l'accès au savoir était détestable, mais un monopole des robots, ou pire d'un seul robot, le serait tout autant. L'extraction et l'interprétation du sens doivent rester la responsabilité des humains. Face à ce défi, les bibliothécaires ont un rôle clé à jouer.