

La bibliothèque virtuelle de Google

« *La société Google s'est fixé pour mission d'organiser l'information mondiale.* »

L'annonce faite par Google le 14 décembre 2004 a fait l'effet d'une bombe dans les domaines de l'information et de la documentation. Le géant américain a alors annoncé le lancement d'un vaste projet de numérisation des collections de cinq grandes bibliothèques anglo-saxonnes. Étala sur plusieurs années, le projet Google Print¹, dont le coût minimum est estimé à 150 millions de dollars, vise à numériser un total de 15 millions d'ouvrages imprimés, soit un coût moyen par volume de minimum 10 dollars.

Les partenaires directs de ce projet pharaonique sont quatre bibliothèques universitaires — celles de Harvard, du Michigan, d'Oxford et de Stanford — ainsi que la bibliothèque publique de New York (NYPL). L'université du Michigan permettra à Google de numériser la totalité de ses 7 millions de livres ; Oxford et la NYPL ne proposeront que des ouvrages passés dans le domaine public ; Stanford mettra un quart de son fonds à disposition, soit 2 millions d'ouvrages. Quant à l'université de Harvard, elle n'offrira « que » 40 000 de ses 15 millions de volumes... Il faudra compter 6 ans rien que pour la numérisation des 7 millions de livres de l'université du Michigan. Il faut dire qu'en métrage, cela représente plus ou moins 132 kilomètres linéaires de volumes² !

Le projet prévoit la réalisation de deux copies numériques pour chaque ouvrage : l'une appartiendra à Google et sera indexée dans son moteur de recherche, l'autre appartiendra à la bibliothèque propriétaire de la version papier. Celle-ci pourra ainsi par exemple l'adjoindre à la notice bibliographique correspondante dans son catalogue.

Business is business

Un tel projet n'a évidemment pas été mis sur pied pour le bien de l'humanité, dans le seul but philanthropique d'archiver une importante part du patrimoine culturel du monde occidental³. La société Google vise principalement à engranger de plantureux bénéfices grâce aux publicités qui apparaîtront à l'écran avec les résultats de recherche. Outre cet objectif, tout à fait normal en somme pour une entreprise privée, le géant américain tente par ce nouveau service d'amener chercheurs et internautes à une utilisation (encore) plus accrue de son moteur de recherche. Là aussi, c'est de bonne guerre dirons-nous. C'est tout bénéfice pour l'entreprise et ses actionnaires mais c'est également dangereux pour notre société car dans une situation de (quasi) monopole, individus et besoins particuliers entrent de moins en moins en ligne de compte.

L'un des membres de l'équipe dirigeante de Google, Adam Smith — cela ne s'invente pas —, l'a clairement annoncé : Google veut proposer un maximum de livres dans un maximum de langues. La quantité primera sur la qualité ! Quant à la hiérarchie et le classement des résultats affichés à l'écran, ceux-ci seront déterminés par les mystérieux algorithmes qui font le succès et la fortune de Google. Algorithmes tellement obscurs qu'ils constituent, avec la recette du Coca-Cola, l'un des secrets les mieux gardés au monde...

Outre les cinq grandes bibliothèques partenaires, Google collabore également avec les éditeurs qui le souhaitent. Après tout, c'est bien de leur production, de leur fonds de commerce qu'il s'agit avant toute chose. À côté des publicités, les internautes auront à l'écran un lien vers le site de l'éditeur qui a publié l'ouvrage scanné par Google. Ils pourront ainsi directement acheter en ligne le livre auprès de l'éditeur. Les librairies sur Internet comme Amazon, Dover et Barnes & Noble sont également partie prenante dans ce projet. Ils auraient bien tort de s'en priver. En effet, selon Jean-Noël Jeanneney, « [d]es enquêtes sur les publics virtuels de Gallica⁴ [...] ont montré que beaucoup de consultations conduisent à

¹ La version bêta a été lancée en mai 2005. *Google Print*, <<http://print.google.com/>> [consulté le 1^{er} octobre 2005].

² Chiffres de Peter Stuber, « Le gigantesque projet de bibliothèque de Google » in *BiblioAcid* 2(2005)1, pp. 9-11, <<http://biblioacid.typepad.com/ba/pdf/BAv2n1.pdf>> [consulté le 1^{er} octobre 2005].

³ Vu les bibliothèques impliquées, il est clair que c'est principalement tout un pan de la culture occidentale qui va se voir ainsi diffusé sur la Toile.

⁴ Lancée en 1997, Gallica est le nom de la bibliothèque numérique de la Bibliothèque nationale de France (BnF), <<http://gallica.bnf.fr>> [consulté le 1^{er} octobre 2005].

l'achat de l'ouvrage concerné, neuf ou d'occasion, dans les magasins en ligne ou chez les libraires ou bouquinistes traditionnels »⁵.

Une partie des bénéfices engendrés par les publicités sera rétrocédée aux éditeurs partenaires. À ce jour, rien n'est prévu pour les bibliothèques participantes... En ce qui concerne les publicités, il semble bien que plus il y aura d'argent sur la table, plus les sociétés auront la liberté de choisir avec quels livres et à quelle fréquence leur nom apparaîtra à l'écran. Petit bémol toutefois, l'idée jadis avancée de publicité intelligente, c'est-à-dire une publicité qui coïnciderait avec le produit culturel affiché, semble pour le moment être mise de côté. Ce qui aurait été un moindre mal car le livre est le seul média qui ne comporte pas de publicité extérieure à son domaine.

Technique, limites et limitations techniques

Outre à des moyens financiers colossaux, ce gigantesque projet de numérisation fait également appel à des moyens hautement techniques. Numériser autant de pages en six ans n'est possible qu'avec des appareils à grand débit. Il existe actuellement des robots capables de tourner les pages d'un livre et de les photocopier à un rythme allant de 600 à 1200 pages de l'heure. On sépare le texte de l'image, puis le texte est indexé, ce qui permet à l'utilisateur d'effectuer des recherches sur un mot bien précis. Signalons toutefois que cette indexation ne permettra de trouver le document que via les moteurs de recherche Google (Google, Google Print, Google Scholar⁶...) et non via ses concurrents (Yahoo, Altavista, A9, MSN...).

Ici aussi, c'est la quantité qui prime. Les textes ne sont pas numérisés en haute mais en basse définition : plus rapides à numériser, plus facilement téléchargeables, moins lourds. On peut comparer la qualité de l'image à celle d'une copie de fax.

L'indexation des textes des livres a cependant ses limites. En effet, la recherche est des plus sommaires. À titre d'exemple, en effectuant une recherche au moyen des termes « Candide » et « Voltaire », on trouvera bien sûr l'ouvrage de Voltaire en ligne, mais également tous les livres qui contiennent ces termes dans le corps du texte, en note de bas de page, dans la préface, la bibliographie, dans la liste des autres livres appartenant à la même collection, etc., à savoir un livre sur l'opéra, un autre sur les aspects historiques du catalogage et de la classification, *Moby Dick* de Melville, une étude sur la féminité dans la littérature africaine... Il est également impossible de chercher par date de publication (une recherche sur 1870 donnera tout naturellement les ouvrages où cette date apparaît dans le texte). Bref, il s'agit plutôt d'un immense catalogue de librairie en ligne où il est quasi impossible de trouver un livre précis selon des critères de recherche bibliographique traditionnels (auteur, titre, sujet, date...).

À ces limites qui résultent de choix stratégiques opérés par Google, s'ajoutent deux contraintes qui navrent les internautes :

- o impossible d'enregistrer une copie du livre trouvé ou de parties de celui-ci ;
- o impossible de pouvoir imprimer proprement certaines pages (bien sûr une impression via le navigateur [Fichier > Imprimer] reste toujours possible, mais elle est fastidieuse et la qualité du résultat laisse pour le moins à désirer).

La réponse européenne

Le projet est à peine porté sur les fonds baptismaux que certaines personnalités européennes crient au loup ! L'un d'eux, Jean-Noël Jeanneney, directeur de la Bibliothèque nationale de France (BnF), se fend d'une tribune dans *Le Monde* du 22 janvier 2005 dans laquelle il pointe les dangers qu'implique le projet Google Print. En mai, il publie un pamphlet aux accents gaulliens et au titre sans équivoque : *Quand Google défie l'Europe, plaidoyer pour un sursaut*. Jeanneney met en garde l'Europe, et en particulier les autorités françaises, sur le poids et l'influence qu'un tel projet mené exclusivement par une entreprise privée américaine peut avoir sur la conservation des savoirs, la diffusion des connaissances

⁵ Jean-Noël Jeanneney, *Quand Google défie l'Europe, plaidoyer pour un sursaut*. Mille et une Nuits, 2005, p. 26.

⁶ Google Scholar est un moteur de Google spécialisé pour la recherche d'informations et de documents scientifiques et académiques (périodiques scientifiques, thèses, livres, *preprints*, rapports techniques, comptes rendus, etc.). Pour ce faire, il moissonne les sites des éditeurs scientifiques, des sociétés savantes, des répertoires institutionnels et des universités, <<http://scholar.google.com/>> [consulté le 1^{er} octobre 2005].

et la liberté d'accès à toute culture. Selon lui, cela risquerait en effet de mener à une déformation de notre propre vision du monde :

« Des choix se feront toujours obligatoirement. On peut évaluer à plus d'une centaine de millions d'ouvrages imprimés la production totale de l'humanité depuis Gutenberg (je ne parle que de l'Occident). La quantité promise par Google, si impressionnante en termes absolus, ne correspond qu'à un petit pourcentage de cette immensité. Il faut donc s'interroger sur les livres qui y figureront, sur les critères qui conduiront à en déterminer la liste. »⁷

À cette inquiétude sur le quoi et le combien s'ajoute celle sur l'origine des ouvrages publiés et présents dans les bibliothèques partenaires. Sachant que les traductions réalisées aux États-Unis représentent moins de 3% de l'ensemble de ce qui se publie outre-Atlantique, il est fort à parier que le prisme de lecture américano-américain y constituera la norme. Si l'on ajoute à cela, l'absence d'emprise des citoyens sur la hiérarchisation des résultats de recherche et la profession de foi de la société sur son site (« *Google's mission is to organize the world's information* »⁸), on peut comprendre le malaise du directeur d'une des plus grandes et riches bibliothèques d'Europe.

Dans les premiers mois, la réaction de Jeanneney passe presque inaperçue en France. Puis, en avril, dix-neuf grandes bibliothèques européennes envisagent sérieusement un projet de numérisation en commun. Lors des *Rencontres pour l'Europe de la culture* ces 2 et 3 mai, le Président Chirac réactive le débat en appelant à la création d'une bibliothèque numérique européenne. L'Allemagne, l'Espagne, la Hongrie, l'Italie et la Pologne marquent directement leur accord. Il est inconcevable que le patrimoine mondial culturel (donc aussi européen) soit résumé par et chez Google à la seule culture anglo-saxonne. Pour Chirac, c'est de plus l'occasion de mettre l'accent sur la cohésion et la nécessité d'une Union forte, et ce à quelques semaines à peine d'un référendum qui s'annonçait bien mal...

Peu de temps après, c'est au tour de Viviane Reding, commissaire européenne à la Société de l'information, d'entrer dans la danse. Elle annonce le déblocage de fonds : numérisation et mise en réseau des archives (écrites et audiovisuelles) de bibliothèques européennes (60 millions d'euros), lancement de projets et développement d'un moteur de recherche grand public (36 millions d'euros). En passant à l'offensive, l'Europe vient de se liguer contre l'unilatéralisme googlien, et donc américain⁹... En tout cas, tout le monde s'accorde à dire que l'initiative de Google permet enfin, diront certains, de lancer sérieusement le débat dans toute l'Europe.

Ne nous y trompons pas. Ce « plaidoyer pour un sursaut », cette dénonciation de l'hyperpuissance américaine et de l'ingérence du privé dans la conservation du patrimoine culturel, cette mise en garde contre la publicité écrasante, cet appel à ne pas oublier les peuples à la culture moins diffusée (Afrique, Inde, Chine...), cette invitation à plus de collaboration entre les pays européens, c'est aussi (et surtout ?) une façon pour Jeanneney de réclamer plus de sous. Noblesse oblige, il flotte autour de la BnF une aura de grandeur et de splendeur qu'il lui faut conserver. Et cela ne doit pas être chose facile pour cette institution dont les énormes investissements financiers des dernières années en matière de mobilier et d'immobilier n'ont pas toujours été des plus judicieux.

Face au droit d'auteur

Outre à la réaction européenne, Google doit également faire face à un problème plus délicat, celui du droit d'auteur. Début août, le géant américain a annoncé une pause jusqu'en novembre dans son programme de numérisation des ouvrages soumis au droit d'auteur et décide de ne fournir pour ceux-ci que 3 pages seulement (exception faite des index, bibliographies et tables des matières).

⁷ *Quand Google défie l'Europe, plaidoyer pour un sursaut*, page 11.

⁸ *About Google Print*, <<http://print.google.com/googleprint/about.html>> [consulté le 1^{er} octobre 2005].

⁹ Dans leurs articles sur la question, les médias francophones utilisent par ailleurs dans leurs titres et manchettes des métaphores à caractère guerrier : « Faut-il avoir peur de la bibliothèque virtuelle ? » (Le Vif/L'Express), « La bataille des contenus Internet » (RFI), « Bibliothèque numérique : l'Europe se ligue contre Google » (Le Soleil), « La riposte des bibliothèques européennes à Google » (Le Figaro), « La guerre autour de la première bibliothèque universelle a commencé » (Télérama), « L'Europe se lance dans la guerre de la numérisation » (PC-INpact), « L'Europe vote non à Google : [...] la riposte à Google est devenue le symbole de l'Union européenne sur fond de campagne référendaire. L'appel de vingt bibliothèques nationales européennes emmenées par la BNF est entériné par six chefs d'Etat » (Livres hebdo).

La stratégie de Google est la suivante : laisser le temps aux ayants droits — c'est-à-dire les auteurs et éditeurs — de manifester leur refus d'être intégrés dans le projet. C'est donc à ces derniers qu'il revient de faire connaître leur opposition, et non à la société de partir en quête des éventuels partenaires intéressés par Google Print. Le monde à l'envers ? C'est en tout cas ce que pense l'*Association of American University Presses* qui a clairement fait savoir qu'elle n'acceptait pas cette solution où il revient aux ayants droits de déclarer les titres des ouvrages qu'ils ne désirent pas voir numériser.

Cette méthode peut nous sembler surprenante et quelque peu cavalière, il n'empêche qu'elle correspond à une pratique fort répandue aux États-Unis dans la collecte de données personnelles des internautes (adresse électronique, etc.) : l'*opt-out* (basée sur le consentement implicite). A contrario, l'Europe a plutôt tendance à privilégier la pratique de l'*opt-in* (basée sur le consentement préalable). Cette procédure de l'*opt-out* est en fait tout à fait assimilable à celle que Google pratique quant à l'indexation des sites Web : toutes les pages Internet que ses moissonneuses-indexeuses trouvent sur la Toile sont susceptibles d'être indexées. Si un site souhaite refuser son indexation, partielle ou totale, il lance une procédure technique afin que le moteur ignore les pages en question.

À l'heure actuelle, on trouve donc sur Google Print des livres intégralement numérisés (car tombés dans le domaine public) et des livres dont seules 3 pages sont accessibles (ouvrages toujours sous copyright), cette dernière catégorie étant bien évidemment la plus représentée.

Les e-books

Peut-être l'initiative de Google a-t-elle au moins permis au grand public de réentendre parler des *e-books* ? Le livre électronique n'est en effet pas encore très répandu chez nous. Le concept date pourtant du début des années 70¹⁰ ! Tout a commencé pour lui en 1971, lorsque Michael Hart (University of Illinois) lance le célèbre Projet Gutenberg¹¹ qui a comme but initial la création d'une bibliothèque publique électronique de 10 000 livres. Seize ans plus tard, en 1987, le premier ouvrage disponible sur floppy disk est mis sur le marché, il s'agit du roman *Afternoon* de Michael Joyce. Progressivement, les formats ont évolué et l'*e-book* est devenu un produit de plus en plus fréquent.

Il y a quelques années, beaucoup prédisaient un succès fulgurant au livre électronique, un succès tel que celui-ci aurait littéralement écrasé son pendant papier ! Lors du congrès annuel sur le livre électronique de 1999, tenu au *National Institute of Standards and Technology* (Etats-Unis), Dick Brass, vice-président de Microsoft, prédit la mort prochaine du livre papier : en 2018, 90% de l'ensemble des livres vendus au monde seraient des *e-books* ! Ses propos apocalyptiques ont bien entendu provoqué l'ire des lecteurs cellulosophiles. Les paris sont ouverts !

Si le livre électronique a fait timidement son entrée dans le monde non-anglo-saxon il y a plusieurs années, on ne peut pas dire qu'à l'heure actuelle le succès soit vraiment au rendez-vous. Et pour cause, le livre électronique n'a pas que des qualités. Tout archaïque qu'il soit (?), le format papier reste plus maniable et plus passe-partout. « Un vrai livre de poche, même avec du sable entre les pages, sera toujours plus facile à lire à la plage qu'un écran de téléphone portable »¹². On peut également s'interroger sur la pérennité du support numérique. Celui-ci a-t-il complètement fait ses preuves ? Dans une interview au *Vif/L'Express* du 6 mai 2005, l'essayiste et écrivain Alberto Manguel exprime, non sans humour, son avis sur la question : « Je suis tout aussi dubitatif sur la technologie électronique elle-même. Elle évolue tant qu'elle devient vite obsolète. Au musée de Naples, on peut encore lire sur les cendres d'un papyrus sauvé de Pompéi. En revanche, beaucoup de mes disquettes vieilles de cinq ans sont déjà inutilisables. Les technologies modernes nous condamnent à nous occuper d'elles en permanence. Si nous misons tout sur elles aujourd'hui, qui sait si, demain, les textes ne seront pas perdus ? ». En outre, il faut reconnaître que, par rapport au format papier, relativement peu de titres sont disponibles sur le marché. Existe-t-il par ailleurs des *e-books* sans aucun correspondant papier ? Quant aux technologies de numérisation, elles ne sont pas non plus totalement au point. Pour les livres courants, cela ne pose généralement pas de problème. Par contre, comme le souligne Jean-Michel Salaün, professeur à l'École nationale supérieure des sciences de l'information et des bibliothèques (Villeurbanne), pour les ouvrages plus spécifiques (livres anciens, livres illustrés, livres d'art), la recon-

¹⁰ Voir pour plus d'informations, la brève chronologie de l'*e-book* publiée par *The Guardian* en janvier 2002, <<http://books.guardian.co.uk/ebooks/story/0,11305,627231,00.html>> [consulté le 1^{er} octobre 2005].

¹¹ *Project Gutenberg*, <<http://promo.net/pg/>> [consulté le 1^{er} octobre 2005].

¹² Christine Ferrand dans son éditorial du 6 mai 2005 de *Livres Hebdo* (n° 600).

naissance automatique des formes, des images et des couleurs doit encore s'améliorer. Et sur les livres plus anciens, où l'impression est irrégulière, les taux d'erreurs sont encore importants¹³. De plus, selon une enquête menée à l'université de Galway, cliente du *Safari Books Online*¹⁴, 63,4% des utilisateurs trouvent qu'il est moins facile de lire un livre électronique sur *Safari* que de le lire en version imprimée. Et 51,6% des utilisateurs reconnaissent imprimer une version papier plutôt que de lire à l'écran, contre 32,8% donnant la préférence à la lecture à l'écran et 15,6% sans avis particulier sur la question¹⁵. Ces deux derniers chiffres doivent cependant probablement être relativisés si des coûts sont associés à l'impression des documents par les utilisateurs (par exemple des étudiants). Enfin, le droit d'accès restreint à la consultation, pour les ouvrages sous copyright, constitue lui aussi un obstacle majeur : par exemple un seul téléchargement possible sur un seul poste (impossible d'enregistrer la moindre copie, voire d'imprimer). Et dans le cas d'encyclopédies ou d'autres ouvrages de référence en ligne, gratuitement accessibles si couplés à la version papier, la lecture ou la consultation à l'écran nécessite parfois la présence de la version papier près de soi : le système demande à l'utilisateur de lui fournir la preuve qu'il dispose bien de la version imprimée. Il est dès lors difficilement envisageable d'installer une version électronique sur son portable et d'aller travailler à plus de quelques mètres des 20 volumes papier...

En guise de conclusion

Il serait ridicule de résumer cette affaire aux clichés dichotomiques traditionnels : USA vs Vieille Europe, privé vs public, argent vs culture. Le projet Google Print a l'immense mérite d'avoir secoué le cocotier européen où l'on a souvent tendance à trop tergiverser plutôt que d'innover. Peut-être cela va-t-il servir à mettre un peu d'ordre dans nos pays où l'on a malheureusement encore parfois le mauvais réflexe de trop travailler dans son coin, de numériser de façon éclatée et désordonnée (bibliothèques, communes, centre d'archives, centres de recherches, universités, initiatives personnelles). Quid de la conservation à long terme des fonds numérisés ? L'interopérabilité des différentes bases de données a-t-elle été envisagée ?

L'information n'est d'ailleurs pas la connaissance. Ce n'est donc pas parce qu'une masse d'informations est disponible quelque part qu'elle sera lue, comprise, assimilée et qu'elle rendra plus malin... « Dans les grandes bibliothèques, les livres sont bien plus que la somme de leurs parties » a dit Michael Gorman, président de l'association des bibliothécaires américains.

À l'heure actuelle, Google Print n'est après tout qu'un vaste catalogue de livres (partiellement) numérisés avec des liens vers les éditeurs. Selon moi, ce type de numérisation à grande échelle n'a réellement d'intérêt que dans des cas bien précis, pour des documents spécifiques ou des usages bien particuliers (quand on sait précisément ce que l'on cherche). Et pour des ouvrages d'une certaine ampleur, le confort de la lecture n'est probablement pas prêt à céder le pas au confort d'accès. Les grands gagnants ne sont pas tant les lecteurs, mais plutôt la société Google (qui renforce ainsi sa position sur le marché), les éditeurs participants (qui vendront plus facilement leurs publications) et les cinq grandes bibliothèques qui héritent de leur propre copie pour chacun de leur livre scanné.

Là où le bât blesse, c'est plutôt sur la façon dont Google indexe les livres : il règne une totale opacité sur la hiérarchisation de ses contenus. Moyennant finance, ne serait-il pas tenté de favoriser l'un ou l'autre résultat ? « *Google's mission is to organize the world's information* ». C'est bien cette absence de transparence qui constitue un risque potentiel, bien plus que le « [d]anger que, du côté des ouvrages patrimoniaux tombés dans le domaine public, la liste des priorités favorisent la pesée de la culture anglo-saxonne »¹⁶.

François Renaville

¹³ « Google, la bibliothèque virtuelle et la numérisation ». Interview de Jean-Michel Salaün par Fabrice Impériali in *Journal du CNRS* n° 188, septembre 2005, <<http://www2.cnrs.fr/presse/journal2404.htm>> [consulté le 1^{er} octobre 2005].

¹⁴ *Safari Books Online*, <<http://www.safaribooksonline.com>> [consulté le 1^{er} octobre 2005].

¹⁵ John Cox, « E-Books : Challenges and Opportunities » in *D-Lib Magazine* 10(2004)10, <<http://dlib.org/dlib/october04/cox/10cox.html>> [consulté le 1^{er} octobre 2005].

¹⁶ Jean-Noël Jeanneney, *op. cit.*, p. 12.