« JE COMMENCERAI par une confession. Je ne suis pas informaticien, ni technologue. Je ne suis pas non plus juriste, spécialisé dans la propriété intellectuelle et les subtilités du copyright. Je me considère comme un numéricien par accident, un simple utilisateur d'ordinateur qui a suivi les changements de l'environnement numérique au cours des vingt dernières années. » C'est sur ces mots que s'ouvre le livre de Milad Doueihi qui initie ses lecteurs à *La Grande Conversion numérique.*

Avec déjà un milliard d'usagers, le numérique a une histoire qui se fabrique au jour le jour. Puissance globale qui a métamorphosé tous les systèmes de communication, le numérique fragilise les spécificités nationales et locales, suscitant de nouvelles réalités en politique, dans les médias comme en économie.

Ce livre propose des éclairages précis sur la façon dont une technologie, essentiellement collective, modifie radicalement la vie de chacun, le lien social même, mobilisant nos repères les plus tangibles : écriture et lecture, identité, présence, propriété, archives et mémoire.

Qu'en sera-t-il du savoir historique, de nos bibliothèques, et comment assurer désormais la permanence de nos archives, leur intégrité ?

Ni utopie ni fausse prophétie, le numérique est la vulgate moderne. Avec ses faiblesses, ses aveuglements, ses richesses et ses promesses, le numérique est une culture pour tous.

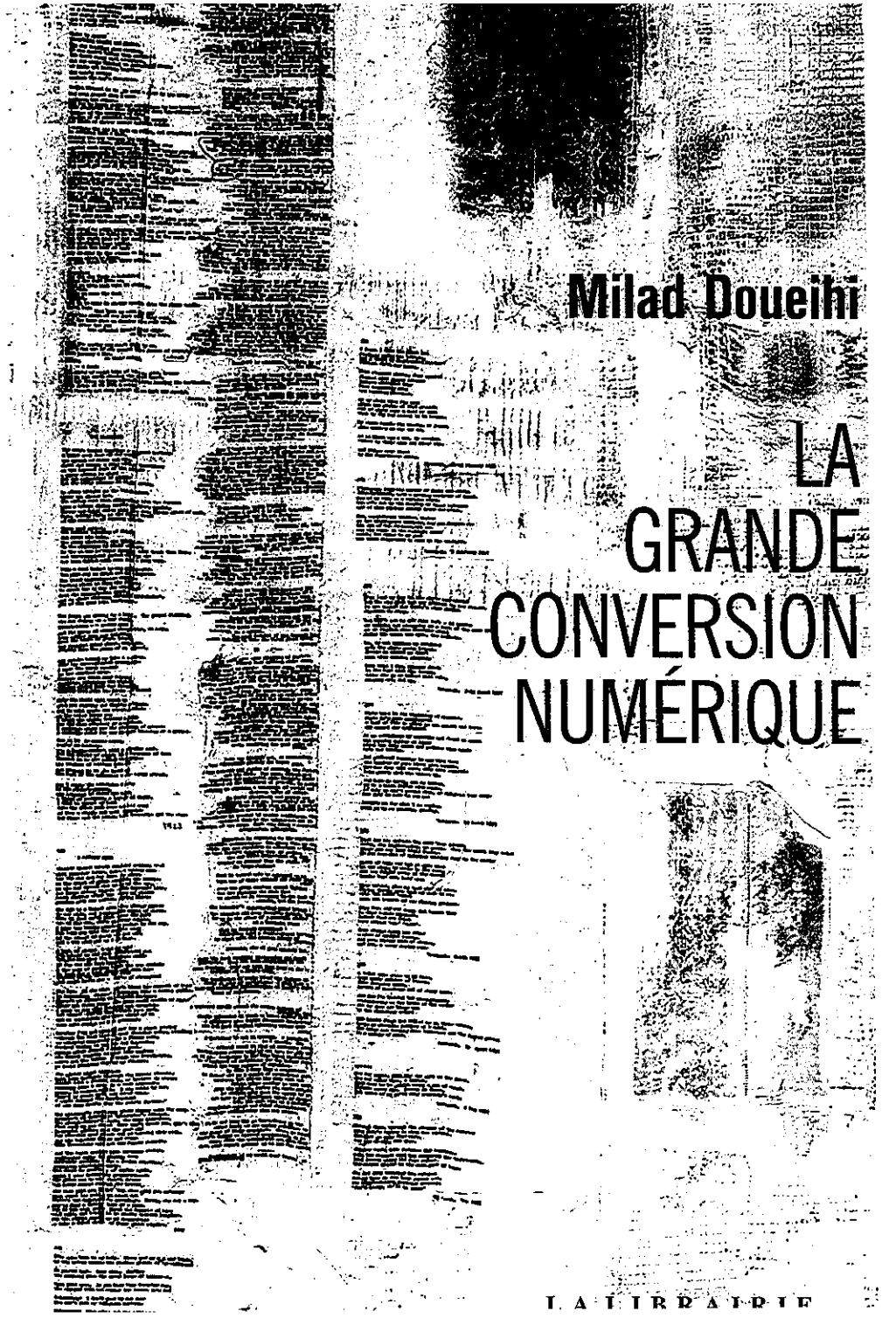
Historien de l'Occident moderne, Milad Doueihi est Fellow à l'université de Glasgow. Il a publié au Seuil, dans la même collection, *Une histoire perverse du cœur humain* (1996) et *Le Paradis terrestre. Mythes et philosophies* (2006).

Traduit de l'anglais par Paul Chemla

wwww.seuil.com

Couverture : Peter Sacks, *Ancestor* (Mandelstam),

Milad Doueihi LA GRANDE CONVERSION NUMÉRIQUE



Milad Doueihi

LA GRANDE CONVERSION NUMÉRIQUE

LA LIBRAIRIE

Archiver l'avenir

Deux articles récents, un peu anecdotiques, constituent peut-être la meilleure introduction au thème de notre dernier chapitre. L'un et l'autre illustrent la fragilité des archives numériques, et les difficultés persistantes que pose à certains grands secteurs d'activité la transition vers ce type d'archivage.

Le premier, une dépêche de l'Associated Press, nous apprend qu'un technicien a effacé définitivement, par erreur, 800 000 images électroniques des archives du Fonds permanent de l'Alaska[1] :

> Peut-être avez-vous déjà ressenti la détresse qui nous envahit quand une seule frappe sur le clavier abolit accidentellement des heures de travail. Imaginez ce que peut faire l'effacement d'un lecteur de disques contenant un compte d'une valeur de 38 milliards de dollars.
> C'est ce qui est arrivé à un informaticien qui reformatait un lecteur de disques au Service des recettes fiscales de l'Alaska. Pendant ce travail de maintenance de routine, il a effacé par accident toute l'information sur les postulants aux bénéfices d'un compte financé par le pétrole – l'un des plus grands avantages sociaux dont jouissent les

1. Il s'agit d'un fonds établi par le gouvernement de l'État d'Alaska, qui reçoit une partie des redevances des compagnies pétrolières actives dans l'État et fait en sorte que tous les habitants de l'Alaska y résidant depuis plus d'un an en profitent effectivement [*NdT*].

résidents de l'Alaska – et il a reformaté par erreur le lecteur de secours aussi. Il restait un dernier espoir, mais le Service des recettes a constaté que sa troisième ligne de défense avait cédé : les bandes de sauvegarde étaient illisibles.
Neuf mois d'informations sur le versement annuel du Fonds permanent de l'Alaska s'étaient évanouis : environ 800 000 images électroniques laborieusement introduites par scannage dans le système des mois plus tôt, les formulaires papier de 2006 que les gens avaient envoyés par la poste ou remplis au guichet pour postuler aux versements du Fonds, et les documents joints pour attester leurs droits, par exemple des certificats de naissance et des preuves de résidence. Il n'y avait qu'un seul secours : ces documents papier eux-mêmes – conservés dans plus de 300 boîtes en carton.
Cette erreur informatique, commise en juillet 2006, coûtera en fin de compte au Service des recettes plus de 200 000 dollars[1].

Trois faits saillants ressortent de cette histoire. D'abord, la facilité avec laquelle les erreurs peuvent tourner à la catastrophe dans l'environnement numérique. Ensuite, la défaillance du matériel et de l'équipement, et surtout le fait qu'il y a une sauvegarde et qu'elle est illisible : il reste une bande de réserve, mais les changements de format et de système l'ont rendue obsolète ; c'est une archive dont l'interface est morte, ou pour laquelle on ne peut trouver aucune interface. Autrement dit, dans l'archivage numérique, il faut non seulement que les objets numériques existent mais aussi qu'ils soient accessibles. Le dernier point est le coût de la renumérisation des documents papier initiaux. Mais, au fait, que se passe-t-il quand il n'y a plus d'archives papier ? Que se passe-t-il quand les archives sont uniquement numériques ?

1. <http://www.msnbc.msn.com/id/17702021>.

Le second article, récemment paru dans *Variety*, traite de la rencontre de Hollywood avec l'archivage numérique. Le problème qu'il analyse est symptomatique des difficultés de l'industrie cinématographique, mais aussi des institutions culturelles, des États et autres entités du même ordre qui ont besoin de stocker à grande échelle et d'archiver des documents multimédias réutilisables. Il résume ainsi le fond de la question :

Pour les films, on croyait que le numérique était comme les diamants, qu'il allait durer toujours. Fini, les couleurs qui passent, les stocks de films qui se détériorent ou prennent feu. Des données numériques pures comme au premier jour, préservées pour l'éternité, et des copies d'exploitation aussi claires et précises que les images prises par la caméra.
Un seul problème : pour la conservation à long terme, il s'avère que le numérique – jusqu'à présent – est une bombe à retardement. À part la peinture de sable[1], il n'existe pas grand-chose de moins durable. C'est tout simple : il n'y a pas de mode de conservation généralement admis permettant de stocker de la « pellicule » numérique plus de quelques mois. Après quoi l'industrie cinématographique utilise un salmigondis de solutions improvisées, certaines assez coûteuses, d'autres pas très fiables.
Le problème paraissait mineur quand le cinéma numérique se limitait aux indépendants à petit budget, aux producteurs de films d'animation et à des pionniers de la technologie comme James Cameron et George Lucas. Mais aujourd'hui, ce problème mineur connaît une croissance exponentielle, puisque les grands studios pas-

1. Allusion, notamment, aux dessins en sables colorés que les Navajo et les Pueblo réalisent pour certaines cérémonies religieuses [*NdT*].

sent à la prise de vues et de son numérique. Des films récents comme *300*, *Apocalypto*, *Zodiac* et *Superman Returns* ont été tournés en numérique. Leur matrice numérique pourrait se dévaloriser très sérieusement si l'on ne s'attaque pas rapidement au problème.
En fait, le problème est si grave que le conseil scientifique et technologique de l'Academy of Motion Picture Arts and Sciences (Académie des arts et sciences cinématographiques) a lancé un cri d'alarme en 2005 : dans quelques années à peine, les films tournés en numérique risquent d'être perdus. Non qu'il n'y ait aucun moyen de conserver les données numériques. C'est le contraire : il existe des dizaines de façons de les conserver, dont la plupart deviennent obsolètes en quelques années. Vous vous souvenez des disquettes 5 pouces et des disquettes zip ? et de celles qui sont encore là ? Pas si fiables.
Les bandes données peuvent cesser brusquement de fonctionner et se désagréger. Une longue histoire nous apprend que les DVD et CD de données se "corrompent" et qu'on ne peut pas compter sur eux pour durer aussi longtemps que leurs cousins sortis de presses commerciales. Sans compter qu'il n'y a aucune raison d'espérer que dans vingt ans – *a fortiori* dans cent – les ordinateurs pourront se connecter aux disques durs d'aujourd'hui[1].

Nous sommes confrontés ici à la vie éphémère des formats numériques, à l'incompatibilité des systèmes et au problème général de l'interopérabilité. La solution, du moins pour l'instant, semble passer par une conversion à rebours, un retour au film à l'ancienne en attendant une réponse technologique dont la plupart des experts reconnaissent qu'elle ne sera jamais pleinement satisfaisante. L'important dans cette situation, c'est que ni le coût, ni le stockage proprement

1. <http://www.variety.com/article/VR1117963533.html>.

dit ne sont en cause. Manifestement, c'est la technologie, la *numéricité* elle-même qui est à l'origine de cette impasse, même quand le film a été tourné en numérique.

Tout utilisateur qui a dû migrer d'un ordinateur à un autre sait sûrement combien il est difficile de préserver les données et de les transférer entre des systèmes. Les changements d'applications et de formats, la croissance exponentielle des données collectées et sauvegardées par les individus et les institutions, le déploiement des données personnelles entre les bureaux des machines et les services hébergés en ligne, sans parler du rythme accéléré de passage à de nouveaux supports de stockage qui rendent les anciens obsolètes, ne sont que de faibles indications du bourbier où se trouve aujourd'hui l'archivage numérique. Nous pouvons en fait affirmer avec certitude que l'un des aspects les plus souvent négligés ou oubliés de la culture numérique est, en dernière analyse, la non-permanence ou la fragilité de l'information et de son support matériel. Les utilisateurs se comportent comme si l'information numérique sera toujours la même, toujours accessible, toujours interrogeable. Un rapide regard sur la brève histoire de la culture numérique suffit pourtant à démontrer l'inconsistance de cette idée. En dépit de leur expérience, la plupart ont une foi utopique dans la capacité des mystérieux pouvoirs de la technologie à convertir et préserver leurs informations essentielles. Comme si la facilité d'utilisation, l'aisance avec laquelle nous pouvons ouvrir de nouveaux comptes en ligne, adopter de nouveaux systèmes et formats, constituait une conversion furtive mais naturelle de notre passé numérique. Mais l'histoire numérique est littéralement jonchée de vestiges de médias et formats anciens et oubliés, de ruines de machines et de systèmes inaccessibles.

Plus nous entrons dans l'environnement numérique, plus nous produisons de données, stockons de documents, échangeons de courriels – et finalement nous devenons des consommateurs de données. L'essor de cette culture de consommation

de données a entraîné une diversification des informations que nous utilisons et collectons, et a accru aussi le volume de celles qui sont collectées à notre sujet. C'est, en un sens, l'une des caractéristiques fondamentales de l'environnement numérique : chaque transaction, chaque recherche, chaque visite sur un site, chaque action liée à la présence en ligne peut être suivie, enregistrée, stockée et devenir finalement un objet accessible par recherche et exploitable. Si cette dimension de l'activité en ligne pose des problèmes de respect de la vie privée et de sécurité, elle incite aussi à se demander qui collecte des données sur nos habitudes numériques, et à quelles fins. Et elle conduit à examiner de plus près les lois qui régissent actuellement ces collectes de données et leurs usages potentiels. Quel doit être le rôle des États, tant pour légiférer sur la conservation des informations que pour protéger la vie privée des citoyens et les prémunir contre une exploitation abusive des données collectées ? Ces données sont devenues d'une richesse et d'une diversité remarquables. Elles englobent non seulement l'information stockée par les FAI, les sites et les fournisseurs du Web marchand, mais aussi par les services de téléphonie mobile, les moteurs de recherche et les États. Elles sont devenues le lieu de production de nouvelles catégories : nouvelles catégories de comportements et nouvelles catégories d'objets numériques, qui, en raison de leur nature, ne peuvent être archivés et conservés qu'à l'aide de concepts propres à leur cas. L'archive numérique est comme la culture numérique : si elle partage avec la culture imprimée certaines caractéristiques essentielles et fondamentales, elle a aussi des traits nouveaux et problématiques qui nous imposent de réévaluer soigneusement nos politiques et stratégies d'archivage, tant individuelles qu'institutionnelles et nationales. Si un trait définit l'histoire des premiers pas de la culture numérique au XX^e siècle, c'est la pauvreté de ses archives. Il est paradoxal qu'une culture ancrée dans l'automatisation de l'enregistrement et du « pistage » de toute

action dans son environnement n'ait pas d'archive exhaustive et fiable sur elle-même. Cet oubli n'est ni innocent ni négligeable par réduction sommaire à un simple accident de l'histoire : il indique un grave aveuglement culturel, et en définitive économique et politique. L'enregistrement des débuts de la culture numérique se distingue par ses lacunes et omissions, car elle a été victime de son propre succès précoce, de son adoption rapide. Si la culture imprimée a engendré une archive historiquement fiable (qui, par comparaison, exigeait peu d'interventions), la culture numérique, fondée sur le changement et les conversions continuelles, a produit, au mieux, une archive fragmentée. Ce type de changement peut être capital, parce qu'il est potentiellement capable, à long terme, d'altérer notre perspective historique, de changer notre façon de définir et de comprendre la notion d'archive, d'archive historique, et d'influencer les récits que nous serons en mesure de produire sur les événements cruciaux de notre histoire culturelle. Jusqu'ici, la dynamique naturelle de la culture numérique n'a pas impulsé une nouvelle dynamique d'enregistrement historique des actes et des objets qui sont la matière première du nouveau savoir-lire.

Le problème des archives numériques ne se réduit évidemment pas aux difficultés de conservation et de transmission du passé récent et de l'état actuel du cyberenvironnement. Il comprend aussi les obstacles auxquels se heurtent les nations, les États, quand ils cherchent à concevoir des méthodes réalistes et pas trop coûteuses de conversion des archives traditionnelles aux formats numériques. Ici, les problèmes sont nombreux, et vont de la sélection des documents à numériser au procédé de numérisation lui-même. Ils portent également sur la mise au point de plates-formes et de mécanismes pour assurer l'accès à ces archives sans compromettre leur intégrité et leur sécurité. De plus, ils nous obligent à réfléchir au statut de l'objet converti ou numérisé : est-ce une simple copie (notamment quand il est disponible dans de multiples ver-

sions et formats) ou un équivalent juridique de l'original ? Doit-il reproduire ce dernier en tout point, ou nous contenterons-nous d'une reproduction acceptable ? Les livres imprimés peuvent exister dans plusieurs éditions et en plusieurs versions. Les versions numériques ont tendance à éliminer ce type d'écarts, elles convertissent l'ancien en nouveau sans nécessairement garder trace de leurs différences.

La technologie numérique est assez puissante pour reproduire une image des textes, par exemple, mais ne peut pas encore reproduire la texture. Il semble donc, pour l'instant, que nous allions vers un univers archival hybride, où l'imprimé et le numérique coexisteront, mais où l'accès au numérique s'élargira tandis que l'accès à l'imprimé deviendra l'apanage d'une petite élite de privilégiés. Cela dit, mon intention essentielle ici est simplement de faire apparaître les conséquences potentielles de l'émergence de l'archivage numérique et certaines lacunes qui lui sont inhérentes. Sera-t-il possible d'avoir un équivalent numérique de nos archives historiques actuelles ? Et ce type d'archives est-il souhaitable, ou devons-nous nous dire qu'avec l'essor de la culture numérique et de son savoir-lire il faut s'attendre à voir l'archivage pertinent prendre de nouvelles formes, qui ne correspondront pas nécessairement à celles de la longue période où nous nous sommes fiés à l'imprimé et à ses prédécesseurs ? Les archives numériques, si notre histoire récente n'est pas un pur mensonge, ont une vie opérationnelle limitée, car les normes, les formats de fichier, les protocoles changent continuellement. Il faudra toujours, non seulement les conserver au sens traditionnel du terme, mais aussi les conserver numériquement, c'est-à-dire les convertir dans des formats plus récents, émergents et, espérons-le, plus fiables. L'interopérabilité et la compatibilité sont donc les caractéristiques de base de tout cadre concevable pour l'archivage numérique, et cette exigence devient un argument fort en faveur des formats ouverts. Si aujourd'hui les débats font rage sur les formats ouverts entre les

entreprises privées, les groupes du logiciel libre et les États, il me paraît essentiel de garder à l'esprit le besoin historique de formats ouverts, non seulement pour les documents, mais aussi pour toute une série d'autres manipulations numériques importantes, de la compression (qui sert à réduire l'espace de stockage) au cryptage* (parfois utilisé pour sécuriser des fichiers ou des archives) et aux formats de bases de données. La conversion en archive numérique est porteuse d'un potentiel prodigieux, mais aussi de problèmes et d'obstacles qui lui sont propres, et nous n'avons plus, je pense, la naïveté de croire à la promesse selon laquelle les nouvelles technologies vont constamment atténuer et simplifier les difficultés passées, à défaut de les résoudre. Si la technologie numérique a prouvé quelque chose, c'est qu'elle évolue en oubliant et même parfois en ignorant son propre passé récent : elle avance en abandonnant ce qui n'a pas fonctionné ou pas plu. Ce principe peut être utile pour réduire l'énormité de l'information produite, mais pas pour maintenir et préserver celle que nous devons ou voulons conserver.

L'objet de ces remarques préliminaires est de souligner le besoin d'une réflexion sur la valeur de l'archivage numérique, et sur les valeurs que les archives numériques incarnent ou engendrent. Nous pouvons déjà voir émerger certaines de ces grandes tendances, des diverses analyses quantitatives de structures spécifiques au réseau à la « longue traîne[1] », pour ne nommer que les plus évidentes[2]. Mais, pour l'essentiel,

1. La « longue traîne » est, dans la courbe des ventes, la longue ligne basse des nombreux articles peu vendus. Alors que, dans l'économie « matérielle », ces articles ne seraient plus produits et deviendraient introuvables, dans l'économie numérique ils restent présents, contribuent à la diversification de l'offre et représentent finalement, si l'on cumule leurs petites ventes, un gros pourcentage des ventes totales (plus de la moitié chez Amazon) [*NdT*].

2. Pour la « longue traîne », voir le livre de Chris Anderson, *La Longue Traîne. La nouvelle économie est là*, trad. fr. de Brigitte Vadé

on peut affirmer sans risque, en ce moment historique, que les aspects globaux, quantitatifs, de la culture numérique ont dominé pour des raisons compréhensibles, avec quelques efforts, réduits mais significatifs, pour étudier les réseaux relationnels émergents et leurs écosystèmes.

1. Histoires du Web : entre l'archive et l'index

Les premières archives de l'environnement numérique sont les index de recherche. Les moteurs de recherche ont été les premiers outils automatisés à exploiter l'ouverture inhérente au réseau, en allant visiter les sites Web publics, en les indexant, puis en gardant en mémoire des clichés de leurs changements et de leur évolution. Ne soyons donc pas surpris si, aujourd'hui, les deux archives les plus exhaustives d'Internet tel que nous le connaissons appartiennent à des entreprises de recherche en ligne. D'abord, la Wayback Machine et son « Internet Archive »[1]. La Wayback Machine rend un service public remarquable en donnant accès à ses archives. Elle collecte des clichés de sites depuis 1996, et ses archives sont faciles à interroger, ce qui permet d'étudier ou de suivre l'évolution d'un site particulier, ou d'observer les effets de technologies nouvelles ou de tendances émergentes sur la

et Michel Le Séac'h, Paris, Village mondial, 2007 (éd. originale, *The Long Tail: Why the Future of Business is Selling Less of More*, New York, Hyperion, 2006). Un premier article a été publié dans *Wired* (octobre 2004) et se trouve en ligne à l'adresse : <http://www.wired.com/wired/archive/12.10/tail.html>. Voir aussi certains textes de Clay Shirky, par exemple « Power Laws, Weblogs, and Inequality » (http://www.shirky.com/writings/powerlaw_weblog.html).

1. <http://www.archive.org/web/web.php>, et, pour plus de détails sur l'archive et son histoire, voir <http://www.archive.org/about/faqs.php #The_Wayback_Machine>. Certains sites produits par des bases de données ne sont pas indexés et archivés. L'index ne note que les pages statiques et ne peut pas enregistrer un contenu produit de façon dynamique.

conception et la fonctionnalité des sites. Mais, comme il s'agit d'un moteur de recherche fondamentalement bien élevé, il respecte les Robots.txt, donc permet aux sites de préciser s'ils acceptent d'être indexés et inclus dans l'Internet Archive ou s'ils préfèrent ne pas y figurer[1]. Entre le choix des sites et les lacunes dans l'archive elle-même, nous avons donc un enregistrement précieux mais au mieux fragmentaire d'Internet depuis 1996. Et comme l'outil utilisé pour visiter les sites est essentiellement un robot de recherche, les résultats varient très sensiblement, de l'analyse de l'ensemble d'un document à l'indexation de balises Méta[2] ou simplement des premiers paragraphes des pages html. Autrement dit, nous pouvons avoir un enregistrement fiable de certaines parties seulement des sites indexés, et la sélection peut varier en fonction de la conception et de la méthode de chaque robot.

En même temps, grâce à la Wayback Machine, on peut écrire des histoires d'Internet et de son évolution, bien qu'écrire une histoire des exclusions et lacunes dans l'Archive soit un choix aussi défendable : puisque nous pouvons voir ce qui manque dans l'Archive, nous pouvons spéculer sur les raisons de ces absences ou omissions et les interpréter en conséquence. Dans les deux cas, nous avons avec la Wayback Machine un premier outil pour *une* archive (il est regrettable, malgré sa valeur, que ce soit la seule accessible) de l'histoire des premiers temps de l'environnement numérique, une archive qui nous en dit autant sur la culture d'Internet que sur la difficulté de créer et de maintenir son archive. D'un point de vue historique, la situation est assez remarquable, puisque, pour la plupart d'entre nous, il s'agit d'événements et d'objets que nous avons contribué à créer ou dont nous avons été

1. Pour plus de détails sur les Robots.txt, voir <http://www.robots txt.org/wc/faq.html>.

2. Il s'agit d'une balise html intégrée dans l'en-tête de la page Web et qui en décrit le contenu [*NdT*].

témoins, mais sans qu'il y ait eu enregistrement complet de l'expérience. La mémoire collective d'Internet et de sa culture contient un nombre exceptionnel de pages blanches[1]. Mais l'archive est dans le domaine public, et contribue donc éminemment à rendre accessible à tous un instantané de l'histoire numérique. De plus, son « libre accès » illustre, malgré la nature partielle de l'archive, l'évolution de la compétence numérique et ses liens étroits avec des technologies et des pratiques, du Web textuel des débuts aux concentrateurs multimédias d'aujourd'hui. Nous pouvons aussi retracer le basculement des pages d'accueil statiques, points d'entrée relativement stables de sites personnels et collectifs, au modèle actuel dynamisé par le RSS et autres paradigmes semblables d'abonnement-transmission. Ces changements sont révélateurs, parce qu'ils montrent comment les pratiques et les technologies convergent pour engendrer des valeurs et des attentes, et créent par là même de nouveaux modèles économiques et forums de sociabilité numérique. L'histoire culturelle de l'environnement numérique reste à écrire, et des archives comme celles qu'a accumulées la Wayback Machine contribueront éminemment à l'entreprise.

La seconde archive, également née de l'indexation d'un

1. Internet Archive, le propriétaire de la Wayback Machine, propose à présent un service d'archivage payant à l'intention des institutions culturelles : « Puisque, de plus en plus, l'information est créée, communiquée et consommée numériquement, les institutions de mémoire comme les archives, les bibliothèques et les musées sont confrontées à la tâche redoutable de préserver des quantités infinies de pages Web. Internet Archive a déjà compris ce défi, et il a sorti aujourd'hui Archive-It 2.0, la dernière version de son service d'archivage sur abonnement. Servant un large éventail d'institutions savantes à un coût sensiblement inférieur à celui d'autres plates-formes d'archives, Archive-It 2.0 permet aux abonnés de collecter, gérer, interroger et conserver des documents en ligne issus tant des sites de leur propre institution que du World Wide Web » (http://www.archive.org/iathreads/post-view.php?id=59978).

moteur de recherche, est celle que possède et maintient Google. On peut raisonnablement supposer que Google, au fil du temps, a construit une archive parallèle d'Internet dans ses grappes de serveurs. L'existence de cette archive est à la fois très précieuse et de nature à inspirer des inquiétudes légitimes, parce qu'elle est propriété privée, mais Google a commencé récemment à l'exploiter de plusieurs façons, qui sont prometteuses. Nous voyons là, à nouveau, le développement d'une stratégie archivale solidement fondée sur les exigences d'une recherche en ligne efficace. Exactitude, vitesse, pertinence et fiabilité sont essentielles au succès d'un moteur de recherche, et essentielles aussi pour une archive. Google a récemment créé son service « Historique Web[1] », à l'intention des utilisateurs qui ont un compte Google[2]. Ce service exige des utilisateurs qu'ils autorisent Google à garder trace de leur activité en ligne. Mais il leur offre une histoire riche et détaillée de leurs recherches, de leur navigation, et des outils leur permettant d'organiser, de visualiser et d'analyser leur

1. C'est le nom de la version française de ce service. La version anglaise s'appelle Web History [*NdT*].

2. « Aujourd'hui, nous avons le plaisir d'annoncer le lancement de l'Historique Web, un nouveau service offert aux utilisateurs de comptes Google qui leur permet de visualiser les pages qu'ils ont visitées et d'y faire facilement des recherches. Si vous vous rappelez avoir vu quelque chose en ligne, vous pourrez le retrouver plus vite avec l'Historique Web, et sur n'importe quel ordinateur. L'Historique Web vous permet de regarder en arrière dans le temps, de revisiter les sites où vous avez navigué, et de chercher dans le texte intégral des pages que vous avez vues. C'est votre tranche du Web au bout de vos doigts. Comment fonctionne l'Historique Web ? Tout ce qu'il faut, c'est un compte Google et la barre d'outils Google avec la fonctionnalité PageRank activée. La barre d'outils, intégrée à votre navigateur, nous aide à associer les pages que vous visitez à votre compte Google. Si vous êtes actuellement un utilisateur de Search History, vous remarquerez que nous avons rebaptisé Search History en Historique Web pour refléter cette nouvelle fonctionnalité » (http://googleblog.blogspot.com/2007/04/your-slice-of-web.html).

présence en ligne enregistrée par Google. L'Historique Web de Google est en fait un index historique, un agrégat de transactions créé en associant les demandes de recherche et les sites Web visités, qui peut produire un récit historique personnel pour l'identité numérique associée au compte Google en question. Dans ce modèle, l'archive n'est plus passive, ce n'est plus un simple dépôt d'informations, un simple rappel des centres d'intérêt passés, elle est activée par son association dynamique à la navigation et à la recherche actuelles ; elle est le site de la mémoire numérique, ou le site des souvenirs d'une mémoire de l'identité numérique. À ce titre, l'Historique Web fournit différents instantanés de l'existence numérique : il permet la reconstruction de centres d'intérêt microcosmiques, et l'analyse de leurs relations tant avec l'histoire personnelle qu'avec des événements et évolutions d'ordre général. Il s'agit assurément d'un outil puissant et éclairant, qui se concentre sur les événements numériques impulsés par une identité (et les utilisateurs peuvent avoir plusieurs identités Google), et qui nous donnera une vision pénétrante des mécanismes de l'identité numérique telle qu'elle se déploie sur le réseau à travers l'acte de chercher et de trouver. Il est clair que, dans cet environnement, émergent des régularités qui reflètent des choix ou des préoccupations personnelles, et que les modifications de ces structures signalent des changements de perspective. Avec un tel outil, nous pouvons imaginer des biographies numériques d'identités écumant le cyberespace en quête de ses trésors cachés, ou simplement lancées dans un périple d'aventure et de découverte. Il est probable que l'outil Historique deviendra très vite le lieu d'une version ouverte de Second Life[1], où les utilisateurs se paieront le luxe de gérer leurs préférences et de déterminer ce qui doit servir à la construction de leur Historique Web. Mais, si Second Life mime le monde réel dans un environnement virtuel,

1. <http://www.secondlife.com>.

l'Historique Web est numérique d'un bout à l'autre : il saisit la dimension polyphonique de l'identité numérique et lui permet de gérer ses propres représentations.

La Wayback Machine et l'Historique Web de Google sont deux outils qui ont émergé de la recherche et abouti à l'archivage de données massives sur l'environnement numérique et ses participants. Tous deux négocient avec la transposition numérique de la temporalité en passant par son expression numérique, et tous deux promettent de nous donner accès à une nouvelle historicité du cyberenvironnement, qui ne se fonde ni sur l'absolu du passé ni sur les privilèges du présent. Chacun à sa façon, ils conçoivent des mécanismes de conversion numérique de l'histoire, pour produire des voix historiques indigènes à l'environnement numérique. Si la Wayback Machine est marquée par des lacunes et des absences, l'Historique Web de Google se définit par une exhaustivité supposée. Tant que l'utilisateur est connecté sur son compte, le système enregistre et suit à la trace ses demandes et les actions qui leur sont associées. Des tendances générales émergent facilement, mais ce qui reste mystérieux, ce sont les intentions, les motivations qui les expliquent, bien que, du point de vue numérique, ce silence ne soit pas nécessairement un problème ni un souci. L'index, cette condensation absolue du contenu de l'environnement numérique, prend dans ce cas une forme personnalisée ; il est individualisé et réduit, pour refléter une par une de multiples trajectoires. Bref, l'Historique Web montre que l'archive numérique est potentiellement capable de produire des outils numériques permettant de réfléchir sur soi-même.

Le problème, bien sûr, est la propriété et le contrôle de cette archive. S'il autorise les utilisateurs à décider ce qui est à conserver dans leur profil, Google garde néanmoins le contrôle total de l'intégralité de l'archive et peut choisir d'en déployer le contenu ou l'analyse comme il voudra. L'archive de Google et l'index qui lui est associé sont la propriété de la compagnie,

ainsi que les enregistrements de leurs utilisations. Il semble qu'actuellement Google détienne certains des souvenirs collectifs et individuels les plus précieux de l'environnement numérique. Et, avec l'introduction des outils bureau de Google (et d'autres sociétés), cette archive peut aujourd'hui accumuler en toute transparence des enregistrements d'activité et d'intérêt numériques qui associent des actions locales et en ligne. La croissance phénoménale de Google atteste l'attrait de son modèle et la pertinence de l'archive comme principe d'organisation de la présence dans l'environnement numérique. De Gmail à la Suite bureautique de Google, la promesse initiale a toujours été la recherche rapide et exacte sur un contenu hébergé et le libre stockage. Autrement dit, la promesse de Google est celle de l'archive numérique elle-même, de l'archive créée par les utilisateurs et leurs actes, de l'archive produite par la diffusion de la compétence numérique. On ne saurait donc s'étonner de voir Google chercher à étendre son modèle fondateur aux archives traditionnelles, celles qui sont normalement hébergées par les bibliothèques sous forme de livres imprimés et de documents papier.

2. « Catalogue de bibliothèque » mondial ou index anthologique ?

Google Recherche de livres[1] a suscité un tel tollé qu'on a du mal à se souvenir avec précision du projet initial, ou de sa mise en œuvre réelle. Les réactions ont été si vives – en particulier celles des éditeurs, de certaines bibliothèques nationales et des rivaux de Google – que l'importance d'une numérisation massive des livres imprimés est passée inaperçue. Dans les remarques qui suivent, je vais surtout m'intéresser à ce projet en tant qu'extension du modèle global de l'index et de l'archive, caractéristique de Google (qui affirme explicitement qu'il s'agit en partie d'un prolongement numérique du catalogue de bibliothèque traditionnel sur fiches en carton, par exemple[1]), et aux problèmes qu'il pose sur le plan de la propriété et du contrôle des documents d'archive, et sur celui de leur accessibilité et de leur diffusion numérique. Avant d'examiner les tentatives des institutions publiques pour élaborer des systèmes d'archivage à l'échelle nationale, et leur affiliation à des projets de « patrimoine numérique », il est instructif d'observer les mécanismes internes du système de Google, parce qu'ils nous permettent de mieux comprendre la conservation et la protection des documents numérisés et le nouveau rôle des bibliothèques et bibliothécaires.

Google Recherche de livres associe la numérisation des ouvrages imprimés et leur indexation : il rend accessibles et interrogeables en ligne des versions numériques des livres, en intégrant leur index à l'index général de Google. Il présente les ouvrages de façon différente selon leur statut au regard des lois sur le copyright. S'ils sont dans le domaine public, on peut les lire en ligne et les télécharger en format PDF. S'ils sont sous copyright, Google communique aux lecteurs une description bibliographique complète du livre et leur donne accès à un nombre de pages limité (qui peut varier selon les souhaits des détenteurs du copyright). Ce dispositif a l'avantage de rendre accessibles en ligne des livres épuisés et du domaine public tout en élargissant l'accès à des publications plus récentes. En un sens, il obéit à une « stratégie Google » cohérente, où l'index de l'archive est l'attrait principal. Vu sous cet angle, ce n'est qu'une expansion naturelle,

1. <http://books.google.com/intl/en/googlebooks/about.html>. Les réactions négatives au projet sont en partie dues à son modèle « ouvert » ; Google et ses partisans répondent qu'il relève de l'« usage loyal ».

1. « Google Books Library Project – an Enhanced Card Catalog of the World's Books », *ibid.*

qui consiste à accroître l'index existant en lui ajoutant des éléments avec, dans certains cas, un accès complet aux documents indexés, puis à rendre accessibles ces contenus *via* les services Google en plein essor. Car, en dernière analyse, la grande différence entre l'offre de Google et les autres, c'est justement la recherche et sa fiabilité. Le programme de numérisation de la BNF, par exemple, et sa bibliothèque numérique Gallica[1] démontrent l'importance de la fonction recherche (et, à cet égard, la mise en ligne récente d'une première version de la composante française d'Europeana montre combien il est difficile de fournir des livres numérisés totalement interrogeables). La plupart des éditeurs ne sont pas bien préparés non plus à la tâche. Seuls des projets de coopération entre éditeurs et compagnies des technologies de l'information semblent en mesure d'apporter une solution, et de créer des concurrents à Google Recherche de livres.

Jusqu'à présent, les adversaires de Recherche de livres ont essentiellement concentré leurs critiques sur deux points : le risque de violation du copyright et la peur d'un monopole Google sur la culture numérique du livre. Pour la violation du copyright, on ne sait pas encore comment les tribunaux trancheront, et comment ils apprécieront le bien-fondé de la thèse de Google sur l'« usage loyal ». Mais, outre l'infraction possible à la loi, la plupart des éditeurs et certaines compagnies des technologies de l'information ont avancé des arguments qui portent sur les méthodes correctes ou acceptables pour gérer le passage aux archives numériques dans le respect de la « créativité » et de la propriété intellectuelle. À leur avis, en décidant d'adopter un modèle d'« inclusion automatique, exclusion sur demande », Google « s'empare » purement et simplement des œuvres et les numérise sans consulter vraiment les éditeurs et surtout sans protéger pleinement les droits des détenteurs du copyright et des créateurs. Ce point de vue a été exprimé très clairement par Thomas C. Rubin, conseiller juridique adjoint de Microsoft sur le copyright, dans un discours prononcé en mars 2007 devant l'assemblée générale de l'Association des éditeurs américains[1] :

> Voici donc la question que nous devons nous poser : quelle voie, en tant que société, allons-nous choisir pour rendre accessibles en ligne les livres et publications du monde entier ? Une voie qui nourrit la créativité et l'innovation à long terme, et qui préserve les incitations pouvant amener les auteurs à proposer en ligne leurs meilleures œuvres ? Ou une voie qui encourage des entreprises à « s'emparer » purement et simplement des œuvres des autres, sans se soucier du copyright ni des retombées de leurs actes sur les auteurs, et aussi sur les éditeurs ? Cette technologie [Turning the Pages, développée conjointement par Microsoft et par la British Library[2]], que la British Library a l'intention d'utiliser uniquement sur des textes qui ne sont plus sous copyright, permet aux habitants du monde entier d'avoir un contact direct avec certains des livres les plus rares et les plus vénérables du monde, qui sont au fondement même de notre histoire et de notre culture.

1. Pour plus de détails sur Gallica, voir <http://www.bnf.fr/pages/zNavigat/frame/infopro.htm?ancre=numerisation/po_chartegallica.htm>, et sur Europeana, voir <http://www.europeana.eu>.

1. <http://www.microsoft.com/presspass/exec/trubin/03-05-07AmericanPublishers.mspx>.

2. Rubin omet de préciser que Turning the Pages fonctionnera uniquement (du moins pour l'instant) sur Internet Explorer et sous Windows : « Turning the Pages 2.0™ permet de tourner "virtuellement" les pages de nos livres les plus précieux. On peut agrandir les détails, lire ou écouter le commentaire d'un expert sur chaque page, et conserver ou partager ses propres notes. Turning the Pages 2.0™ fonctionne avec Internet Explorer sous Windows Vista ou Windows XP SP2 avec .NET Framework version 3, sur une connexion ADSL » (http://www.bl.uk/ttp2/ttp2.html).

> La voie choisie par Google permettra sans aucun doute de rendre davantage de livres interrogeables en ligne plus vite et à moindre coût que d'autres, et, à court terme, ce sera bon pour Google et pour ses utilisateurs. Mais la question est: quel sera le coût à long terme? À mon avis, Google a pris le mauvais chemin pour le long terme, parce qu'il enfreint systématiquement le copyright et prive auteurs et éditeurs d'un canal important pour commercialiser leurs ouvrages. Ce faisant, il fragilise des incitations cruciales à la création. Cela viole le deuxième principe que j'ai mentionné. Et Google a pris ce chemin sans même avoir essayé de parvenir à un accord avec les auteurs et éditeurs touchés avant de commencer à copier. Ce qui viole à la fois les deuxième et troisième principes[1].

Par ces remarques, Rubin exprime une position très répandue dans le métier de l'édition aux États-Unis et en Europe. Le deuxième volet du raisonnement consiste à dire, au fond, que l'échelle et le rayon d'action de Recherche de livres aboutiront à un monopole Google, sentiment que beaucoup partagent, notamment ceux qui sont sensibles aux différences culturelles, et au besoin de tenir compte de la diversité des coutumes et des traditions dans l'élaboration numérique d'une archive qui est fondamentalement culturelle, qui est une construction numérique de l'identité et de l'histoire des cultures. Enfin, il y a une troisième dimension: les efforts technologiques pour créer une expérience fonctionnelle et pratique de la lecture numérique. Nous voyons donc émerger dans cette critique trois problèmes fondamentaux: l'impact que doivent avoir les lois actuelles concernant le copyright et la propriété intellectuelle sur la structure de l'archive numérique naissante; le rôle que doivent jouer les professionnels actuels de l'édition et des métiers du patrimoine culturel pour

1. <http://www.microsoft.com/presspass/exec/trubin/03-05-07AmericanPublishers.mspx>.

prévenir la création d'un monopole, qui peut évoluer non seulement en mainmise absolue sur l'accès, mais aussi en monoculture d'interfaces et de modèles pour la vie en ligne des artefacts culturels numérisés; et enfin les tentatives d'élaborer des technologies capables de transformer l'environnement numérique en environnement de lecture pour des objets comme les livres.

Mais Microsoft et les éditeurs américains ne sont pas les seuls à refuser la démarche de Google Recherche de livres. Nous trouvons une critique différente dans *République 2.0*, le rapport rédigé récemment par Michel Rocard. Le passage vaut d'être cité *in extenso* parce qu'il exprime clairement la dimension culturelle et politique du débat, dans des termes qui associent l'identité culturelle à la conception et au fonctionnement de l'archive numérique. Il s'agit, en l'occurrence, d'opposer le modèle Google, non spécifié mais que l'on suppose américain, à une alternative européenne que préconise l'auteur:

> Peu de Français se satisfont de la domination culturelle américaine et la plupart des Européens devraient être solidaires du projet de Bibliothèque numérique européenne qui est une juste réponse face au projet «Google Book Search[1]». Voilà pourquoi le projet de Bibliothèque numérique européenne, ou Europeana, mérite tout notre soutien. Face à Google, la réponse ne peut être qu'européenne. Or, le moteur de recherche européen Quaero est en difficulté depuis le retrait de la partie allemande et le système européen de navigation par satellite Galileo n'est toujours pas lancé, quand la plupart de nos voitures européennes sont déjà équipées de GPS. Il ne faut pas que le projet de Bibliothèque numérique européenne connaisse le même sort, surtout que pour l'heure seuls la Hongrie et

1. Nom de la version originale anglaise de Google Recherche de livres [*NdT*].

> le Portugal se sont associés au projet français. C'est un aveu sinon d'échec, du moins d'impuissance, de notre Bibliothèque numérique « européenne ». Europeana ne doit pas être seulement une bonne idée, bien française, idéaliste et finalement irréaliste car ne s'appuyant pas sur un modèle viable. Au lieu d'affronter Google avec la seule politique publique, peut-être le ministre de la Culture aurait-il pu imaginer une alliance avec d'autres partenaires de premier plan, pour concurrencer Google Book Search, plutôt que de confier la seule résistance à notre Bibliothèque nationale de France ? Peut-être aurait-il dû négocier un projet véritablement européen au lieu de laisser la France en être seule le moteur ?
>
> Nous ne devons pas laisser à Google le monopole de la numérisation de notre mémoire et de notre histoire. Il est important aussi de garder à l'esprit toutes les questions d'interopérabilité ou les problèmes liés à l'internet et aux libertés que peut faire peser Google s'il est seul à gérer la numérisation des livres[1].

Il semble que la solution proposée ou du moins suggérée par Michel Rocard passe d'abord par l'expansion d'Europeana pour en faire un projet pleinement européen, puis par l'alliance avec des sociétés privées du secteur des technologies de l'information afin de contrer le poids dominant de Google et l'échelle de son entreprise. Mais, du moins pour l'instant, le seul allié réaliste possible est Microsoft, et, malheureusement, cette alliance-là (on en a eu la démonstration dans de nombreux projets, dont Turning the Pages) contredirait aussi la dernière remarque de la citation. Microsoft ne veut, semble-t-il, qu'apporter des technologies, des interfaces et des outils qui s'appuient sur son système d'exploitation et sur son navigateur : il a un long passé d'exploitation de sa situation privi-

1. *République 2.0, op. cit.*, p. 21.

légiée en qualité de fournisseur de la majorité des systèmes de bureau, dont il s'est servi comme levier pour promouvoir son modèle. Tandis que le projet de Google, malgré ses visées universelles et ses menaces potentielles d'homogénéisation, communique son archive dans un format ouvert, qui n'exclut aucun système d'exploitation ni navigateur. Contrairement à Microsoft, son interface est une interface Web qui est devenue synonyme d'interopérabilité et de compatibilité avec la plupart des systèmes et navigateurs. Donc, s'il est effectivement nécessaire de faire contrepoids à la domination de Google, il faudra aussi tenir compte d'une autre nécessité : donner accès de façon ouverte et standard. Car le modèle du libre accès, dans ce cas, ne se réduit pas au simple accès au texte numérisé. Il implique aussi la possibilité d'interroger, télécharger, imprimer et manipuler les documents du domaine public.

Mais quels sont les projets de Google avec Recherche de livres ? Un article du *Times* de Londres éclaire un peu ce qui va suivre et les conséquences potentielles de l'opération sur la numérisation et sa relation avec l'imprimé : « Avec les 380 millions de personnes qui utilisent Google chaque mois, cette initiative va beaucoup stimuler le développement des livres électroniques, et elle aura un gros impact sur le secteur de l'édition et les libraires. Jens Redmer, directeur de Google Book Search en Europe, nous a confié : "Nous travaillons à une plate-forme qui permettra aux éditeurs de donner aux lecteurs totalement accès à un livre en ligne." Selon lui, le téléchargement de livres ne signifie pas la fin de l'imprimé, cela donne seulement plus d'options d'achat aux lecteurs. "Ils veulent peut-être louer un guide de voyage seulement pour les vacances, ou acheter un seul chapitre d'un livre. En définitive, ce sont les lecteurs qui diront comment les livres seront lus", a-t-il conclu[1]. » Ce qui apparaît clairement dans cet

1. Pour l'article, voir <http://business.timesonline.co.uk/tol/business/article1294870.ece>. Et pour un commentaire perspicace sur

article, c'est que Google cherche à étendre l'environnement numérique, à travers son interface à base d'index et d'archive, pour y inclure finalement les livres électroniques. C'est sous cet angle-là qu'il faut voir et interpréter ses objectifs, et pas seulement du point de vue de possibles violations du copyright, abus de la clause d'« usage loyal » ou ingérences dans les affaires des éditeurs traditionnels. La numérisation massive, couplée à une indexation puissante et à une « livraison » simple mais efficace des objets numériques, s'inscrit dans un effort concerté pour redéfinir radicalement et ressusciter le paysage exsangue des livres numériques. Google démontre ainsi la viabilité commerciale de la dimension anthologique de la compétence numérique, conformément à des modèles assez familiers et désormais bien établis (l'iTunes Store avec publicité AdSense). Les enjeux financiers sont énormes, mais, ce qui nous intéresse ici, ce sont les composantes et les conséquences culturelles de ce vaste effort d'archivage en ligne de livres numérisés, car c'est la lecture elle-même qui en sera transformée, transfigurée. Les lecteurs pourront louer ou acheter des chapitres ou des parties d'un livre, tout à fait comme on achète une chanson seulement sur iTunes. Mais, dans le cas du texte, la recherche est plus puissante et détaillée, et permet de le fragmenter et subdiviser à des niveaux inférieurs aux chapitres ou aux parties. On peut donc imaginer qu'un utilisateur achète quelques pages, ou loue une combinaison de chapitres. (Les opérations de ce genre exigeront, évidemment, une gestion des droits complexe pour les livres sous copyright.) Pour Google, le bénéfice est énorme : tout cela accroît le trafic vers ses sites, la taille et la pertinence de son index, et transforme l'activité de lecture, au moins pour une importante catégorie de livres, en résultat d'une recherche et d'une indexation.

---

les projets d'avenir de Google, voir Nate Anderson, « Google Book Search : Buy Your Books by the Chapter » (http://arstechnica.com/news.ars/post/20070123-8685.html).

Bref, cette pratique va redéfinir sur des points importants le livre numérique, et avec lui la lecture numérique et le savoir-lire dans l'environnement numérique. Dans ce cas au moins, l'archive en ligne est l'agent d'une conversion radicale : celle du livre imprimé en un objet numérique dont l'interface première est la recherche et l'indexation. Le livre passe ainsi de l'état de composition relativement autonome et cohérente à celui de structure tabulaire et indexicale. Et le chercheur / lecteur potentiel devient tout-puissant : il peut composer sa propre sélection à partir de divers livres et travaux, et créer ainsi l'espace d'un environnement anthologique personnel ultime qui est reproductible et échangeable en ligne. Les livres numériques et leur indexation vont ressusciter les anthologies personnelles de la Renaissance et des Temps modernes. Comment les auteurs et les éditeurs réagiront-ils à cette mutation potentielle du livre traditionnel ? Les auteurs se mettront-ils à écrire avec cette lecture à l'esprit ? Écriront-ils pour l'index et l'archive dans leur incarnation numérique ? Les livres ne s'y prêtent pas tous, évidemment, ils ne se laissent pas tous « anthologiser » facilement, mais cette possibilité pourrait réorienter sur des voies surprenantes la production des textes à lire. Que sera la création littéraire, en tant que pratique numérique ?

Dans tous les moteurs de recherche, la matérialité de l'index se réduit à son interface iconique, qui détermine l'accès au contenu et la forme de sa réception. Google Recherche de livres – et c'est peut-être l'un de ses traits les plus radicaux – parvient à effacer, ou du moins à faire oublier à la plupart d'entre nous, la différence entre numérique et numérisé, entre les objets qui, lorsqu'on les transfère dans l'environnement numérique, sont réellement transformés en autre chose, et ceux dont la construction et la conception doivent leur origine et une partie de leur intelligibilité à la matérialité de l'imprimé et de sa culture. C'est peut-être dans ce contexte que nous devons penser une critique souvent faite à Recherche

de livres : son impact sur la « créativité ». Dans la plupart des cas, « créativité » est un euphémisme au service d'une défense conservatrice du copyright, dont on veut étendre les dispositions protectrices à l'environnement numérique en se souciant fort peu, ou pas du tout, de ses spécificités. L'expression accompagne donc souvent des appels à « gérer les droits » ou à contrôler l'accès au contenu, qui vont à l'encontre de libertés fondamentales inhérentes à la culture numérique. Mais il faut voir aussi que la compétence numérique, en s'étendant, en numérisant le patrimoine de la culture imprimée avec la mémoire et l'histoire qui lui sont associées, acquiert la capacité potentielle d'engendrer de nouvelles formes de lecture qui, bien qu'elles nous rappellent des pratiques antérieures, vont aussi, probablement, susciter de nouvelles formes de créativité aux dépens de certaines des anciennes. Le livre en tant qu'objet ne risque guère de disparaître dans un avenir prévisible. Mais il est clair aussi qu'il n'est plus le premier ni le seul objet adapté à la production du savoir, à son échange et à sa transmission. Et c'est cette transformation qui crée le besoin et inspire l'effort d'une réflexion de fond sur l'émergence des archives numériques.

3. Le patrimoine et les nouveaux objets numériques

Google Recherche de livres ouvre la voie à une modification potentiellement radicale du paysage culturel grâce à la numérisation et à l'indexation massives des archives imprimées, avec tout ce que cela implique pour la compétence numérique. Mais les systèmes d'archives numériques en sont encore à leurs débuts : il en existe très peu – aucun, peut-être – à l'échelle d'un pays. Dans la plupart des pays développés, des études ont défini les conditions technologiques, politiques, financières et culturelles à réunir pour créer et mettre en service ces archives. De nombreux projets sont aujourd'hui en

cours, du Royaume-Uni à Taiwan[1]. Les pays et les organisations internationales reconnaissent l'importance des archives numériques et la nécessité de bien penser leur mise en œuvre et leur financement. Pour ce qui nous intéresse ici, j'analyserai essentiellement les plans des États-Unis, exposés dans une série de rapports rédigés pour la Bibliothèque du Congrès et sa Digital Preservation Initiative (Initiative de conservation numérique). Ces plans s'inscrivent aussi dans le cadre des Bibliothèques numériques nationales des États-Unis[2].

La plupart des études des futures archives numériques ont en commun une série de préoccupations et de problèmes : faire la distinction entre documents numériques et documents à numériser, avec les spécificités de chaque catégorie ; disposer d'une infrastructure capable de gérer le déploiement à grande échelle d'archives numériques nationales, leur conception et leur financement ; prendre en compte les pratiques numériques et la nature très particulière des objets numériques actuels et émergents, parce qu'ils constitueront le noyau des archives ; mettre en œuvre des dispositifs de sécurité pour préserver l'intégrité des archives et gérer l'accès quand il faut protéger des droits ; concevoir des interfaces adaptées à ces archives numériques et aux usages et manipulations prévisibles de leurs utilisateurs. Dans toutes les études, le point de départ est aussi assez semblable : entamer la transition des archives et méthodes de conservation actuelles à leurs homo-

1. Il y a aussi un rapport français, *La Diffusion numérique du patrimoine, dimension de la politique culturelle*, à l'adresse <http://www.culture.gouv.fr/culture/actualites/rapports/ory-lavollee/ory-lavollee.pdf>. Le NDAD britannique (National Digital Archive of Datasets) donne accès aux ensembles de données et aux documents de l'État (http://www.ndad.nationalarchives.gov.uk). Les archives numériques de Taiwan sont accessibles à l'adresse <http://www.ndap.org.tw/index_en.php>. Ce ne sont que deux exemples parmi bien d'autres.

2. Voir <http://www.digitalpreservation.gov> pour toute information sur le projet et ses plans.

logues numériques, et entreprendre des consultations pour élaborer des méthodologies et parvenir à un consensus sur la sélection des documents à numériser. Nous avons donc des plans nationaux de patrimoine numérique, conscients de la nature hybride du patrimoine actuel et très attentifs aux risques comme aux promesses de la culture numérique. Le premier pas en direction de ces archives de type nouveau consiste à identifier les documents culturels historiquement « les plus significatifs » – c'est l'expression qui a été retenue –, indépendamment de leur support matériel et de leur forme. Viennent ensuite la collecte, la numérisation et l'organisation des objets numérisés pour les rendre, finalement, accessibles au public. La tâche ici (qui peut être longue et complexe) consiste d'abord à se mettre d'accord sur une sélection large et représentative de documents qui n'exclue ni un groupe, ni un point de vue, ni une période historique. On suppose en effet que, lorsque les archives numériques seront opérationnelles, la majorité des utilisateurs vont s'en servir à la place des originaux; il est donc crucial d'avoir une sélection solide et viable de documents numérisés si l'on ne veut pas changer radicalement le paysage culturel. De plus, le processus de sélection postule – à juste titre, je pense – que tout ne peut pas être et ne sera pas numérisé. On reconnaît donc que la coexistence va se poursuivre, que l'archive restera hybride. Le rêve numérique de bibliothèque absolue où tout objet imprimé sera accessible n'est pas réaliste. Il y aura probablement des bibliothèques et des archives divisées, avec des préoccupations et des centres d'intérêt différents, mais – sauf peut-être dans quelques-unes, très peu nombreuses – la majorité des accès et consultations finiront par avoir lieu dans l'environnement numérique. Même quand la numérisation croissante, voire excessive, renforcera l'attrait « exotique » des archives à l'ancienne.

Les imprimés et les autres objets non numériques sélectionnés pour être inclus dans les archives numériques naissantes

posent une série de problèmes difficiles, qui sont fonction de la nature de chacun d'eux, de sa forme ou présentation et de sa matérialité. Ces documents vont des manuscrits et des livres rares aux premières photographies et aux premiers enregistrements sonores. S'il est facile de les convertir en objets numériques, il ne l'est pas toujours autant d'assurer que les objets numérisés soient équivalents à l'original. Les notes en bas de page, par exemple, sont beaucoup plus faciles à gérer dans l'imprimé que dans le numérique, et il faudra en tenir compte quand on décidera du mode de conservation numérique d'un livre où elles sont longues et nombreuses, ou d'un ouvrage où ces notes et leurs structures complexes jouent un rôle important[1]. Mais la matérialité des objets imprimés impose aussi aux bibliothécaires et aux conservateurs des choix fondamentaux: doivent-ils se contenter de prolonger la pratique actuelle qui consiste à protéger un objet rare ou fragile en le remplaçant pour les consultations par une copie fac-similé, ou mettre à profit la technologie numérique et ses outils en associant à l'objet numérisé une information supplémentaire, donc en orientant l'accès à cet objet et sa réception par le choix de ces outils et informations ajoutés? C'est une question grave, car cette pratique peut revenir, selon les choix opérés et leur généralisation, à imposer indirectement une certaine lecture et interprétation des documents historiques et des textes fondateurs. Il n'y a pas d'outils innocents ou neutres, notamment quand ils structurent et organisent directement le mode de lecture de textes et d'images. Ce qu'il faut bien voir en l'occurrence, c'est que le contexte joue un rôle déterminant dans la formation des opinions et des idées, et que l'environnement numérique est une construction contextuelle, fondée sur la compétence et sur des pratiques numériques parfois assez séduisantes pour être retenues, sans

1. C'est le cas, par exemple, du *Dictionnaire historique et critique* de Pierre Bayle.

véritable réflexion, comme paradigmatiques pour la présentation du matériel numérisé. Les bibliothécaires sont conscients de ces potentialités de la technologie numérique[1], mais, sur les questions « que numériser et comment ? », la décision ne devrait pas être du seul ressort des bibliothèques et de leur personnel. La technologie numérique transforme la bibliothèque elle-même en un espace nouveau : d'abord, l'environnement numérique étant par nature un réseau, elle la mue en site d'accès à d'autres collections en ligne et pas seulement aux collections locales ; deuxièmement, elle redéfinit sa mission de conservation ; troisièmement, elle change la nature de l'index ou du catalogue de la bibliothèque, en introduisant des possibilités de recherche d'un autre type. L'organisation du catalogue par titres, sujets ou auteurs, essentielle à la longue histoire de la bibliothèque en tant que lieu d'hébergement et de conservation des catégories du savoir et de leurs représentations culturelles, doit s'adapter à l'index numérique,

1. « Un problème qui ne peut pas être correctement traité ici fait l'objet de discussions constantes à la Bibliothèque : la possibilité d'utiliser des versions numériques comme copies de sauvegarde. Traditionnellement, la conservation du contenu privilégie la création d'un fac-similé, d'une copie aussi fidèle à l'original que possible, sur un support de longue durée. La méthode la plus largement admise pour conserver l'information de documents textuels est le microfilm, et pour les documents illustrés c'est la reproduction photographique. Un aspect du débat porte sur une autre question : quand convient-il de produire une version numérique qui s'efforce d'être une copie fidèle d'un document, et quand faut-il tirer profit du potentiel d'enrichissement de l'accès au contenu ? Faut-il améliorer la lisibilité d'une page de manuscrit en ajustant le contraste ? Pour un tirage photographique, la copie fidèle est peut-être le choix le plus adapté. Toutefois, si la Bibliothèque possède le négatif mais aucun tirage, est-il convenable de faire des ajustements numériques pour présenter une image plus forte ? Peut-on élaborer des principes généraux pour guider ce type de décisions ? » (*Historical Collections for the National Digital Library*, <http://www.dlib.org/dlib/april96/loc/04c-arms.html>).

avec sa conceptualisation de l'objet et les notions qu'il associe au contenu de celui-ci.

Pour trouver des livres dans une bibliothèque, la structure imposée des représentations et ses catégories générales sur la production du savoir facilitent les choses. Trouver de l'information en ligne est aussi une activité régie par des représentations et des concepts, mais ils sont différents. Ils sont actuellement produits par un effort collectif qui associe à l'analyse du contenu des métadonnées : la représentation numérique d'un objet, de son ou ses auteurs, de son sujet, de son lieu de publication, etc. Ces données sur les données sont essentielles, et elles vont jouer un rôle de plus en plus important dans les systèmes automatisés et d'autodécouverte ; mais donner un nom aux conventions et aux standards est capital si nous voulons une archive numérique interopérable. L'identification des objets numériques par des marqueurs uniques et permanents (comme l'identificateur d'objet numérique, DOI[1]) lèvera en partie la difficulté de trouver les objets numérisés. Dans le cas des archives numériques, les métadonnées sont à la fois un moyen standard de marquer et d'identifier les objets, et un ensemble d'instruments de gestion qui permettent de séparer les collections des outils d'accès et de leurs interfaces. Autrement dit, les métadonnées serviront de pont

1. *Digital Object Identifier* : « Le système DOI sert à identifier les objets-contenus dans l'environnement numérique. Des noms DOI® sont assignés à toutes les entités pour utilisation sur les réseaux numériques. On les emploie pour donner des informations actualisées, dont l'endroit où l'on peut trouver ces objets sur Internet (ou s'informer à leur sujet). L'information relative à un objet numérique peut changer au fil du temps, y compris sa localisation, mais son nom DOI ne changera pas. Le système DOI offre un cadre d'identification durable pour gérer le contenu intellectuel, les métadonnées, assurer le lien entre clients et fournisseurs de contenu, faciliter le commerce électronique et permettre la gestion automatisée des médias. Les noms DOI peuvent servir à toutes les formes de gestion de toutes les données, commerciales ou non » (http://www.doi.org).

entre les systèmes automatisés et leurs interfaces vers l'archive et son catalogue, ou entre les utilisateurs et les outils qui leur permettront de chercher des objets, de les identifier et d'y avoir accès. Certaines réalisations illustrent déjà des applications raffinées, et prometteuses, de la technologie aux archives numériques scientifiques. Un chercheur peut générer un abonnement RSS qui l'informera de toute nouvelle publication « taguée » par les mots clés qui l'intéressent, en liant le nouvel objet à ceux qui lui sont proches parmi les documents déjà présents dans l'archive ou identifiés par le chercheur. Des outils comme ceux-ci, qui combinent des technologies issues de divers horizons de l'environnement numérique, enrichissent l'expérience de la recherche en ligne ; ils permettent d'identifier facilement un nouveau document et de l'associer tant à un contexte défini par l'utilisateur qu'aux métadonnées de l'index général.

Mais que dire des données et objets « numériques de naissance » ? Comment les capturer, les conserver et les rendre accessibles ? Les données numériques, nous l'avons dit, sont à la fois fragiles et toujours changeantes. Selon la Bibliothèque du Congrès, 40 % des sites présents sur Internet en 1998 avaient disparu un an plus tard[1]. La production numérique de l'information n'égale pas la relative permanence des imprimés. Mais un autre facteur est tout aussi important : l'échelle de l'information numérique est phénoménale. La mise en œuvre d'archives numériques durables, fiables et capables d'en enregistrer et maintenir ne serait-ce qu'un petit pourcentage exige une nouvelle infrastructure numérique, mieux faite pour supporter le stockage massif et les innombrables demandes d'accès. Par leurs structures et leur architecture, les bibliothèques d'aujourd'hui sont monumentales, et nous avons besoin de nouveaux monuments numériques qui soient, dans notre cyberenvironnement, aussi robustes et durables

1. <http://www.digitalpreservation.gov/importance>.

qu'elles. Les bibliothèques sont construites et conçues comme des voies d'accès au savoir historique ; ce sont des lieux de découverte et d'étude. Les nouvelles archives numériques devront remplir les mêmes fonctions, mais il leur faudra aussi intégrer les nouveaux objets numériques. Certains d'entre eux sont générés dynamiquement : ils n'existent pas en tant que tels, mais sont assemblés à la suite d'une demande venue d'un utilisateur. Comment enregistrer et archiver un tel objet ? Comment décider s'il mérite ou non d'être inclus dans une archive numérique ? Et qui doit prendre cette décision ? La plupart des sites d'achat actuels sont générés dynamiquement. Certains objets numériques sont très éphémères, leur conservation exige donc de nouveaux protocoles de capture. Leur maintien, leur nomination et leur authentification posent de gros problèmes. Comment les convertir pour qu'ils restent accessibles malgré la rapidité des changements de systèmes et de formats ? Comment élaborer un mécanisme de nomination qui gardera sa cohérence à travers archives et réseaux ? Enfin, comment authentifier ces objets ? Pour des archives numériques, les problèmes de sécurité sont essentiels, en raison de la nature même de l'objet numérique. Si l'intégrité des objets n'est pas maintenue et authentifiée, si les lieux de stockage sont attaqués par des pirates, il peut y avoir des modifications et des altérations de document. Dans ce cas – et nous avons vu suffisamment de brèches de sécurité, et de données modifiées et perdues, pour comprendre l'importance de stockages sûrs, qui serviront de modèles de référence pour les archives numériques accessibles –, c'est la légitimité de l'archive qui deviendra suspecte. Puisque les archives numériques sont destinées à devenir, non seulement la mémoire de notre culture numérique, mais aussi le lieu de conservation d'une large part de nos identités historiques et culturelles, il est crucial de concevoir et d'assurer leur protection.

L'environnement numérique promet l'accès universel, mais il accroît aussi les risques et modifie radicalement certains

éléments de base de nos activités culturelles. Il est en train de convertir (au double sens technique et religieux) nos espaces culturels à ses propres concepts et catégories, et, ce faisant, il rend plus numériques notre patrimoine et nos identités traditionnelles. Les portails, interfaces et privilèges d'accès se transforment, ils passent d'un univers structuré par la culture imprimée à un monde régi par la compétence numérique et son environnement évolutif. Les travaux préparatoires des archives numériques conçues pour préserver le patrimoine culturel vont bon train, mais nous assistons aussi, parallèlement, à la conception et à la concrétisation d'autres archives, qui auront peut-être une influence déterminante sur leur orientation et sur leur mise en œuvre.

4. L'hippocratisme numérique : données biométriques, conservation sur longue durée et bases de données

Aux États-Unis et en Europe, un ensemble controversé d'archives ne cesse de grandir : celles qui sont inspirées par le souci de sécurité. On demande depuis peu aux fournisseurs d'accès ou de services de communication de conserver plus longtemps leurs données ; on collecte des informations sur les voyageurs étrangers qui entrent aux États-Unis ; on accumule des données biométriques en vue de délivrer les passeports et cartes d'identité biométriques que l'on projette d'instaurer[1]. Indépendamment du débat politique sur leur nécessité, tous

1. Pour la directive de l'Union européenne sur la conservation longue des données, voir <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006L0024:EN:NOT>. Les États-Unis n'ont pas actuellement de législation de ce genre, mais une nouvelle loi a été proposée. Le Programme des visas des États-Unis, géré par le Département de la Sécurité intérieure, collecte des données biométriques sur les voyageurs qui entrent dans le pays (http://www.dhs.gov/xtrvlsec/programs/content_multi_image_0006.shtm). Le Royaume-Uni a un

ces efforts sont essentiellement des projets d'archivage. Il s'agit de collecter des données identifiantes qui serviront à authentifier des identités, ou de compiler des enregistrements de transactions, en ligne ou autres, pour pouvoir en extraire des informations dont l'analyse permettra de repérer des structures suspectes ou d'autres menaces. Nous sommes donc en train de développer des archives de grande envergure, pas nécessairement si éloignées des archives culturelles qui nous intéressent dans cet essai, mais à des fins de sécurité et de contrôle d'identité. La mise en place de ces contrôles, bien qu'elle soit naissante et parfois maladroite, donne la possibilité de repérer les difficultés et les obstacles auxquels se heurte toute entreprise d'archivage à grande échelle. Ces archives sécuritaires révèlent aussi l'existence de liens étroits et complexes entre certaines protections par copyright, comme celles de leurs bases de données, et la gestion des informations privées et confidentielles. Mais notre principale raison de jeter un bref regard sur leur histoire immédiate tient à ce que nous pouvons en apprendre sur la constitution et la fiabilité de ce type d'archives, sur leur efficacité et, en définitive, sur leur sécurité.

Les bases de données biométriques, que leur collecte soit opérée, à des fins d'identification certaine, sur les étrangers ou sur les nationaux, illustrent les problèmes d'utilisation des informations archivées mais aussi les difficultés à acquérir ces données. Dans le cas de la loi britannique sur la carte d'identité nationale, le ministère de l'Intérieur a commandé une étude de l'enrôlement biométrique, afin d'évaluer et de mesurer le succès de l'acquisition de données auprès des individus[1]. Il est peut-être plus difficile de collecter des données sur des êtres

projet de passeport et de carte d'identité nationale biométriques (<http://www.opsi.gov.uk/ACTS/acts2006/20060015.htm> et voir *supra*, p. 91-94).

1. <http://www.passport.gov.uk/downloads/UKPSBiometrics_Enrolment_Trial_Report.pdf> et voir *supra*, p. 93.