RAPPORT SUR LA NUMÉRISATION DU PATRIMOINE ÉCRIT

Remis par
Marc Tessier au Ministre de la culture et de la communication le 12 janvier 2010

SOMMAIRE

Introduction

La mission sur la numérisation des fonds patrimoniaux des bibliothèques s'est réunie, sous la présidence de Marc Tessier, du 19 octobre 2009 au 7 janvier 2010 (cf. Annexe 1 : Lettre de mission). Elle a procédé à une trentaine d'auditions, y compris de représentants de grandes bibliothèques étrangères.

Elle a analysé le cadre technique, économique et juridique dans lequel s'inscrivent les accords et projets d'accords passés entre la société Google et les bibliothèques. Cette analyse a été conduite dans une perspective de renforcement de la présence et de l'accessibilité des œuvres du patrimoine écrit sur l'internet.

La mission a estimé que cet objectif prioritaire conduisait à s'interroger sur un certain nombre de points, à commencer par l'examen des plates-formes de diffusion de livres numériques existantes, et plus particulièrement celle de Gallica, développée par la Bibliothèque nationale de France (BnF). Cette analyse de l'existant a ensuite permis d'examiner les possibilités d'étendre cette plate-forme et d'en modifier les modes de gestion et les fonctionnalités, afin que le principal acteur public français en la matière puisse être en mesure d'engager des discussions avec des partenaires privés sur une base équilibrée. L'axe européen, enfin, a retenu toute l'attention de la mission, car une alternative crédible à de grands projets internationaux ne peut pas par définition se construire sur une base exclusivement nationale.

Le présent rapport s'articule donc en trois temps :

- **un état des lieux des principales bibliothèques numériques** – y compris, s'agissant de Google Livres, de la situation juridique complexe dans laquelle se trouve sa maison mère ;

- **une analyse des accords** passés entre les bibliothèques et Google, qui ne semblent pas apporter de réponse suffisamment adaptée aux missions des bibliothèques ;

- **des pistes d'action**, se déclinant en trois axes : le changement d'échelle de la numérisation des ouvrages et du mode de fonctionnement de Gallica ; une proposition de partenariat avec Google Livres qui passerait notamment par un échange de fichiers numérisés, sans exclusivité sur les fichiers échangés ; enfin, la relance d'une impulsion européenne, tant en direction des autres bibliothèques européennes que du portail Europeana.

Une conclusion en forme de résumé exécutif reprend ces différentes solutions.

*
* *

I. Etat des lieux : des avancées hétérogènes dans un environnement incertain

I.1. Où en sont les bibliothèques numériques ?

I.1.1. Une idée relativement ancienne qui a connu un réel essor à partir de 2004

■ **L'idée de numériser des livres pour constituer des bibliothèques numériques est relativement ancienne**[1] : dès 1971, Michael Hart, étudiant de l'Université de l'Illinois (aux États-Unis), développe la première initiative de bibliothèque numérique, le « projet Gutenberg ». Il s'appuie sur une équipe de volontaires pour relire et vérifier l'océrisation[2] des ouvrages numérisés, qui relèvent tous du domaine public[3]. Le site annonce aujourd'hui plus de 100.000 livres disponibles via un réseau de partenaires, et 30.000 ouvrages disponibles gratuitement et directement depuis le site. Essentiellement anglophone au départ, le projet a commencé à s'intéresser à des ouvrages dans d'autres langues depuis 1997.

Ce projet a inspiré ensuite la création ou les projets de création de grandes bibliothèques numériques – à commencer par l'idée, émise par Jacques Attali lors des toutes premières réflexions sur la création en France d'une Très Grande Bibliothèque, de sauter une étape pour directement élaborer une « Bibliothèque numérique francophone ». Ce projet ne verra pas tout de suite le jour, mais la Bibliothèque nationale de France (BnF) lancera cependant la première version de Gallica dès 1997, avec au départ une approche sélective et une numérisation en mode image uniquement. Dans un premier temps, Gallica a ainsi proposé 3.000 livres en mode image, avant d'évoluer progressivement (cf. infra).

■ Les projets de grandes bibliothèques numériques ont connu une nouvelle actualité avec les initiatives des **grands moteurs de recherche**. Les moteurs de recherche ont en effet un intérêt spécifique à ce que la plus grande masse de contenus possible soit moissonnée par leurs robots, puisque ces contenus élargissent leur base de recherche et l'efficacité et la pertinence de leurs résultats.

Google a été le premier à lancer, non sans controverse, une nouvelle plate-forme en octobre 2004, alors appelée Google Print, avant de devenir Google Book Search en novembre 2005. L'ambition affichée était de numériser 15 millions d'ouvrages en dix ans, en s'appuyant principalement[4] sur les ouvrages conservés dans les fonds des cinq premières bibliothèques partenaires – la New York Public Library, et les bibliothèques des universités de Harvard, Stanford, du Michigan, ainsi que la Bodleian library à Oxford.

En réaction à Google Book Search, qui n'autorise pas les autres moteurs de recherche à indexer les éléments présentés sur sa plate-forme, d'autres acteurs du secteur se sont lancés dans des projets initialement assez comparables. **Microsoft** a lancé, en décembre 2006, son propre programme de numérisation de livres : son moteur de recherche Live Search était désormais associé à une famille de services, dont une plate-forme de livres numérisés, « Live Book Search », qui devait être alimentée grâce à des

---

1 Voir notamment l'article de Jean-Michel Salaün, « Bibliothèques numériques et Google Book Search », in *Regards sur l'actualité n° 316*, La Documentation française, décembre 2005.

2 L' « océrisation », de l'acronyme anglais OCR (reconnaissance optique de caractères), désigne l'opération consistant, après avoir scanné un livre, à utiliser des logiciels informatiques permettant de reconnaître les caractères imprimés sur le document (lettres, signes ou espaces) et de répertorier chaque mot. C'est un procédé essentiel pour permettre ensuite des recherches sur tous les mots contenus dans le texte (recherche dite « plein texte »).

3 Au sens de la loi américaine – il s'agit donc d'ouvrages publiés antérieurement à 1923.

4 Mais pas uniquement : dès l'origine, des accords avec des éditeurs ont également été signés.

partenariats avec la British Library, la New York Public Library et, là aussi, des bibliothèques universitaires américaines (universités de Cornell, de Toronto et de Californie). Mais le projet a finalement été abandonné en mai 2008, à la faveur d'une réorganisation profonde des activités de Microsoft, qui a choisi de séparer le développement de son moteur de recherche (devenu Bing au lieu de Live Search) de la famille de services Live Search.

Yahoo! a lui aussi, cherché à développer ses activités de numérisation en s'appuyant sur l'Internet Archive – un organisme à but non lucratif, qui existe depuis avril 1996 et dont le but est d'archiver le web. Ils créent ensemble l'Open Content Alliance (OCA), qui rassemble des partenaires nombreux (bibliothèques des universités de Californie et de Toronto, Archives nationales britanniques, Research Library Group, ainsi que diverses sociétés informatiques). Le site expérimental d'OCA[5] permet d'accéder à plus d'un million de livres du domaine public, là encore essentiellement anglo-saxons.

■ **L'initiative de Google a également fait réagir les États** au travers d'organisations internationales.

À l'initiative de la France et de cinq autres États européens dont l'Allemagne, l'Union européenne a ainsi lancé, en mars 2006, la création de la Bibliothèque numérique européenne (BNUE), qui s'inscrit dans le cadre de la Stratégie de Lisbonne (volet « i2010 »). Le portail **Europeana** est ouvert en 2008. L'objectif est à la fois d'offrir un accès gratuit au patrimoine numérique européen à travers 10 millions de documents mis en ligne d'ici à 2011, et éventuellement de proposer un accès payant aux contenus sous droits des éditeurs partenaires.

L'Unesco a de son côté annoncé en décembre 2006 le lancement de la World Digital Library, qui en réalité s'apparente davantage à une vaste banque de données culturelles et multilingues très sélective qu'à une bibliothèque de livres numériques.

I.1.2. Une idée qui s'est concrétisée par des avancées hétérogènes

L'état actuel de la situation des différents projets de bibliothèques numériques aujourd'hui révèle des avancées hétérogènes, selon les plates-formes et les acteurs. Le panorama suivant n'est pas exhaustif, mais est principalement centré sur les sites contenant des ressources francophones importantes[6].

■ **Google Book Search, aujourd'hui appelé Google Books – en français, Google Livres,** dénomination qui sera retenue dans la suite de ce rapport – est une plate-forme hébergeant une base de données et dotée d'un moteur interne. Cet outil stocke et indexe le contenu des livres scannés, traités et stockés au format numérique par la société Google.

En termes d'utilisation, l'internaute peut soit se rendre sur le site de la plate-forme et y effectuer directement ses recherches, s'il cherche uniquement du contenu en provenance de livres, soit utiliser le moteur Google, où il pourra accéder à des résultats composés à la fois de pages web et d'extraits de certains livres pertinents. Le contenu de Google Livres est donc important non seulement du point de vue de la plate-forme mais également de celui du seul moteur, puisqu'il lui permet d'accroître la base à partir de laquelle il effectue ses recherches et, partant, la richesse et la pertinence de ses résultats.

Lorsqu'un résultat en provenance de la base Google Livres apparaît, l'usager, en cliquant sur le lien, ouvre une interface qui lui permet de visualiser des niveaux d'informations différents selon le statut de l'œuvre. Pour les livres du domaine public, l'ouvrage peut être vu en entier et téléchargé au format image PDF et texte Epub ; pour les œuvres sous droit, l'expérience sera différente selon que des

[5] La partie du site permettant l'accès aux ouvrages est accessible uniquement en version bêta depuis l'Europe.

[6] L'annexe 3 fait une comparaison approfondie entre les fonctionnalités offertes par Gallica et par Google Livres.

accords auront été conclus entre la société Google et les éditeurs ou pas : soit l'usager peut lire quelques pages de l'ouvrage et suivre un lien renvoyant vers le site de l'éditeur (éditeurs partenaires), soit il n'aura accès qu'aux seules références de l'œuvre éventuellement assorties de courts extraits (« *snippets* »), pour les éditeurs n'ayant pas signé d'accord. Dans tous les cas, l'affichage des données s'accompagne de liens renvoyant vers des sites de librairies et de bibliothèques, sur le côté gauche de l'écran.

Le site est alimenté principalement par deux sources. D'une part, les bibliothèques ayant signé des accords de numérisation qui proposent généralement à la numérisation des livres hors droit. Mais Google a aussi été en mesure de numériser, via les fonds de grandes bibliothèques américaines, des ouvrages sous droits, sans obtention préalable du consentement de leurs ayants droit, ce qui a suscité un contentieux important tant aux États-Unis qu'en Europe, notamment en France (cf. infra, I.2.1). L'autre source est celle des éditeurs partenaires. Enfin, Google se procure également des métadonnées – informations d'identification de l'ouvrage – et reconstitue une image banalisée de couverture, lorsqu'il ne détient pas le contenu numérisé, afin de pouvoir donner accès à un minimum d'informations (titre, auteur, éditeur, ISBN, nombre de pages…) sur le livre. Une recherche sur un ouvrage récent d'un éditeur non partenaire donnera donc accès à une page d'informations assortie, le cas échéant, d'avis d'internautes et de liens vers des sites de librairies et bibliothèques.

Début 2010, Google Livres annonce que la plate-forme permet d'effectuer des recherches sur l'intégralité de plus de 10 millions de livres[7]. Parmi ces livres, 2 millions ont été numérisés en partenariat avec les éditeurs et 1,5 millions relève du domaine public. Les autres ouvrages, sous droits, ont été numérisés sans accord des ayants droit.

■ **La bibliothèque numérique Gallica** est développée par la BnF depuis le milieu des années 1990, dans le cadre du grand projet voulu par François Mitterrand. Elle a été inaugurée en 1997 avec une offre de quelques dizaines de milliers de documents, principalement en mode image. Conçue à l'origine comme une bibliothèque numérique sélective à vocation encyclopédique proposant des corpus de documents (les revues des sociétés savantes, les voyages en Italie, ...), elle a profondément changé à compter de 2005, en contrepoint des projets de numérisation de Google. La BnF a alors développé à son tour une politique de numérisation de masse (marché Jouve dit « des 30.000 », marché Safig dit « des 100.000 » en 2007) et validé un passage au mode texte (marché d'océrisation des contenus déjà présents dans Gallica, dit « des 60.000 »).

Une autre évolution importante a été l'ouverture de discussions avec le Syndicat national de l'édition (SNE) fin 2007, en vue de permettre un accès à des contenus numériques sous droits *via* Gallica. Les éditeurs français sont désormais présents sur Gallica à travers le signalement dans ce portail de près de 20.000 livres contemporains numérisés. Les documents sont consultables, sous conditions, sur le site de distributeurs numériques.

À partir de 2005, Gallica s'est également enrichi de contenus de presse (presse quotidienne du XIXe siècle de grand format) avec un important marché de numérisation spécifique (3,5 millions de pages, une vingtaine de titres concernés) qui a obtenu un soutien financier du Sénat.

Fin 2009, Gallica donne accès à plus de 950.000 documents dont environ 370.000 en mode texte. Parmi ces documents : 145.000 livres (monographies), 650.000 fascicules de périodiques, 115.000 images.

[7] http://googleblog.blogspot.com/2009/10/tale-of-10000000-books.html

930.000 documents sont issus des collections de la BnF, les autres provenant soit des éditeurs associés au projet, soit de bibliothèques partenaires. La BnF a en effet entrepris de donner accès à des documents numériques d'autres bibliothèques, soit en les hébergeant, soit en les moissonnant par le protocole OAI-PMH. Cette offre demeure cependant encore modeste avec moins de 4.000 documents de bibliothèques partenaires accessibles depuis Gallica (0,4 % du total de Gallica). Les documents libres de droits sont également signalés sur Europeana dont Gallica est l'un des agrégateurs pour la France.

Les principaux chantiers techniques aujourd'hui en cours sont la modernisation de l'interface de consultation (un nouveau visualiseur est ainsi proposé en décembre 2009), la modernisation du moteur de recherche (courant 2010) ou encore le renforcement des capacités de stockage et diffusion afin d'améliorer la qualité de la réponse apportée aux internautes. Un travail sur la structuration des données numériques et des métadonnées associées est également effectué par la BnF, notamment dans un cadre international. Par ailleurs de nouveaux marchés de numérisation (documents spécialisés d'une part, livres rares et précieux d'autre part) ont été lancés en 2009.

Par comparaison, on peut indiquer qu'aux **États-Unis, la bibliothèque du Congrès** a développé, dès le début des années 1990, une politique numérique ambitieuse s'appuyant sur d'importants financements publics (provenant du Congrès) et privés - plus de 45 millions de dollars ont ainsi été obtenus auprès d'acteurs privés, notamment sous forme de dons. Le résultat est le programme « *American Memory* » (http://memory.loc.gov/ammem/index.html) soit une bibliothèque numérique de plus de 5 millions de documents en accès libre, principalement des manuscrits, des documents iconographiques et de la presse, selon les objectifs de la politique documentaire définie par la grande bibliothèque nationale nord-américaine. Ces documents, répartis dans une centaine de collections thématiques, proviennent de la Bibliothèque du Congrès mais aussi d'autres institutions culturelles américaines. Pour sa part, le Japon a récemment refusé un partenariat avec Google en matière de numérisation de livres et décidé fin 2009 d'entreprendre son propre programme national de numérisation sur financements publics avec comme acteur majeur **la Bibliothèque de la Diète** qui joue dans ce pays le rôle de Bibliothèque nationale. Les financements envisagés seraient de 90 millions d'euros pour l'année 2010 et de l'ordre d'1 milliard d'euros pour l'ensemble du programme.

■ **La bibliothèque numérique Europeana** est en fait un portail de consultation et non pas un site hébergeant les contenus eux-mêmes. Son développement a été confié à une fondation de droit néerlandais, EDL (European digital library), dont le financement est actuellement assuré en partie par la Commission européenne dans le cadre d'appels à projet, et en partie par un certain nombre d'États membres.

Le portail Europeana a été inauguré en novembre 2008 (version bêta, www.europeana.eu). Il propose à la consultation environ 6 millions de documents, dont en réalité assez peu de livres (moins de 200.000). Les contenus proposés par la France, principalement à travers le portail Collections du ministère de la culture, la bibliothèque numérique Gallica (cf. *supra*) et le site de l'INA, représentent actuellement environ la moitié du total des documents accessibles *via* Europeana. La mise en service de la version opérationnelle est prévue au deuxième semestre 2010 avec un objectif de 10 millions de documents en ligne. Plus d'un millier d'institutions culturelles européennes participent à Europeana mais avec des degrés d'implication et des offres de contenus extrêmement inégaux. La Commission européenne réfléchit actuellement à l'évolution d'Europeana et a lancé pour cela, à la fin du mois d'août 2009, une consultation publique « *Europeana - next steps* ». Les principales questions portent notamment sur les contenus que le portail doit offrir aux internautes, les modes envisageables de financement et de gouvernance, les solutions possibles et acceptables pour mieux associer le secteur privé à ce projet et accroître son rayonnement.

■ De leur côté, outre l'offre proposée dans Gallica, **les principaux éditeurs français**[8] ont entrepris la constitution d'une offre numérique susceptible de répondre aux attentes des internautes et respectueuse du droit d'auteur. La mise en place de cette offre suppose une évolution des différents métiers de l'édition, de lourds investissements financiers (avec un soutien des pouvoirs publics notamment à travers les nouvelles aides numériques du Centre national du livre, créées en 2008) et l'identification des droits effectivement détenus par chacun pour l'exploitation numérique des œuvres.

Le développement de cette offre numérique (que l'on peut évaluer fin 2009 à environ 40.000 titres de l'édition française disponibles) s'est traduite en 2009 par le lancement de plusieurs plates-formes de distribution (Numilog, d'Hachette, Eden-Livre regroupant Flammarion, Gallimard et La Martinière, site E-Plateforme d'Editis, L'Harmathèque de L'Harmattan, etc.). Ces différentes plates-formes de distribution (« B to B ») s'ajoutent à une offre plus ancienne, constituée plutôt par des agrégateurs numériques indépendants des éditeurs (essentiellement Cyberlibris et Numilog, avant son rachat par Hachette) et tournée directement vers les internautes (« B to C »)[9]. À moyen terme les éditeurs français préparent la transition vers une filière de production nativement numérique. L'offre numérique éditoriale devrait être surtout constituée, au moins dans un premier temps, de titres récents[10].

I.1.3. Une idée portée par l'évolution des usages

L'émergence de ces différentes bibliothèques et plates-formes de livres numériques n'aurait pu avoir lieu sans le développement d'usages nouveaux, spécifiques à la recherche sur la toile. L'essor rapide de l'internet a en effet entraîné des changements profonds dans les modes d'accès au savoir et à l'information. Deux types d'usages expliquent en partie l'intérêt suscité par le développement de bibliothèques numériques et peuvent profondément influencer les réflexions en matière d'élaboration de telles bibliothèques.

■ Le premier de ces usages est le recours désormais prioritaire aux **moteurs de recherche**.

Les moteurs de recherche sont aujourd'hui des outils universellement reconnus comme particulièrement efficaces pour permettre aux internautes d'accéder à la masse de connaissances disponibles sur la toile. Outre Google, certains moteurs de recherche ont été largement utilisés dans le passé ou le sont encore à des degrés divers aujourd'hui (Altavista, Yahoo!…), d'autres émergent (Bing) ; mais l'outil que représente le moteur de recherche est incontournable pour les internautes, y compris dans leurs usages de consommation culturelle. Ce succès a été remporté principalement par la conjugaison d'un modèle économique très robuste, gratuit pour l'utilisateur, par la simplicité d'utilisation de ce type d'outil et par la puissance de l'algorithme, fondée sur une conception spécifique de la pertinence et une infrastructure technologique extraordinairement puissante et performante.

Deux éléments définissent l'efficacité d'un moteur de recherche : sa pertinence et sa puissance. Or les principaux moteurs du web – en particulier Google – ont d'abord fait le choix de la puissance. On rappellera brièvement **les grands principes de fonctionnement d'un moteur de recherche** du type de Google, qui sont utiles pour comprendre la stratégie qu'ils peuvent avoir en matière de

[8] Les éditeurs de STM (sciences-techniques-médecine) ont déjà engagé depuis plusieurs années l'accès numérisé à leurs collections, particulièrement dans les domaines du droit, de la médecine ou des sciences. La nouveauté est l'extension au domaine de la littérature générale, dont il s'agit principalement ici.

[9] Hachette a annoncé le lancement du site « Myboox », magazine « B to C » assorti de fonctions commerciales et communautaires.

[10] Un projet tel que celui des éditions Gallimard (numérisation de plus de 25.000 titres du fonds) fait pour l'instant plutôt figure d'exception.

numérisation :
- tout d'abord, un robot explore de façon automatique et régulière la toile ; le robot suit tous les liens hypertextes qu'il rencontre, pour récupérer et indexer toutes les ressources utiles. La première étape consiste donc à visiter extensivement la toile[11], afin d'y repérer des documents et des pages web ;
- l'étape suivante consiste à indexer les documents collectés, qui pourront ensuite être recherchés par les internautes grâce à des mots clés y figurant. Il faut donc extraire les mots significatifs de chaque document, qui sont ensuite classés selon un dispositif d'indexation propre au moteur de recherche. Parallèlement, les mots extraits sont affectés d'une pondération, qui correspond généralement à la fréquence d'apparition de ce mot dans le document (mais d'autres critères peuvent être utilisés) ;
- l'étape de recherche est celle qui, après requête des internautes, restitue les résultats par ordre de pertinence. Celle-ci est appréciée en fonction d'algorithmes propres à chaque moteur. Google se fonde, notamment, sur le modèle du « *page rank* » : la pertinence d'un document y est définie notamment au regard de sa notoriété sur la toile, laquelle fait l'objet d'un calcul complexe tendant à fixer pour chaque page web un indice de popularité « fiable ». Ce score est, par essence, évolutif, dans la mesure où il résulte d'une analyse globale et permanente des pratiques de liens et de consultations sur le web.

C'est donc notamment à l'aune de leur popularité que les résultats sont présentés. Des recherches sont cependant en cours pour développer d'autres types de moteurs, davantage fondés sur la pertinence grâce à des analyses sémantiques. L'idée est notamment d'associer au terme recherché d'autres mots dont le contenu sémantique ou logique est proche, afin de répondre à la question posée – alors que les moteurs actuels se bornent à rechercher la concordance entre les mots-clés de la requête et leur index. Enfin, l'évolution vers un « web sémantique » devrait permettre la création automatique de liens entre les documents numérisés (par exemple, la version numérique d'un livre, la mention de ce livre dans un article, une biographie de l'auteur sur Wikipedia, etc.), voire de hiérarchiser ces documents entre eux. Mais la mise en place de ce « web sémantique » implique encore un important travail initial sur la qualification des différentes données du Web et semble tarder à voir le jour.

Cependant, l'accès potentiellement universel aux ressources en ligne proposé par les moteurs actuels paraît suffisamment séduisant pour les internautes, qui plébiscitent ce mode de recherche. À plus forte raison lorsqu'il s'agit d'ouvrages : le fait de trouver immédiatement des contenus en ligne, permettant de s'affranchir des contraintes de temps et de déplacement liés à la mise à disposition des livres « papier », semble présenter pour les chercheurs comme pour le grand public un intérêt largement supérieur à d'éventuelles faiblesses de qualité tenant aux modes de recherche des moteurs.

Le développement de ces usages est donc **suffisamment incitatif** pour que l'on s'attache à numériser le plus rapidement possible des contenus afin de les mettre à disposition en ligne. Pour les moteurs de recherche, cette motivation est encore renforcée par la volonté de disposer d'un plus grand nombre de documents disponibles pour améliorer la richesse et la pertinence de leurs réponses et, partant, accroître l'assiette documentaire de leurs ressources publicitaires.

■ Le développement des **réseaux sociaux**

Dans l'univers des réseaux sociaux, l'utilisateur constitue son propre univers et le contextualise. La logique est très différente de celle du moteur de recherche : dans certains domaines, et notamment en matière d'information, ce n'est plus l'internaute qui va chercher lui-même l'information – il attend au contraire que l'information lui arrive par l'intermédiaire de son réseau. Il bénéficie ainsi d'une information filtrée et contextualisée en fonction de ses propres centres d'intérêts ou des personnes « ressources » de son réseau. L'internaute peut aussi, à son tour, proposer aux membres de son réseau

[11] Pour diverses raisons, notamment techniques, une partie du *web* n'est cependant pas accessible à ces robots : on parle alors de « Web profond » ou « Web invisible ».

sa bibliothèque idéale, et devenir lui-même source d'information (une application en ce sens est par exemple proposée par Facebook).
Cet usage relativement nouveau a un intérêt spécifique pour les livres numérisés : il s'agit d'un autre mode d'accès possible, différent de celui des moteurs. L'information sur le livre et son contenu passe par d'autres modes de recherche que les algorithmes ou le web sémantique. Le développement rapide de ce nouvel usage doit dès lors être pris en compte par les bibliothèques numériques, en proposant des services *ad hoc*.

I.2. Un environnement incertain

I.2.1. Google se trouve dans un contexte juridique complexe

■ La numérisation, dans les fonds des bibliothèques partenaires, d'œuvres sous droits sans consentement préalable de leurs ayants droit a suscité dès 2005 **un contentieux aux États-Unis.** Le projet de **règlement transactionnel** auquel sont parvenues les parties le 28 octobre 2008 **doit encore être validé par le juge** alors qu'il a soulevé une émotion internationale justifiant l'introduction **d'amendements**.

Les œuvres sous droits numérisées à partir des collections des bibliothèques universitaires américaines sont entièrement indexées par le moteur ; la recherche « plein texte » conduit à l'affichage de courts extraits présentés sous la forme de bandelettes de papier déchirées (les « *snippets* »). Dès 2005, les associations américaines d'ayants droit (*American Publishers Association* et *Author's Guild*) ont intenté contre la société Google une « action de classe » dans laquelle elles se sont portées parties au nom des « classes » entières qu'elles représentaient (c'est-à-dire tous les éditeurs et tous les auteurs).

Il s'agissait d'un procès en contrefaçon de droits d'auteurs : la société Google se voyait reprocher de violer le « *Copyright* » par la reproduction et la représentation de ces livres sans autorisation préalable. Elle opposait à ces accusations l'argument de l'exception dite de « *fair use* » (utilisation loyale), exception très générale appliquée dans le droit américain. Elle soulignait également qu'elle était disposée à retirer les ouvrages à la demande de leurs ayant droits qui en feraient la demande (« *opt out* », pratique très contestée parce que contraire aux principes de la propriété intellectuelle ; la lenteur avec laquelle Google semble donner suite aux demandes de retrait a également été mise en avant).

Le juge n'a pas eu à se prononcer sur le fond : en octobre 2008, après trois ans d'une procédure très coûteuse, les parties ont rendu public un projet d'accord transactionnel de classe (*Class action settlement agreement*) visant, s'il était validé par la cour, à éteindre le contentieux. Par une spécificité du droit américain, cet accord aurait le pouvoir de lier tous les membres des « classes » représentées, sauf ceux qui s'en seraient explicitement retirés. Une vaste campagne de publicité a alors été entreprise dans le monde entier pour signifier aux auteurs et aux éditeurs qu'un document de plus de 300 pages, rédigé en anglais juridique, était sur le point de modifier leurs droits sur leurs propres livres.

Ce premier projet comportait deux grands volets. D'une part, il avait pour effet d'éteindre, par un système de dédommagement, toutes les poursuites passées et à venir contre la société Google pour les faits initialement reprochés. D'autre part, il mettait en place des modalités pour l'exploitation commerciale par Google de tous les livres numérisés. Si les livres n'étaient pas disponibles dans les grands canaux de vente américains, Google les exploitait par défaut, sauf objection expresse, formulée titre par titre par les éditeurs. Cette exploitation devait se limiter au territoire américain, sur la foi de l'adresse IP des consommateurs. Dans le cas des livres « revendiqués » par leurs ayants droit, si ces derniers autorisaient l'exploitation par Google, ils se voyaient reverser 33% du chiffre d'affaires généré

(publicité, accès payant individuel par titre, accès payant institutionnel à la base toute entière). 33% du chiffre d'affaires généré par les livres « non revendiqués » était reversé à un « Registre des droits sur les livres » (*Books Rights Registry*) qui aurait pour mission d'inciter les ayants droit du monde entier à s'enregistrer.

Des critiques nombreuses et très vives ont été portées à ce premier projet. Il méconnaissait les principes de la propriété intellectuelle en obligeant les ayants droit à l'*opt out* s'ils ne voulaient pas voir leurs livres exploités par Google. Il mettait cette même société dans une situation de monopole, sur le territoire américain, pour l'exploitation des ouvrages « non revendiqués » : l'accord transactionnel lui attribuait en effet, et à elle seule, une licence d'exploitation exorbitante au droit commun. La société Google présente cet aspect du projet d'accord comme la seule façon qui s'offrait à elle de résoudre la question des œuvres « orphelines », œuvres dont les ayants droit n'ont pas été identifiés et qui, en l'absence d'accord possible de leur part, ne peuvent pas, en principe, être exploitées (cf. *infra* I.2.2).

Le juge chargé de l'éventuelle validation du projet d'accord transactionnel a reçu un nombre considérable d'objections et de remarques. Les gouvernements français et allemand, puis le gouvernement américain, lui ont adressé des courriers exprimant les plus grandes réserves sur les termes et les effets de la transaction. Les parties ont donc décidé d'amender le projet et un nouveau document a été rendu public le 13 novembre 2009. La modification la plus significative est que l'accord ne porterait plus, dorénavant, que sur les livres initialement publiés aux États-Unis, au Royaume-Uni, au Canada et en Australie, ainsi que sur les livres inscrits - par les éditeurs du monde entier - au Bureau du copyright des Etats-Unis (soit, pour un certain nombre d'éditeurs français, une part importante de leur catalogue). Quelques aménagements visent par ailleurs à modérer le caractère monopolistique du système. L'économie générale du projet n'est cependant pas remise en cause.

Le juge doit donner son avis sur ce projet amendé le 18 février 2010. Si le texte de cet accord transactionnel est validé, Google sera à même de mettre en place une immense plate-forme de commercialisation de livres, très majoritairement anglophones et elle disposera de l'exclusivité entière d'exploitation d'une grande part de ces livres, c'est-à-dire tous ceux qui n'auront pas été « revendiqués » auprès du « Registre des droits sur les livres ». Cependant, les remarques faites au juge par le gouvernement américain semblent loin d'avoir été prises en compte dans le projet amendé et une intervention des autorités américaines de la concurrence n'est pas à exclure.

■ Un contentieux similaire a une traduction judiciaire en France

Une procédure a été engagée le 6 juin 2006 par le groupe La Martinière contre les sociétés Google Inc. et Google France sur des chefs globalement similaires à ceux qui avaient été avancés par les ayants droit américains (cf. *supra*), c'est-à-dire la contrefaçon de droits d'auteur par reproduction sans autorisation et mise à disposition de courts extraits des livres sous la forme de « *snippets* ». Les plaignants ont été rejoints en octobre 2006 par une intervention volontaire du Syndicat national de l'édition (SNE) et en novembre 2006 par une intervention volontaire de la Société des Gens de Lettres (SGDL).

En défense, Google estimait que, pour les actes de numérisation, il convenait d'appliquer le droit américain dans la mesure où les opérations ont eu lieu sur le territoire des États-Unis ; que, par conséquent, cette numérisation n'était pas une contrefaçon dans la mesure où elle relevait de l'exception dite du « *fair use* » ; et que, pour les actes de représentation, cette pratique entrait dans le cadre de l'exception de courte citation reconnue par le droit français[12].

L'audience a eu lieu le 24 septembre 2009 et le tribunal de grande instance de Paris a rendu son

[12] Article L 122-3 du Code de la propriété intellectuelle.

jugement le 18 décembre. Ce dernier considère que le droit applicable est le droit français, autant pour les opérations de numérisation que pour la représentation de courts extraits. Ceci posé, il conclut que la société Google Inc. s'est effectivement rendue coupable de contrefaçon de droits d'auteur par la reproduction sans autorisation préalable puis par la représentation d'œuvres protégées. Il considère en effet que l'exception de courte citation n'est pas applicable à la démarche de Google dans la mesure où les extraits sont affichés de manière aléatoire et excluent donc tout but d'information.

La société Google Inc. a, par conséquent, été condamnée en première instance à verser un dédommagement de 300.000 € au groupe La Martinière et d'1 € symbolique à la SGDL et au SNE. Le tribunal lui interdit par ailleurs de poursuivre ces agissements sous astreinte de 10.000 € par jour de retard.

Il convient de souligner que le jugement ne porte que sur une liste précise et bien identifiée de livres qui avait été établie par un constat d'huissier pour initier la procédure. Cependant, Google court maintenant le risque de faire face à une multitude de procès similaires intentés par des éditeurs français considérant que le jugement est transposable à leur propre situation. Il a déclaré son intention d'interjeter appel de ce jugement.

I.2.2. Une coordination insuffisante des autres acteurs

■ **Au niveau européen**, certaines questions restent à préciser. **La première d'entre elles est celle des « œuvres orphelines »** : un droit sur une œuvre est orphelin si au moins un des titulaires de ce droit n'a pu être identifié malgré des recherches avérées et sérieuses. La question des « œuvres orphelines » a pris une acuité particulière avec les grands projets de numérisation des institutions culturelles : une institution ne peut en effet, sauf exceptions particulières[13], numériser et mettre à la disposition des internautes des œuvres protégées sans avoir obtenu auparavant l'assentiment des titulaires des droits d'exploitation numérique de ces œuvres. Or la recherche de ces ayants droit peut être extrêmement coûteuse et parfois infructueuse. Tout projet de numérisation de masse est confronté à cet obstacle juridique. On a vu que Google a contourné cet obstacle en numérisant systématiquement des livres protégés sur le territoire américain sans les mettre initialement à disposition des internautes, ce qui permettait à l'entreprise de s'abriter derrière le principe du « *fair use* », d'où le contentieux avec les ayants droit. Si le projet de règlement à l'amiable était validé par le juge américain, Google serait en situation d'exploiter le produit de sa numérisation sans avoir à rechercher les titulaires des droits sur les livres (cf. *I.2.1.*).

Cette perspective laisse craindre un monopole sur l'exploitation numérique des livres non revendiqués, d'où le vif désir qui anime la Commission européenne de donner aux institutions culturelles européennes la sécurité juridique nécessaire pour mener à bien leurs propres projets de numérisation de masse. En France, à la suite des recommandations de la commission spécialisée du Conseil supérieur de la propriété littéraire et artistique[14], le ministère de la culture et de la communication s'engage en 2010 dans la mise en place d'un système permettant la gestion collective de ces œuvres orphelines, au moins pour le domaine de l'imprimé et de l'image fixe. D'autres pays européens s'orientent vers des solutions conformes à leur sensibilité culturelle nationale (« licences étendues » en Scandinavie, par exemple). La tâche de la Commission sera de veiller à ce que la multiplicité de ces systèmes ne nuise pas à la circulation et à la libre diffusion des œuvres concernées, sans céder à la tentation d'une nouvelle exception communautaire peu compatible avec la diversité de traditions nationales souvent sensibles.

[13] En France, par exemple, on peut citer l'exception « bibliothèques » (CPI, L 122-5 et L 211-3) ou encore le droit particulier s'appliquant aux institutions en charge du dépôt légal (Code du patrimoine, L 132-4).

[14] CSPLA, Commission sur les œuvres orphelines, 18 mars 2008.

La seconde question concerne le financement d'Europeana. Le portail européen tire jusqu'à présent l'essentiel de ses ressources d'appels à projets communautaires, par nature non pérennes, ce qui constitue un point de fragilité majeur. Cette situation pourrait évoluer à compter de 2013 avec l'entrée de l'Union européenne dans un nouveau cycle budgétaire. Afin de consolider le fonctionnement d'Europeana et de lui permettre d'entrer dans une démarche stratégique pluriannuelle, il est en effet nécessaire d'asseoir le financement du portail sur des ressources régulières. Les contributions peuvent avoir trois sources : Commission, États membres, partenaires privés. Un équilibre satisfaisant entre ces trois axes reste à trouver, de même que l'organisation d'un pilotage nouveau pour le projet. Ces conditions sont nécessaires pour le futur développement de l'interface. L'équilibre dépendra surtout de la volonté plus ou moins grande des différents acteurs en présence de soutenir le projet.

■ **Des initiatives privées non coordonnées** : l'offre des éditeurs en France

Dans la mesure où le livre numérique demeure un marché émergent, sans un chiffre d'affaire ni une régulation par les pouvoirs publics comparables à ce qui existe depuis assez longtemps dans l'univers physique (livre papier), les actions des principaux acteurs économiques, en particulier les éditeurs de littérature générale, sont restées jusqu'à il y a peu du domaine de l'expérimentation.

Des solutions interprofessionnelles garantissant l'interopérabilité des offres et leur plus grand déploiement n'ont pu pour l'instant émerger. Les principales initiatives demeurent du ressort des différents groupes économiques – avec notamment, depuis 2009, la mise en place de différentes plates-formes de distribution ou de diffusion de livres numériques (cf. *supra*). Il reste désormais à assurer le lien entre ces plates-formes « B to B » et les portails de vente de libraires (« B to C »). Une conciliation interprofessionnelle doit intervenir pour garantir à chaque détaillant, dès lors qu'il s'est mis en mesure, avec l'accord des diffuseurs, de vendre des livres numériques, un accès simple, homogène et exhaustif à l'offre de tous les éditeurs français. Le projet de « hub » récemment exposé par Dilicom, structure interprofessionnelle opérant déjà comme centrale d'information et de prise de commande pour le livre physique, s'inscrit dans ce mouvement. Son succès semble toutefois conditionné par l'unanimité dont feront preuve les éditeurs français à voir leur offre représentée sur le plus de plates-formes de commercialisation possible, sans exclusive ni restriction.

La réflexion interprofessionnelle existe cependant, animée par les syndicats professionnels (SNE avec sa commission numérique, SLF avec le projet de portail de la librairie indépendante) et accompagnée par les pouvoirs publics (réflexion organisée à travers différents rapports et missions[15], soutien financier du Centre national du Livre à diverses initiatives).

Toutefois, du point de vue de l'internaute, le panorama de l'offre peut paraître encore peu lisible et plutôt éclaté.

I.2.3. Une introuvable définition du livre numérique

[15] Cf. Rapport sur le livre numérique établi par Bruno Patino en juin 2008 ; rapport d'Hervé Gaymard, Situation du livre – Evaluation de la loi relative au prix du livre et questions prospectives (mars 2009) ; rapport de la mission « Création et internet », confiée à Patrick Zelnik, Jacques Toubon et Guillaume Cerruti, remis le 6 janvier 2010 au ministre de la culture et de la communication – ce dernier rapport a recommandé notamment la mise en place d'une plateforme commune de distribution numérique selon un modèle B to B. Consultable en format PDF à l'adresse : http://www.culture.gouv.fr/mcc/Actualites/A-la-une/Remise-du-rapport-de-la-mission-creation-et-internet

Enfin, de façon plus générale, **le « livre numérique » reste un objet difficile à définir** de façon figée, ce qui explique en partie la numérisation plus lente dans ce domaine que dans d'autres secteurs culturels comme la musique ou le cinéma. Un livre numérique peut recouvrir des réalités très diverses, allant du simple fac-similé digitalisé de l'imprimé au produit numérique présentant des fonctionnalités très raffinées (recherche, intertextualité, interactivité, enrichissements éditoriaux…). Alors que la réglementation fiscale définit le livre comme un « ensemble imprimé, illustré ou non, publié sous un titre, ayant pour objet la reproduction d'une œuvre de l'esprit d'un ou plusieurs auteurs en vue de l'enseignement, de la diffusion de la pensée et de la culture », le contenu numérisé d'un livre peut être proposé sous des formes et des supports très divers. Le fichier peut être intégralement téléchargeable ou accessible *via* un abonnement à un univers thématique ; il peut être « figé » ou annotable et transformable… - autant de possibilités diverses qui rendent difficile une définition unique[16] et multiplient les choix possibles de présentation aux internautes. Cette labilité du livre numérique met au défi l'idée traditionnelle du livre imprimé ; elle en renouvelle potentiellement les usages.

Cette absence de définition précise peut aussi contribuer à renforcer le sentiment de flou qui paraît marquer les accords passés entre Google et les bibliothèques partenaires, et tenant notamment aux conditions dans lesquelles ces bibliothèques pourront ou non utiliser leurs copies des fichiers numérisés à partir de leurs fonds. La deuxième partie de ce rapport s'attache à en analyser la portée.

*
* *

[16] La définition du livre numérique semble un peu plus simple dans le cas où il s'agit de la simple transposition d'un livre imprimé : c'est la piste que semblent avoir choisi les auteurs du rapport Zelnik lorsqu'ils préconisent une extension au livre numérique du système du prix unique existant dans l'univers physique, nécessitant une telle définition (proposition 10 : « Etendre le prix unique du livre aux livres numériques dits homothétiques »).

II. Les accords actuels avec Google : une réponse inadaptée

II.1. Une réponse inadaptée au regard des missions des bibliothèques

En matière de patrimoine numérisé, les missions des bibliothèques publiques sont principalement de deux ordres :

- Assurer la pérennité, à long terme, du patrimoine écrit numérisé - autrement dit la conservation et la mise à jour des fichiers numérisés, dans un conteste d'obsolescence plus ou moins rapide des technologies.
- Favoriser l'accès le plus large possible au patrimoine numérisé. Cela implique, d'une part, une mise à disposition et une visibilité de ce patrimoine numérisé sur Internet, au terme d'un processus de « masse » nécessitant des moyens importants ; d'autre part, un niveau de qualité suffisant des supports et des outils numériques, permettant de répondre à la diversité des usages des internautes.

Les projets d'accords passés entre Google et des bibliothèques[17] **posent, au regard de ces missions, un certain nombre de questions**.

II.1.1. La mission de conservation

L'objectif de conservation du patrimoine numérique est fortement lié à la mission de mise à disposition de ce patrimoine par la bibliothèque. Il suppose, d'une part, que la pérennité des données numérisées soit assurée, et d'autre part, que ces données puissent être stockées et transférées sur différents supports numériques, amenés le cas échéant à évoluer. De la qualité de conservation des données et de leurs supports dépendra la qualité de la mise à disposition et la possibilité de multiplier les formes de cette mise à disposition.

Les accords existants montrent une réelle insuffisance des accords sur ce point, puisque le sujet de la conservation et de la mise à jour éventuelle des fichiers mis à la disposition des bibliothèques par Google n'y est pas abordé – qu'il s'agisse de l'accord concernant la bibliothèque municipale de Lyon ou du projet d'accord (« *memorandum of understanding* ») qui avait été élaboré pour la BnF.

Ces accords ou projets d'accords, s'ils prévoient la remise à la bibliothèque d'une copie du fichier numérisé par Google, ne comportent, en particulier, aucune obligation pour Google de faire bénéficier les fichiers remis à la bibliothèque d'éventuelles innovations qu'il apporterait à ses propres fichiers.

Il revient donc, en l'état actuel des choses, aux bibliothèques de prévoir et de gérer les questions de conservation et de pérennisation des contenus numérisés – ce qui est un poste de coût non négligeable.

En outre, ces accords ne sont pas toujours suffisamment précis sur les fichiers qui seront effectivement transmis par Google. S'agissant de l'accord avec la bibliothèque municipale de Lyon notamment, le cahier des clauses techniques particulières indique seulement la transmission du fichier image et du fichier du texte brut (texte non structuré), sans aucun engagement sur la nature de l'océrisation et les traitements effectués ; il n'est donc pas évident que la bibliothèque puisse ensuite faire les liens nécessaires entre image et texte. Cette imprécision des accords est liée aux restrictions exigées par Google quant aux usages possibles des fichiers par la bibliothèque elle-même, qui ne peut proposer en

[17] On notera à cet égard qu'aucune bibliothèque nationale, à l'exception peut-être de la Bibliothèque de Catalogne qui jouit cependant d'un statut particulier, n'a à ce jour conclu d'accords de numérisation avec Google.

téléchargement que les images des livres, et non les fichiers textes (cf. *infra*). La bibliothèque ne sait donc pas avec suffisamment de précision ce qu'elle va exactement recevoir de son cocontractant, ce qui pose question à la fois en termes de conservation et en termes d'usages des fichiers par la bibliothèque[18].

II.1.2. La mission d'accessibilité

En raison de leur mission consistant à favoriser l'accès au patrimoine numérisé, les bibliothèques doivent pouvoir conserver une liberté réelle de travail sur les fichiers et sur leur utilisation, impliquant la possibilité de contracter différents partenariats.

Or les accords actuels proposés par Google posent un certain nombre de limitations et de clauses d'exclusivité qui paraissent à cet égard excessives, qu'elles soient explicites ou implicites.

■ Il n'est pas anormal qu'un partenaire privé ayant pris à sa charge la numérisation de collections bénéficie de certaines contreparties - notamment d'une exclusivité d'exploitation commerciale des fichiers. Encore faut-il s'assurer que ces contreparties n'affecteront pas la mise en valeur et l'exploitation de ces fichiers par les bibliothèques elles-mêmes.

La première difficulté des accords – en l'occurrence, projet de protocole d'accord avec la BnF et marché conclu par la bibliothèque municipale de Lyon – provient des **formulations imprécises qui sont utilisées** (ou envisagées). Tout d'abord, la portée de l'exclusivité commerciale, si elle est définie pour la bibliothèque, ne l'est pas pour Google, qui ne précise pas nécessairement la manière dont il compte utiliser les fichiers – au mieux figure une stipulation indiquant que les prestations exécutées pourront être rémunérées par les recettes publicitaires tirées de la mise en ligne des fonds numérisés[19]. Ainsi, la bibliothèque ne doit pas faire payer l'accès aux œuvres du domaine public.

Les téléchargements systématiques (« *systematic downloading* ») sont également prohibés, or il s'agit là encore d'une notion assez imprécise : on peut penser qu'elle désigne les pratiques d'opérateurs tiers qui téléchargeraient en masse les fichiers proposés sur le site de la bibliothèque, à des fins de revente notamment. Mais la portée exacte de la notion est peu claire et des usages légitimes pourraient se trouver handicapés par cette restriction si elle n'est pas précisée.

L'autre difficulté tient à des **limitations explicites**, qui peuvent brider les initiatives de la bibliothèque pour renforcer l'accessibilité à son patrimoine numérisé. Ainsi, la bibliothèque ne peut partager ou fournir le contenu numérisé à une tierce partie sans avoir obtenu préalablement l'autorisation de Google : si l'idée générale de la clause paraît assez logique (les fichiers numérisés gratuitement par Google ne peuvent être cédés à un concurrent), cette stipulation peut cependant handicaper des projets en cours avec d'autres partenaires, y compris publics. On relèvera que, lors des discussions avec la BnF, Google avait envoyé une lettre indiquant expressément son accord pour que les fichiers numérisés par ses soins soient reversés dans Gallica et répertoriés sur le portail Europeana.

■ Les accords passés par Google prévoient toujours que **les autres moteurs de recherche ne pourront pas accéder aux fichiers numérisés par lui pour les indexer et les référencer[20]**. Autrement dit, cette exclusivité se traduit, concrètement, par l'absence d'indexation et de référencement du texte des livres par d'autres moteurs de recherche. Seules les métadonnées, généralement produites

[18] Cf., sur ces questions, l'annexe 3 précitée (pp. 47 et suivantes).

[19] Cf. article 6-2 du contrat entre la bibliothèque municipale de Lyon et Google.

[20] Dans tous les contrats connus, une clause stipule en effet que la bibliothèque prendra les mesures techniques appropriées pour empêcher « l'accès automatisé » aux fichiers livrés par Google.

par les bibliothèques partenaires, sont accessibles aux moteurs, ce qui réduit considérablement la visibilité sur Internet des fichiers exploités par les bibliothèques et fait peser un lourd handicap sur les bibliothèques numériques que celles-ci pourraient vouloir développer de façon autonome. On peut comprendre les motivations de Google, qui prend à sa charge financièrement et techniquement les opérations de numérisation, et souhaite, en contrepartie, bénéficier d'une exclusivité sur ce contenu numérisé, lui permettant d'étendre sa base de recherche d'indexation et de rémunération. Mais cela revient aussi à permettre à un acteur en position dominante sur le « marché » de la recherche d'information et de l'accès aux contenus numériques de renforcer cette position dominante.

Il faut par ailleurs s'interroger sur l'indexation et le référencement, par Google, des fichiers des bibliothèques qui n'ont ou n'auront pas été numérisés par lui. Cette question n'est pas abordée dans les accords. Google se réserve la pleine propriété des fichiers numériques résultant de la coopération proposée, sans qu'il soit fait référence en contrepartie aux autres fichiers numérisés par les bibliothèques, les uns et les autres vivant, en quelque sorte, « leur vie propre ». Or le partenariat avec la bibliothèque municipale de Lyon et surtout le projet de protocole d'accord avec la BnF ne portent que sur une partie minoritaire du fonds patrimonial - le choix des ouvrages à numériser, qui revient aux partenaires publics, est nécessairement limité aux ouvrages relevant du domaine public et pouvant supporter le processus de numérisation. Se pose dès lors la question de leur intégration éventuelle dans le mode de consultation et de requêtes mis en œuvre par Google Livres. Il est dommage que les accords n'envisagent pas, dans le même temps, les aspects de référencement et d'indexation dans Google Livres, la bibliothèque devant alors effectuer des démarches supplémentaires pour assurer le référencement de ses fichiers propres dans Google Livres et dans le moteur Google.

■ **La durée des clauses d'exclusivité** est également excessive : des durées de plus de vingt ans, qui semblent aujourd'hui la référence dans les contrats passés avec la société Google[21], sont extrêmement longues, notamment à l'ère d'Internet, et peuvent aller à l'encontre de la mission d'accès impartie aux bibliothèques.

■ Le **niveau de qualité minimum de numérisation** n'a pas non plus été défini avec précision. Or pour les bibliothèques, il doit nécessairement être élevé et clairement connu. La question des possibilités et des outils de recherche sur les fichiers numérisés qui seront exploités par les bibliothèques est en effet fondamentale au regard des usages et de leurs évolutions[22]. La pratique qui s'est dégagée est semble-t-il de proposer le fichier numérisé à la bibliothèque, qui peut l'accepter ou le refuser – en cas de refus, le processus d'océrisation (reconnaissance de caractère) est repris, et un nouveau fichier est alors à nouveau proposé. Mais aucun seuil minimum de qualité n'a été fixé, ni aucune pénalité prévue au cas où un tel seuil ne serait pas atteint.

■ Enfin, la **confidentialité des accords** revendiquée par la société Google, outre qu'il s'agissait d'un principe difficilement acceptable s'agissant de bibliothèques publiques[23], n'a pas contribué à lever les doutes sur le bilan coûts/avantages de tels accords pour les bibliothèques.

II.2. Au regard de l'articulation entre logique privée et logique publique

II.2.1. Une prise en compte insuffisante des atouts des bibliothèques

Les bibliothèques françaises - en particulier la BnF - disposent d'atouts importants, qu'il ne faut pas

[21] Le marché passé avec la bibliothèque municipale de Lyon prévoit une exclusivité commerciale de 25 ans.

[22] Cf. Annexe 3 précitée, notamment pp. 48-52.

[23] Les pièces du marché passé avec la bibliothèque municipale de Lyon n'ont été rendues publiques qu'après une démarche volontaire du rédacteur en chef de la revue *Livres-Hebdo* auprès de la Commission d'accès aux documents administratifs.

sous-estimer dans le cadre d'une négociation avec un partenaire privé.

Le premier de ces atouts est celui du **fonds**, très complet grâce à l'apport du dépôt légal dans le cas de la BnF. La BnF dispose également de **métadonnées bibliographiques** déjà constituées à travers un catalogue multimédia riche de plus de 10 millions de notices de documents. Ce catalogue est le produit de près de deux siècles d'effort de catalogage par des professionnels expérimentés et de plus de vingt années de conversion rétrospective dont le coût pourrait être évalué avec précision et qui dépasse certainement une dizaine de millions d'euros. Les données les plus récentes sont rédigées dans un format bibliographique très détaillé dont la valeur est unanimement reconnue.

Par ailleurs, le catalogue de la BnF s'appuie également sur des métadonnées d'autorité (5 millions de notices) décrivant avec une assez grande précision les auteurs et les sujets des documents présents dans la bibliothèque et représentatifs de l'ensemble de la production nationale puisque issus pour une grande partie du dépôt légal (70.000 livres et 40.000 titres de périodiques, plusieurs dizaines de milliers de documents spécialisés en 2008). Cet ensemble est en soi d'une haute valeur puisqu'il permet notamment de distinguer les auteurs les uns des autres et de contribuer le cas échéant à une meilleure gestion des droits à l'heure où la question des œuvres orphelines se pose avec une certaine acuité. On comprend l'intérêt de Google pour ces métadonnées d'autorité dans le projet de protocole d'accord envisagé avec la BnF à l'été 2009.

La BnF dispose également d'un **savoir-faire en matière de numérisation de masse**. Avec la mise en place des premiers marchés de masse en 2007, l'établissement public a réussi à passer d'un rythme de 5.000 documents auparavant numérisés chaque année à un rythme d'environ 100.000 documents dont environ 40.000 livres en 2009. Cette expérience a permis de préciser les étapes-clés du processus de numérisation et les exigences à avoir à l'égard des prestataires, d'identifier les points de blocage et freins potentiels, de définir progressivement les niveaux de qualité les plus souhaitables, de spécialiser un certain nombre d'agents dans les différents domaines liés au processus numérique, depuis la chaîne de numérisation proprement dite jusque, dans une moindre mesure, aux phases de diffusion et de conservation des documents numériques ainsi produits. Cette première expérience a donné des résultats tangibles avec une bibliothèque numérique de plus de 950.000 documents disponibles en ligne en décembre 2009.

Enfin, on peut considérer que l'établissement bénéficie d'une **« marque » reconnue**, au niveau mondial, et qu'il peut exercer un **effet d'entraînement réel** sur les autres bibliothèques européennes, notamment via Europeana.

Or les accords semblent peu prendre en compte ces aspects – à l'exception du projet concernant la BnF, où une clause spécifique prévoyait que la société Google reverse à la bibliothèque les fichiers d'œuvres francophones du domaine public numérisées à partir des fonds des autres bibliothèques signataires. Cet aspect d'échanges de fonds est particulièrement intéressant et peut conduire à envisager d'autres types de partenariats (cf. III). La bibliothèque municipale de Lyon a, quant à elle, mis l'accent sur la mise en place par Google d'une interface sur le site de la bibliothèque (« *hosted solution* »), afin que les fichiers numérisés y soient accessibles, en attendant de construire à terme sa propre bibliothèque numérique à partir des fichiers récupérés auprès de son partenaire – mais l'autonomie de cette bibliothèque ne sera acquise qu'à l'issue de la période d'exclusivité de 25 ans prévue par l'accord.

II.2.2. Une négociation délicate du fait du positionnement bien particulier de Google

■ Toute négociation avec Google est de surcroît particulièrement délicate en raison de la **position dominante qu'occupe le moteur de recherche**. Certes, en droit de la concurrence, seul l'abus de

position dominante est sanctionné. Il n'en reste pas moins que les objectifs stratégiques de Google doivent être pris en compte dans l'appréciation des engagements pris.

Google a intérêt à préserver la position de son moteur, dans un univers en constante évolution. À cet égard, la numérisation de masse de livres présente un très grand intérêt puisqu'elle permet au moteur de recherche **d'accroître sa base d'indexation**, tout en s'assurant de l'exclusivité des contenus numérisés, sur une période longue. Il s'agit d'une forme de recherche « d'intégration verticale » entre l'outil de recherche et les contenus indexés, non plus uniquement à partir des ressources disponibles librement sur le web, mais également de contenus propres et exclusifs, conférant un avantage différentiel.

Le positionnement de Google l'incite donc à rechercher une exclusivité d'utilisation sur les fichiers numérisés à partir des fonds des bibliothèques - seules les bibliothèques elles-mêmes peuvent aussi en assurer une exploitation, mais limitée. Or les ressources qui sont numérisées proviennent de collections publiques, et relèvent du domaine public. Dans ces conditions, alors même que la numérisation est assurée par le partenaire privé, il est difficile d'accepter des clauses d'exclusivité longues. En d'autres termes, d'autres acteurs privés devraient pouvoir accéder aux mêmes ressources. Bien entendu, les livres « papier » n'entrent pas dans les clauses d'exclusivité, d'autres acteurs pouvant les re-numériser s'ils le souhaitent (et dans la mesure où l'état de conservation des ouvrages permettrait à nouveau de les numériser). Mais les fichiers numérisés devraient eux-mêmes pouvoir faire l'objet d'autres utilisations, afin que des initiatives alternatives ou complémentaires puissent surgir. Il paraît donc essentiel de s'assurer que le contrat ne bride pas les initiatives des autres acteurs privés.

À cet égard, est souvent utilisée la **notion de « facilité essentielle »** : cette notion, issue du droit de la concurrence, est souvent employée dans le domaine des réseaux (chemin de fer, télécommunications, transport d'électricité). Les caractéristiques d'une facilité essentielle sont les suivantes : elle n'est pas interchangeable ou substituable ; le coût de mise en place d'une infrastructure équivalente serait prohibitif, en termes d'argent et/ou de temps. L'accès à cette facilité est donc indispensable pour les opérateurs du secteur, et celui qui la détient se trouve, de fait, en situation de monopole ou de position dominante. Appliquée aux livres numérisés, cette notion justifie l'obligation de mise à disposition des fichiers pour les autres acteurs.

■ Par ailleurs il s'agit d'une **entreprise dont les méthodes sont contestées.** Il y a tout d'abord, comme on l'a vu, la pratique de la numérisation de contenus sous droits sans autorisation préalable des ayant droits, en arguant de la notion de « *fair use* ». Le projet de règlement transactionnel de classe, s'il était validé par le juge américain, permettrait à l'entreprise de faire valider les acquis de cette pratique pourtant illégale.

Par ailleurs, Google peut aussi être perçu comme une menace sur les questions de stockage et d'utilisation des données personnelles des internautes. L'inquiétude provient de la capacité de l'opérateur à agréger des données éparses pour établir un profil détaillé de millions de personnes (parcours professionnel et personnel, habitudes de consultation d'internet, participation à des forums...). Google a baissé de 18 mois à 9 mois la durée de conservation des données personnelles de ses utilisateurs. Cette durée demeure cependant supérieure au délai de 6 mois pour la conservation de données personnelles par les moteurs de recherche, recommandé en avril 2008 par le G29, un comité réunissant les différentes autorités chargées de la protection des données personnelles (dont la CNIL) dans les pays européens[24]. En outre, la CNIL reproche à Google de ne pas se conformer à la législation française applicable.

[24] Avis du G29 adopté le 4 avril 2008 (PDF)

Ce positionnement particulier rend donc difficile, d'emblée, les négociations. Des solutions peuvent toutefois être proposées - avec ou sans Google - pour numériser les fonds patrimoniaux des bibliothèques françaises et améliorer leur visibilité et leur disponibilité sur la toile. C'est l'objet de la troisième partie de ce rapport.

*
* *

III. Les solutions possibles

La réflexion s'inscrit désormais dans un contexte renouvelé, marqué par des marges de manœuvre réelles pour mener une politique autonome. L'annonce par le Président de la République, dans les priorités du « Grand emprunt », d'une enveloppe spécifique pour la numérisation du patrimoine culturel et notamment des livres, introduit un changement considérable dans la dimension, le rythme et la « philosophie » des projets de numérisation. Elle permet d'envisager une politique de numérisation du patrimoine écrit à la fois ambitieuse et autonome. Elle oblige en contrepartie à trouver une organisation à la mesure de cette volonté, s'inscrivant dans le cadre de coordination mis en place au niveau du ministère de la Culture et de la Communication, et s'appuyant sur des partenariats public-privé avec les éditeurs ou, le cas échéant, avec certains acteurs des réseaux internet.

Ce choix véritablement stratégique change en tout cas la donne pour les bibliothèques françaises, lesquelles se trouveront dans une situation plus équilibrée pour négocier avec des partenaires privés. Elles pourront en effet mener une réflexion en propre sur leur politique de numérisation, plutôt que de dépendre de propositions extérieures qui ne seraient pas nécessairement adaptées à leurs objectifs. L'autonomie retrouvée devrait également leur permettre de mieux maîtriser leur calendrier de numérisation, ou en tout cas de ne pas dépendre uniquement de celui de grands opérateurs comme la société Google.

L'expérience de numérisation lancée par la BnF avec Gallica est à cet égard un atout, grâce au savoir faire acquis en la matière et à l'existence de fonds d'ores et déjà numérisés d'un volume non négligeable. Si les débuts ont été assez laborieux, le site permet en effet d'accéder aujourd'hui à plus de 950.000 documents, dont 145.000 livres (cf. *supra*, I.1.2.). Son alimentation a trouvé un rythme conforme aux objectifs stratégiques initiaux, avec le passage d'un volume de numérisation de 5.000 documents par an avant 2007 à 100.000 documents par an en 2009. La mise en œuvre des moyens financiers importants dégagés par la décision du chef de l'État doit cependant s'accompagner d'un changement d'échelle et de rythme.

Les objectifs de cette politique de numérisation doivent être définis clairement dès maintenant. Deux objectifs généraux nous paraissent devoir être soulignés d'emblée.

Le premier objectif est d'éviter le risque d'une segmentation du patrimoine, **en se donnant l'ambition d'une numérisation exhaustive, ou en tout cas la plus large possible, des ouvrages libres de droits et sous droits**. Le débat entre numérisation sélective ou numérisation de masse a en effet été tranché par les usages observés, qui témoignent de la confiance des internautes dans la « neutralité » des moteurs de recherche et de la capacité des acteurs culturels à proposer progressivement leurs contenus sur le web. Une large partie des requêtes repose sur l'interrogation de bases larges par un moteur simple à utiliser. La grande majorité des internautes, notamment le grand public, n'attend pas véritablement une « éditorialisation des archives » ; elle veut trouver ce qui l'intéresse. Si l'usage simple du moteur de recherche peut ensuite être combiné avec d'autres modes d'accès, plus structurés, il n'en reste pas moins que l'usage dominant aujourd'hui est celui de l'interrogation *via* une requête simple, fondée sur un ou quelques mots-clés. Une numérisation exhaustive est donc bien un des objectifs à rechercher, en sus des structurations documentaires pour des publics plus spécialisés par les institutions publiques.

Cet objectif est en phase avec la vocation historique de la BnF, attributaire du dépôt légal, et détenant à ce titre un fonds de référence. L'établissement a d'ores et déjà commencé à prendre en compte cet objectif et entamé le processus de numérisation, tout en étalant ses étapes. Après identification des

grands domaines prioritaires et des séries thématiques, la sélection des documents à numériser s'effectue aujourd'hui principalement sur des critères matériels (œuvre hors droits, publication en France, état de l'ouvrage, format), en lien avec l'objectif d'une numérisation de masse.

Cependant, compte tenu de l'ampleur du patrimoine en cause (les collections de la BnF représentent environ 35 millions de documents dont 11 à 12 millions de livres ; parmi ces derniers, 5 millions sont entrés dans le domaine public), il s'agit bien d'un objectif de long terme, d'autant que les « objets » à numériser sont hétérogènes, et que seul un certain pourcentage de livres sont en état de supporter un processus de numérisation de masse – en l'état actuel des techniques. Il conviendrait de s'interroger explicitement sur l'incidence qu'a l'état matériel des collections sur la numérisation réalisée *in fine* afin que les ouvrages en moins bon état, souvent les plus demandés, ne soient pas absents de la bibliothèque numérique.

Le second objectif porte sur **la place du patrimoine français écrit sur l'internet**.. Il est aujourd'hui principalement visible *via* Google Livres, grâce aux fonds francophones numérisés des bibliothèques étrangères, qui ne sont pas complets. Les fonds de Gallica ne sont en revanche que difficilement accessibles lorsque l'internaute averti ne se rend pas d'abord sur le site de Gallica. **Il conviendra donc de veiller à ne pas numériser pour numériser, mais d'assurer l'accès à ces fonds numérisés, ce qui implique de réfléchir très en amont** à la façon dont les documents pourront être trouvés, c'est-à-dire visibles, sur l'internet. Une réflexion approfondie sur l'ensemble des moyens permettant cette visibilité numérique (référencement, indexation, citations dans des blogs ou des sites communautaires, etc.) devra donc être engagée.

C'est dans un cadre profondément modifié que sera définie la stratégie de numérisation, à trois niveaux :

▪ En premier lieu, l'existence d'une plate-forme telle que **Gallica** permet de s'appuyer sur un outil existant, mais dont les performances sont désormais insuffisantes et dont la dimension coopérative reste à mettre en place, essentiellement vis-à-vis des éditeurs et des bibliothèques partenaires.

▪ En second lieu, la mise en œuvre de **partenariats avec des acteurs privés** (éditeurs, moteurs de recherche, plates-formes de diffusion…) est l'une des conditions d'une bonne mise à disposition des fonds numérisés sur l'internet ; il convient donc de définir les contours de partenariats efficaces et équilibrés qui pourraient être conduits avec des acteurs privés, qu'il s'agisse de Google ou d'autres entreprises.

▪ En troisième lieu enfin, **une nouvelle impulsion européenne** est nécessaire, en coordination avec les autres bibliothèques européennes engagées et en s'appuyant sur le portail culturel commun Europeana.

III. 1. Un outil privilégié qui reste à améliorer : Gallica

III. 1. 1. Aspects institutionnels

■ Un bilan mitigé

Un premier bilan montre les limites de l'organisation actuelle de la BnF et de Gallica. Plusieurs aspects méritent plus particulièrement d'être signalés, qui concernent respectivement les moyens, la coopération avec les autres bibliothèques, la coopération avec les autres acteurs, au premier rang desquels les éditeurs, et enfin le pilotage de Gallica.

a) Sur le plan des moyens, l'engagement du processus de numérisation de masse à la BnF a permis d'identifier des difficultés tenant à l'environnement humain et technique. Ce processus exige en effet de mobiliser des moyens importants :
- en amont, pour identifier et sélectionner les documents, les conditionner, les adresser au prestataire - opérations qui ne peuvent pas être confiées à un partenaire extérieur - et pour produire les métadonnées, notamment dans le cas où les métadonnées existantes ne sont pas satisfaisantes ;
- en aval, pour assurer le « contrôle qualité » des documents numérisés. Une fois les fichiers reçus, l'étape supplémentaire du contrôle de la qualité de la numérisation incombe en principe à la bibliothèque, or les flux peuvent être considérables. Actuellement, ce contrôle est en grande partie automatisé pour mobiliser moins de moyens humains, mais cette solution n'est pas entièrement satisfaisante.

Le surcroît de travail nécessairement lié au processus de numérisation est un élément essentiel à prendre en compte. Pour l'instant, la BnF a fonctionné à effectifs pratiquement constants, pour mener de front la numérisation et les activités courantes de la bibliothèque. Le passage à un rythme supérieur de numérisation reposera donc la question des moyens humains alloués à cette politique dans le cadre du plafond d'emploi de l'établissement. La numérisation a de surcroît longtemps été envisagée comme un projet supplémentaire par rapport aux missions auxquelles l'institution doit répondre. Ce n'est que très récemment que les différents départements de la bibliothèque ont commencé à se l'approprier, avec un tournant marqué depuis 2007, année pendant laquelle la BnF a engagé un marché de « numérisation de masse » pour ses collections de livres imprimés. Cette appropriation par les différents départements en charge des collections explique d'ailleurs le choix d'une relative dilution de l'équipe au sein de ces départements, plutôt que d'un service centralisé en charge de Gallica. Une révision de l'organisation accompagnée du déploiement plus ambitieux de moyens humains affectés à la filière numérique sont impératifs pour répondre au nouvel objectif de numérisation de masse à un rythme soutenu.

b) Sur le plan de la coopération avec les autres bibliothèques publiques, malgré le souhait affirmée de Gallica d'abriter plusieurs collections en sus de celles de la BnF, force est de reconnaître l'absence d'association effective des autres bibliothèques, sinon très marginalement – ce qui, en pratique, se traduit par le faible nombre de ressources en provenance de ces bibliothèques partenaires (moins de 7.000 documents, sur les 900.000 que compte le site aujourd'hui). Si elles commencent à être invitées à participer au projet Gallica, les modes de décision et l'accès aux crédits de numérisation existants demeurent centralisés auprès de la BnF et sont donc peu incitatifs pour les bibliothèques partenaires.

La BnF propose la signature de conventions avec les bibliothèques intéressées, mais exerce un droit de regard approfondi et peu transparent sur le choix des bibliothèques retenues, celui des ouvrages à

numériser, selon une forme de coopération « verticale descendante ». Le rapport de l'Inspection générale des Finances sur la BnF[25], publié en janvier 2009, propose, sur ce point, de partager davantage la stratégie de numérisation avec d'autres institutions, notamment le réseau des « Pôles associés de partage documentaire », qui n'a que très marginalement accès aux crédits de la troisième tranche du marché de numérisation de masse.

c) Si la BnF doit intervenir naturellement comme acteur principal du processus de numérisation, il est certainement nécessaire de réfléchir **au mode d'association des autres partenaires, qu'ils soient publics ou privés,** dans la définition des objectifs, des options techniques et, plus généralement, dans l'harmonisation des choix éditoriaux.

À titre d'illustration, il ne serait pas illogique que pour faciliter la mise en œuvre **d'une chaîne de numérisation de masse, non seulement des œuvres hors droits, mais aussi des œuvres sous droits**, **la BnF soit investie de la mission de numériser également cette dernière catégorie.** En effet, la loi du 1er août 2006 a introduit dans le code du patrimoine (articles L. 132-4, L. 132-5 et L. 132-6) le droit pour l'organisme dépositaire du dépôt légal de numériser, à des fins de conservation et de consultation sur place, les fonds qu'il détient, sans autorisation préalable des déposants. Forte de cette mission, la BnF pourrait numériser en masse les ouvrages qu'elle conserve, y compris ceux qui sont sous droits, dont elle détient à la fois les ouvrages et les métadonnées correspondantes.

Bien entendu, la mise à disposition de ces fonds numérisés ne pourrait se concevoir que dans un **cadre contractuel avec les éditeurs et les représentants des ayant-droits**, dès lors que les dispositions légales n'autorisent la consultation des œuvres du dépôt légal que de manière très restrictive.

L'ensemble ainsi créé constituerait une base numérisée significative et s'avérerait particulièrement intéressant à exploiter, notamment pour ce qui concerne les œuvres épuisées. Les ouvrages que les éditeurs ne souhaiteraient pas nécessairement publier à nouveau sous format papier pourraient ainsi trouver une exploitation nouvelle, rémunérée et non exclusive, sous format numérique[26]. Pour le public, la base consultable s'en trouverait considérablement élargie. Il y aurait donc une logique à la fois juridique, industrielle et de commerciale à ce que la BnF numérise ses fonds libres de droit comme sous droits.

Dans cette perspective, Gallica deviendrait un site d'accès à tout le patrimoine écrit, *via* une plate-forme coopérative respectueuse des droits des différents partenaires, les conditions d'accès étant adaptées au statut de chaque œuvre.

La mission estime que la position de Gallica et son rôle vis-à-vis des autres partenaires doivent être redéfinis, tant sur le plan technique que sur son mode opératoire. Une association large et effective des partenaires au **pilotage de Gallica** est une condition indispensable de l'accélération de la politique de numérisation et de diffusion numérique.

■ Vers une plate-forme coopérative de valorisation des fonds patrimoniaux et des œuvres numérisées

A titre liminaire, il paraît essentiel d'indiquer que **Gallica aurait pour vocation d'être une plate-forme de référence, mais non exclusive**, permettant l'accès du public aux fonds numérisés de l'ensemble de ses partenaires. Les bibliothèques et les éditeurs qui le souhaiteraient pourraient ainsi

[25] Inspection générale des Finances, Rapport sur la Bibliothèque nationale de France, n° 2008-M-065-02, janvier 2009 – voir notamment p. 14 du rapport.

[26] Un tel accès pourrait être monétisable, par exemple soit *via* le renvoi au site de l'éditeur, soit au sein des bibliothèques, sous la forme d'abonnement.

s'adresser à Gallica, tout en conservant bien sûr la possibilité de diffuser par ailleurs leurs contenus numérisés sur tous les sites de leur choix – dans la lignée de ce qui existe aujourd'hui. En revanche, **l'accès au financement public pour la numérisation des livres devrait être subordonné à l'adhésion à Gallica,** c'est-à-dire à l'une ou l'autre au moins de ses fonctionnalités : indexation du contenu, feuilletage d'extraits voire, le cas échéant, commercialisation du fichier, directe (sur Gallica) ou indirecte (par renvoi de Gallica vers un site tiers de vente, choisi par l'éditeur titulaire des droits).

Gallica se propose donc comme une plate-forme de diffusion de référence, dont l'intérêt permettrait une visibilité accrue sur l'internet grâce à l'importance des contenus proposés et à la mutualisation des moyens.

a) C'est dans ce cadre que serait conçu **le partenariat avec les éditeurs** et les représentants des ayants droit afin d'améliorer la présence du corpus francophone sur l'internet.

La BnF pourrait, comme on l'a vu, procéder à la numérisation de masse des ouvrages collectés au titre du dépôt légal. Si elle dispose de la légitimité juridique pour le faire au titre de la conservation, voire de l'accès sur place à ces fonds, le dialogue avec les éditeurs n'en demeure pas moins indispensable à plusieurs niveaux. Il convient de souligner que les éditeurs restent libres de poursuivre leurs propres opérations de numérisation, qui concernent aujourd'hui essentiellement les ouvrages récents. Ils pourraient en outre mener des opérations de numérisation plus pointues ou nécessitant des normes plus exigeantes pour leurs fonds, compte tenu notamment de l'exploitation qu'ils comptent en faire et qui demeure, bien entendu, de leur ressort[27].

Le dialogue nécessaire à établir porterait également sur les modes d'accès au contenu, étant entendu que les livres sous droits ne pourraient pas faire l'objet d'accès gratuit *via* Gallica – sauf dans le cas éventuel d'un accord préalable de l'éditeur.

Outre l'opportunité juridique offerte par la loi du 1er août 2006 en matière de numérisation et de mise à disposition des fonds numérisés, on devra également recourir à une démarche contractuelle où la numérisation des livres sous droits sera précédée d'une discussion préalable avec les éditeurs sur la répartition des tâches et des charges à assumer par chaque partie en fonction des usages et des modes d'exploitation finalement retenus.

Enfin, par son positionnement à la frontière du champ patrimonial et du secteur sous droits, Gallica serait naturellement appelée **à jouer un rôle de plate-forme de diffusion et de valorisation des œuvres orphelines, à partir du moment où le code de la propriété intellectuelle permettra les utilisations numériques de ces documents**. Une recommandation du Conseil supérieur de la propriété littéraire et artistique d'avril 2008 prône en effet la mise en place d'une gestion collective obligatoire des œuvres orphelines de l'écrit et de l'image fixe afin de les rendre plus aisément disponibles aux réutilisations notamment sur le web. Le 30 septembre 2009, le ministre de la culture a chargé ses services de proposer une suite législative à cette recommandation.

Afin que Gallica puisse devenir le lieu d'une diffusion efficace de ces livres sous droits mais sans ayants droit identifiés, dans le respect des droits associés à ces œuvres, il apparaît extrêmement souhaitable que les éditeurs soient associés au pilotage même de Gallica.

b) S'agissant des bibliothèques, le processus de numérisation lui-même (choix des ouvrages, volume de la numérisation…) pourra continuer à relever des bibliothèques partenaires.

[27] Et feraient l'objet soit d'un financement propre, soit des fonds spécifiques du Centre national du Livre pour les projets de numérisation « pointus ».

La BnF devra cependant proposer toutefois à ceux qui le souhaiteront des solutions communes, en leur ouvrant notamment ses propres marchés de numérisation – dans la lignée, par exemple, de l'expérimentation en cours avec les bibliothèques d'Alençon, de Compiègne, de Lyon, de Cujas et de l'Institut national d'histoire de l'art (trois bibliothèques municipales et deux bibliothèques universitaires), qui a permis de faire bénéficier ces bibliothèques d'un marché de la BnF tout en décentralisant certains aspects du processus de numérisation. Ainsi, les fonctions de numérisation en tant que telles et surtout de stockage des fichiers numériques pourraient être partagées, rien n'empêchant les bibliothèques partenaires de les assurer elles-mêmes si elles le souhaitent.

■ Quelle gouvernance ?

L'organisation à retenir doit s'inscrire dans le cadre qui sera défini au sein du ministère de la Culture et de la Communication pour piloter l'ensemble du processus de numérisation du patrimoine culturel et, plus largement, devra tenir compter du rôle du commissariat en charge du suivi du Grand Emprunt qui sera chargée de piloter l'ensemble des actions.

Les propositions de la mission conduisent à retenir plusieurs niveaux d'action :

- la numérisation des livres (choix des ouvrages, pilotage de la chaîne de numérisation, passation des contrats de prestation nécessaires, production de fichiers…) ;
- le stockage et la maintenance de ces mêmes fichiers ;
- la mise à disposition au profit des internautes (conception de la plate-forme, choix des techniques d'indexation…) ;
- les exploitations commerciales proprement dites, gratuites ou payantes, directes ou indirectes ;
- les partenariats avec les opérateurs susceptibles de développer l'accès aux œuvres nationales.

a) Il appartiendra à la structure de coordination mise en place au niveau du ministère de la Culture et de la Communication de définir les volumes et les financements associés aux deux premiers niveaux ci-dessus : processus de numérisation et stockage.

b) La mission préconise, en revanche, que les choix des formats et normes techniques, ainsi que les procédures et partenariats à concevoir, le soient au sein de la structure de référence réorganisée, Gallica, sous réserve de leurs approbations par la structure de coordination du ministère et le cas échéant, par le commissariat chargé du Grand Emprunt.

Ainsi, l'ensemble des **questions liées à une nécessaire logique d'harmonisation** serait traité au niveau de Gallica[28] :
- l'harmonisation des métadonnées associées aux documents, celles-ci étant un véhicule d'informations essentielles pour la qualité de la recherche, le référencement et l'accès à l'information ;
- l'interopérabilité des fichiers et plus généralement des formats de données échangées (métadonnées et contenus), notamment à travers la définition de standards communs de numérisation, de diffusion et de stockage numérique ;
- la gestion de l'interface de consultation, du moteur et du référencement ;
- le cadre des marchés ayant trait au développement et au fonctionnement de la plate-forme.

L'une des premières missions de Gallica, dans ce cadre renouvelé, sera de **définir ses nouveaux besoins pour accommoder le changement d'échelle de la numérisation**. Ce changement d'échelle

[28] On relèvera à cet égard que plusieurs chartes ont été définies (charte documentaire, charte de numérisation, charte OCR). Ces travaux de réflexion en commun devront être poursuivis.

nécessitera en effet une évolution profonde du fonctionnement de la plate-forme. Les coûts inhérents à ce changement doivent être pris en compte séparément des coûts de la numérisation proprement dite (nouvelles infrastructures, notamment de stockage, développement de nouvelles fonctionnalités, etc.). Elle pourra ensuite préparer un adresser un dossier de demande de financement auprès de la structure de coordination mise en place par le ministère.

Sur le plan stratégique, il semble nécessaire à la mission que Gallica conserve un rôle moteur en matière de **pilotage des accords éventuels avec des partenaires privés** - la logique de tels accords relevant davantage de l'intérêt collectif que de celui de chacun des membres associés à Gallica. En outre, il paraît nécessaire de parler d'une seule voix, par l'intermédiaire d'un interlocuteur ayant une certaine masse critique. Gallica pourrait ainsi être clairement identifiée et bénéficier d'emblée d'une position de négociation et d'un savoir-faire mutualisé en matière de partenariat. Cette fonction motrice n'est cependant envisageable que si le pilotage de Gallica est lui-même redéfini pour y associer de façon beaucoup plus étroite ses partenaires – au premier rang desquels les bibliothèques et les éditeurs.

La question se pose dès lors de savoir si Gallica doit demeurer intégrée au sein de la BnF ou, autre option, s'il est utile de concevoir une entité nouvelle sous la forme, par exemple, d'un groupement d'intérêt public à vocation culturelle.

Quelle que soit la solution retenue, le rôle de la BnF demeurerait important puisqu'elle en serait un partenaire privilégié.

c) La mission préconise en tout cas que l'autonomie de Gallica soit renforcée afin de devenir une structure dédiée, travaillant pour le compte de partenaires d'origine différente et chargée, en toute concertation, de définir les grands choix techniques, d'assurer la diffusion des contenus et de gérer les financements. Cette structure travaillerait **sous l'impulsion d'une instance qui réunirait les différents partenaires de la plate-forme**. Cette instance collégiale serait décisionnelle.

L'instance de pilotage de Gallica serait notamment chargée de définir les grandes évolutions de la plate-forme, les normes et standards techniques de la mise à disposition des fonds patrimoniaux, Il lui reviendrait d'élaborer une politique de visibilité du patrimoine numérisé, de réfléchir à l'interface de consultation, au moteur - et plus généralement à l'amélioration des fonctionnalités de Gallica afin de mieux répondre aux usages et de s'entendre sur « l'expérience » proposée à l'utilisateur. Il s'agirait aussi, le cas échéant, de concevoir une politique de valorisation commerciale, avec l'accord des organismes ou partenaires propriétaires des fichiers.

La coordination générale entre les différents partenaires engagés dans le processus de numérisation serait assurée par la structure mise en place au sein du ministère de la Culture et de la Communication.

III. 1. 2. Améliorer la présence de Gallica et de ses contenus sur l'internet

Les efforts de numérisation doivent s'accompagner d'une volonté de conquête de visibilité sur le web. **Rien ne sert d'être disponible si l'on n'est pas visible**[29]**.**

À cet égard, l'une des premières démarches qui pourrait être conduite serait de **changer le nom de Gallica,** en saisissant pour cela l'occasion du changement de dimension de la politique de numérisation.

Si Gallica est facilement accessible depuis les moteurs de recherche à partir d'une requête avec le nom

[29] Nous renvoyons là encore à l'annexe 3, qui fait une comparaison approfondie entre les fonctionnalités offertes par Gallica et par Google Livres.

« Gallica » (le lien vers le site arrive alors en première réponse, quel que soit le moteur de recherche utilisé), ce n'est pas le cas lorsque l'on interroge le moteur à partir d'un titre ou d'un nom d'auteur. Ainsi, une recherche sur « le Rouge et le Noir », présent dans les collections numériques hors droits de Gallica, **ne laisse apparaître aucun résultat en provenance de Gallica**. La première occurrence de l'ouvrage numérisé vient de Google Livres, qui propose l'accès à l'exemplaire numérisé de la bibliothèque de l'université de Californie (édition de 1866). Une recherche associant les termes « le Rouge et le Noir » et « Gallica » ne renvoie pas non plus directement à des résultats de recherche directement issus du site, mais permet cependant un accès indirect *via* des liens à partir de blogs d'internautes, ce qui montre que le site a été utilement repéré.

La politique de visibilité du site et de « dissémination » de ses contenus hors droits sur l'internet a commencé, mais elle doit être poursuivie et accentuée afin que ces contenus soient mieux repérés par les moteurs de recherche.

■ Référencement et indexation

Les équipes la BnF viennent d'engager une réflexion approfondie pour accroître la notoriété de Gallica, en privilégiant notamment trois formes d'actions : la multiplication des accès, depuis la base, à des contenus variés (stratégie dite de « liens fins »), l'amélioration du signalement et du référencement, et un meilleur accès pris en compte par les moteurs de recherche des métadonnées et de l'indexation de l'ensemble des contenus (indexation « plein texte »).

Cette politique portant à la fois sur un meilleur référencement du site et sur l'indexation des contenus afin qu'ils soient facilement accessibles *via* des requêtes sur les moteurs de recherche doit être poursuivie et approfondie.

Les moteurs de recherche ne peuvent pas accéder à des pages à contenu dynamique, qui ne se créent que sur requête d'un internaute, à partir d'une interface de recherche propre à Gallica. L'attention portée au nommage des pages, la personnalisation des URL, l'amélioration du référencement naturel, le chaînage des pages entre elles, la création de « pages d'atterrissage », l'utilisation des métadonnées, voire l'achat de mots clé peuvent être cités parmi les techniques à disposition de Gallica pour que ses ressources apparaissent mieux dans l'univers numérique.

Ces actions doivent à la fois permettre une indexation des contenus par ces moteurs de recherche et la consultation complète de l'ouvrage par un lien pointant sur le site Gallica.

■ De nouvelles fonctionnalités, davantage orientées vers les aspects participatifs et communautaires

Les aspects participatifs (« wiki ») ne doivent pas être sous-estimés. Ils peuvent permettre d'enrichir utilement le site, et notamment les métadonnées, de développer des communautés au sein de Gallica et d'en accroître la notoriété.

Les internautes peuvent en effet contribuer à enrichir une ressource (par exemple une photo ou un ouvrage) par des liens, des ressources associées (par exemple des thèses sur l'ouvrage ou l'auteur) ou des commentaires. Les compétences des bibliothécaires trouvent là aussi matière à expansion dans le territoire du numérique avec des possibilités de création de labels (par exemple, comme dans une bibliothèque physique, « les usuels ») faisant ressortir l'intérêt d'un ouvrage. Le développement et l'harmonisation des métadonnées peuvent également faire l'objet d'un prolongement par le développement d'une démarche sémantique, laquelle peut à son tour être complétée par les internautes.

Gallica peut par ailleurs offrir aux communautés - d'intérêts communs ou de recherche - des fonctionnalités d'étiquetage, de partage, d'enrichissement. Si les tentatives d'organisation des communautés de la Toile par des institutions sont souvent vouées à l'échec, la proposition régulière de nouvelles fonctionnalités simples d'utilisation mais innovantes peut permettre à Gallica de remplir dans le monde numérique sa mission d'accueil des passionnés et des chercheurs.

III. 1. 3. Améliorer le service rendu par Gallica

■ Un moteur amélioré

Gallica réfléchit actuellement à l'amélioration de son moteur de recherche. Le moteur utilisé aujourd'hui (Lucene) est dédié à la recherche plein texte : il opère sur un index pré-constitué, unique, et n'interagit avec aucune autre base de données. Si le choix du moteur pouvait se comprendre au regard des volumes traités, il devra être profondément repensé dans la perspective d'un accès de masse[30].

D'une part, le moteur doit évoluer dans la mesure où Gallica veut adapter ses outils documentaires en y adossant des outils d'analyse sémantique plus poussée. D'autre part, une réflexion sur l'ergonomie de la recherche doit continuer à être menée, afin de mieux paramétrer le moteur en fonction des attentes en termes d'usages. **Autrement dit, l'ergonomie doit être revue tout autant pour l'internaute qui viserait une recherche rapide, « en 1 clic », que pour le chercheur souhaitant affiner le périmètre de sa recherche grâce à des outils de tri, de filtrage, de positionnement dans un champ précis de recherche.** Ce dernier élément est un élément important de distinction à l'égard de Google Livres. Mais il faudra également prendre en compte impérativement la facilité d'accès au contenu.

Le changement d'échelle de la numérisation rend cette problématique d'autant plus urgente : les choix technologiques actuels, notamment en matière de moteur, ne paraissent pas pouvoir accommoder pareille évolution. Cet aspect est stratégique, la performance du moteur étant une condition indispensable pour exister à côté de plates-formes telles que Google Livres.

Cela peut passer soit par une extension du déploiement du moteur actuel, soit par l'adoption d'un autre moteur de recherche plus performant.

Ce projet est d'ores et déjà identifié comme prioritaire par les équipes techniques de Gallica.

■ Des métadonnées normalisées

Les métadonnées sont essentielles pour l'accès aux informations sur les contenus numérisés. Il est donc d'autant plus important d'éviter les erreurs et de corriger les omissions éventuelles[31]. Aujourd'hui, les bibliothèques numériques comme Gallica et Google Livres utilisent la même norme internationale de base (le « Dublin Core »), ce qui garantit l'interopérabilité des métadonnées provenance de ressources documentaires diverses.

Mais l'usage d'une norme homogène ne garantit pas la qualité. Le caractère contraignant de cette norme peut en effet inciter les bibliothèques numériques à simplifier les métadonnées dont elles disposent, ou à rassembler dans une seule et même catégorie (« champ ») des métadonnées plus finement structurées. La norme ne peut pas, par ailleurs, pallier les défauts qui se trouveraient dans des métadonnées sources.

30 Cf. Annexe 3, pp. 55 et suivantes.

31 Cf. Annexe 3, pp. 53-55.

Celles-ci sont structurées par champs (« auteur », « titre », « éditeur », « source », etc.). Si, par erreur ou par souci de simplification, certains éléments figurent dans le mauvais champ (par exemple si le nom de l'auteur figure dans le champ « titre »), l'inadéquation du contenu et du contenant rendra moins pertinents les résultats de la recherche. Ces erreurs nuisent à la création de liens entre les livres (livres du même auteur par exemple) et à la qualité des actions qui peuvent être menées sur des résultats de recherche (classement ou filtre par auteur, notamment).

Les sources d'informations des bibliothèques numériques étant multiples et hétérogènes, la normalisation des métadonnées reste donc problématique dans les chantiers de numérisation de masse. Cette normalisation passe par l'usage de listes d'autorités unifiées, le redressement des notices bibliographiques et l'exercice d'un contrôle qualité approfondi sur les métadonnées associées aux fonds numérisés.

Sont en jeu la pertinence des résultats de recherche pour l'internaute, la possibilité de classer les informations recherchées - qui sont en lien avec le gain de pertinence sémantique des moteurs de recherche - mais aussi le développement du « web sémantique ». Ces enjeux relèvent pleinement du périmètre d'activité des institutions de conservation et de valorisation des fonds patrimoniaux, qui doivent s'y impliquer significativement.

III. 2. Conditions d'un partenariat équilibré avec des acteurs privés

Les développements précédents ne doivent pas laisser à penser que les grandes institutions publiques pourront mener leur politique de numérisation sans recherche d'éventuels partenariats avec des acteurs privés opérant sur les réseaux internet. De tels partenariats peuvent avoir un effet très positif, non seulement sur le plan technologique, dans un domaine en évolution constante et rapide, mais également en termes de masse critique de contenus notamment, si un tel partenariat était mis en place avec un acteur ayant déjà pris des initiatives importantes en matière de numérisation de contenus.

De tels partenariats devront cependant respecter un certain nombre de conditions essentielles et garantir à la fois l'équilibre et la réciprocité entre les parties. La BnF doit à cet égard jouer un rôle exemplaire, lié notamment au fait qu'elle est attributaire du dépôt légal.

III. 2. 1. Objectifs et conditions préalables

■ Objectifs

L'objectif principal est d'assurer ***une large visibilité de notre patrimoine sur l'internet et son appropriation***, à la fois en recourant aux méthodes spécifiques d'indexation et de référencement propres à l'univers numérique, mais également en tirant parti des complémentarités de contenus, afin d'atteindre un effet de masse critique dans l'univers francophone en ligne.

Le premier aspect suppose de ne pas écarter des partenariats avec des acteurs du type moteurs de recherche, comme Google mais aussi comme Bing, le moteur développé par Microsoft, ou encore Yahoo! Les nouvelles générations de moteurs s'orientent vers la présentation de contenus de plus en plus structurés. Une plate-forme de contenus ordonnés comme l'est Gallica peut, à ce titre, devenir un partenaire très attractif pour ces moteurs.

Le second aspect intéressera davantage les acteurs ayant déjà numérisé des contenus – au premier rang desquels figurent les éditeurs mais aussi Google Livres. Le partenariat avec les éditeurs devrait faire

l'objet d'une coopération renforcée, comme indiqué précédemment. C'est donc essentiellement sur le dernier aspect que le rapport se concentre – la coopération avec des plates-formes de contenus de type Google Livres - mais il est bien entendu qu'il n'est pas exclusif des précédents.

L'autre objectif est ***la mise à disposition la plus large possible des œuvres du domaine public*** qui auront été numérisées. Parce que ces œuvres relèvent du domaine public, leur contenu peut, en quelque sorte, être considéré comme une « infrastructure essentielle »au sens du droit de la concurrence. À ce titre, l'utilisation des fichiers numériques de ces œuvres ne doit pas être bridée par des clauses d'exclusivité – s'agissant du moissonnage des différents robots des moteurs existants ou des usages des bibliothèques dépositaires de ces contenus libres de droit.

■ Conditions préalables

Deux conditions doivent avoir été préalablement remplies par le cocontractant éventuel de Gallica.

La première est le nécessaire ***respect du droit d'auteur***. Il en va du respect de la loi française et de la loyauté à l'égard des différents partenaires de la plate-forme Gallica
La seconde exigence est relative aux conditions dans lesquelles le cocontractant gèrera les informations et ***données personnelles des internautes***. Il paraît essentiel que soient définies préalablement les conditions – notamment les conditions de durée – dans lesquelles ces données seront conservées, voire utilisées. Le type de consultation – et, en l'occurrence, de lectures – d'un internaute est en effet une donnée sensible, qui exige que soient prises certaines précautions *ex ante* en matière de non-divulgation et de conservation limitée de la donnée.

III. 2. 2. « Un livre pour un livre » : une proposition de partenariat fondée sur l'échange de fichiers numérisés

■ Les principes

Les accords jusqu'ici passés ou proposés aux bibliothèques en matière de numérisation de masse, essentiellement par l'acteur majeur qu'est Google dans ce domaine, ont consisté à faire prendre en charge la numérisation (scan et « océrisation » des ouvrages) par l'acteur privé, ce dernier mettant ensuite à disposition de la bibliothèque une copie du fichier numérisé, avec toutefois certaines restrictions d'utilisation pendant une durée excessivement large.

Pour éviter de telles clauses d'exclusivité sur des œuvres du domaine public, et permettre à la bibliothèque d'être propriétaire à part entière de ses fichiers numériques, il paraît nécessaire que ce soit elle qui procède à ces opérations ou qui les finance. C'est donc sur un autre terrain qu'un partenariat avec un acteur disposant déjà de contenus numérisés doit se fonder.

Un accord avec Google par exemple - ou, plus précisément, avec sa plate-forme de contenus Google Livres - pourrait viser, non pas à faire prendre en charge l'effort de numérisation mais à le partager, en ***échangeant des fichiers de qualité équivalente et de formats compatibles*** (textes et images).

Ces échanges pourraient être prévus dans des proportions à définir, afin d'éviter les redondances mais aussi de tirer profit des complémentarités (par exemple, diverses éditions d'un même ouvrage…).

Ceci permettrait à Gallica d'atteindre plus vite la masse critique nécessaire en langue française – Google Livres disposant de contenus francophones numérisés *via* les fonds des bibliothèques américaines avec lesquelles il a contracté, et également de fonds francophones en provenance de bibliothèques

européennes (Gand, Lausanne et bientôt Lyon). De son côté, Google Livres pourrait enrichir sa propre base de contenus des collections de référence francophones présentes dans Gallica. Cet échange pourrait être complété par l'amélioration réciproque des métadonnées et des notices respectives. Enfin, l'accord pourrait envisager que la provenance du fichier apparaisse sur le site consulté, par exemple par l'affichage d'un logo ou d'une mention.

Chaque partenaire ***resterait libre de disposer des fichiers obtenus par l'échange***, dans des conditions transparentes et définies à l'avance.

■ Equilibre économique du projet

C'est certainement le point où les vues les plus divergentes peuvent se faire jour : les objectifs de chacune des parties peuvent ne pas coïncider, la construction de projections financières sûres et transparentes peut se heurter à un éventuel souhait du partenaire de ne pas divulguer le détail de ses propres sources de revenus, enfin les valorisations envisageables sur le moyen terme ne sont pas forcément aisément identifiables.

Si cela s'avérait nécessaire, il serait possible de procéder par étapes :

1ère étape : un accord sur les standards des fichiers à échanger et une évaluation de la qualité et de la compatibilité des fichiers sur une plus grande échelle ;

2ème étape : un échange d'une base de fichiers portant sur un nombre donné de documents (de l'ordre de 100.000 ouvrages par exemple) et une évaluation des consultations effectives des internautes ;

3ème étape : un échange sur une base plus large de fichiers et éventuellement de corpus complémentaires, le cas échéant assorti d'un accord sur les conditions d'exploitation commerciale des fichiers échangés.

■ Proposition alternative

Il ne faut pas sous-estimer, toutefois, la complexité que peut entraîner l'échange pur et simple de fichiers sur le plan technique – définition des types de fichiers à livrer, avec conversion éventuelle si les fichiers initiaux ne correspondent pas, lourdeur et coûts de cette conversion (passage en format PDF par exemple). Dans l'hypothèse où ce type d'échange de fichiers ne pourrait se faire aisément, en raison par exemple, d'incompatibilités entre les fichiers à échanger et la plate-forme d'accueil ou les usages que souhaite en faire la bibliothèque, un autre type de partenariat pourrait être envisagé, qui aurait l'avantage de respecter l'objectif de visibilité des fonds numérisés.

Les moteurs de recherche voient les pages de livres comme des pages « *web* », il est donc nécessaire qu'elles soient structurées de manière à pouvoir être indexées facilement par ces moteurs. Le processus de numérisation que Google met en œuvre intègre cette structuration spécifique. Il pourrait ainsi être intéressant d'envisager la création d'***une filière de numérisation partagée*** – une sorte de joint-venture – qui permettrait et à Google et à la bibliothèque de réaliser leurs opérations de numérisation au même moment, chacun selon leur procédé, mais en partageant les équipes et en ne mobilisant qu'une seule fois les ouvrages à numériser. On pourrait notamment envisager que cette filière commune permette de scanner une seule fois les documents, l'atelier de scan étant pleinement partagé. Autrement dit, la numérisation « image » pourrait être pleinement mutualisée. En revanche les procédés d'océrisation pourraient rester distincts, afin que chaque partenaire puisse retenir le niveau de qualité et les fonctionnalités nécessaires pour l'intégration harmonieuse sur sa plate-forme.

Certes, les ouvrages seraient océrisés deux fois, mais les coûts seraient répartis, et la bibliothèque serait certaine que les fichiers numérisés par Google seraient traités par lui de façon à être correctement indexés. Or une indexation correcte aide considérablement à retrouver un ouvrage sur la Toile, autrement dit à le rendre accessible et visible. Cette visibilité accrue, de façon quasiment gratuite, est à mettre en balance avec les coûts relativement élevés d'une politique dynamique de mise en ligne efficace sur le web. Enfin, chacun des acteurs conserverait, encore une fois, la maîtrise complète de ses propres fichiers. Il conviendrait d'évaluer l'apport initial de chaque acteur (Google ou la bibliothèque) et de déterminer en fonction du nombre de documents fournis par chacun d'éventuelles compensations.

■ Un partenariat conforme aux objectifs des parties

Une société telle que Google, est-elle susceptible d'être intéressée par un partenariat de cette nature, dans le prolongement du programme engagé avec de nombreuses institutions patrimoniales, aux États-Unis comme en Europe ? Réciproquement, la mise en ligne du fonds littéraire français passe-elle nécessairement par un référencement comme élément de Google Livres ? Subsidiairement, un accord de cette nature peut-il contribuer à accentuer le déséquilibre au profit de Google vis-à-vis de ses concurrents ?

On peut observer que ce type d'accord répondrait aux objectifs officiels de Google : mettre à la disposition des internautes un accès simplifié et homogène à la majorité des œuvres numérisées, et cela sans coût direct pour cette société, contrairement aux accords actuels avec les bibliothèques qui sont onéreux pour Google. Il autoriserait également une exploitation commerciale, une fois déterminées les règles de partage de revenus selon le statut des œuvres. Il améliorerait en outre l'importance de l'offre. Enfin, il pourrait donner accès à des métadonnées et des notices incomparables et compenserait à cet égard les faiblesses de Google Livres sur ce point.

Ce type de partenariat est également intéressant pour la visibilité des livres non anglophones sur la Toile. L'offre numérique accrue et la multiplication des accès possibles à ces contenus (directement, par accès à Gallica ou *via* un moteur de recherche), répond à l'objectif central de diffusion du patrimoine écrit numérisé. Et pour chaque partenaire associé à Gallica, n'est-ce pas une valorisation complémentaire de ses actions de numérisation et une plus grande visibilité du site propre géré par ce partenaire, s'il en a conçu un ?

Demeure cette évidence qu'un accord n'est que la rencontre de deux volontés et que l'on ne peut préjuger des avancées à venir.

III. 3. Rechercher une impulsion nouvelle au niveau européen

En Europe, un certain nombre de bibliothèques de statuts divers ont signé des accords avec Google (cf. Annexe 4), d'autres sont intéressées à nouer un partenariat, le plus souvent pour bénéficier des financements nécessaires à une numérisation de masse, mais aussi par souci d'être présente de manière efficace sur l'internet - et cela pour des ouvrages dans des langues de large audience, comme pour des langues moins diffusées mais référencées par Google.

La mise en œuvre des projets français de numérisation du patrimoine écrit risque par ailleurs de créer un déséquilibre en Europe, même si les instruments existent (notamment Europeana) et si les concertations sont souvent étroites.

La mission a reçu plusieurs dirigeants d'autres bibliothèques européennes. Elle s'est efforcée de mieux analyser les enjeux et de percevoir les doutes, comme les engagements réels, de nos partenaires. Il lui

paraît indispensable de développer les bases d'une coopération active susceptible de faire contrepoids aux initiatives des grands acteurs privés du secteur.

Le défi est triple :
- comment financer une numérisation de masse en Europe ?
- comment procéder, selon quels principes, quelle technologie et quelles normes ?
- Europeana est-il une alternative crédible, de qualité potentielle équivalente, à Google Livres ?

Autant de questions auxquelles doit répondre une politique européenne, qui pourrait s'appuyer sur trois axes : la mutualisation des actions des bibliothèques, la relance d'Europeana et la définition d'une charte commune pour les partenariats publics/privés.

III. 3. 1. Mutualiser les actions des bibliothèques

■ Les bases sur lesquelles fonder une action européenne

En matière d'acquis « individuels » des bibliothèques, tout d'abord, nombre d'entre elles ont déjà engagé des actions de numérisation et développé, à ce titre, un savoir-faire.

En matière de volonté d'agir ensemble, ensuite, on peut souligner que la prise de conscience de l'enjeu de la numérisation des fonds patrimoniaux est réelle, que de nombreuses bibliothèques participent d'ores et déjà à divers projets et programmes communautaires (comme le PCRD, parmi les programmes relativement anciens, ou le projet plus récent ARROW, soutenu par le programme européen e-Content +, pour les œuvres orphelines par exemple), enfin qu'un intérêt partagé pour les partenariats public-privé se fait jour, illustré par le projet de charte des bibliothèques publiques sur les conditions de partenariats public-privé (en cours d'élaboration par la Conference of European National Librarians – CENL).

■ La création d'un réseau de bibliothèques volontaires

À l'instar des initiatives prises dans le domaine des médias, pourrait être créé **un réseau de bibliothèques volontaires adoptant un processus commun de numérisation. Celui-ci pourrait se traduire à plusieurs niveaux :**
- la mise en place, pour celles qui le souhaiteraient, de **centres communs de numérisation** - l'objectif étant d'harmoniser les méthodes et les fichiers afin de les rendre interopérables, et de partager les coûts des chaînes de numérisation – en mettant en commun leur expertise et leurs moyens, les bibliothèques pourraient ainsi bénéficier d'une alternative aux propositions de Google, sans avoir à accepter de clauses d'exclusivité ;
- la mise en place, le cas échéant, de **centres communs de stockage numérique**, permettant là encore une mutualisation des coûts ;
- **l'échange de fichiers numériques** entre les bibliothèques de ce réseau. Dans ce cadre, Gallica pourrait proposer à ses partenaires qui le souhaiteraient d'héberger et de diffuser leurs fichiers numériques – étendant ainsi aux bibliothèques européennes une partie de ce qu'il propose à ses partenaires français. La provenance des contenus serait systématiquement identifiée, soit par la mise en évidence de logos, soit encore par la mise en place d'une interface adaptée à la consultation de ces fonds étrangers et rendant transparente la diffusion *via* Gallica. Autrement dit, Gallica hébergerait en ce cas les contenus mais ceux-ci seraient consultables dans leur environnement spécifique, choisi par la bibliothèque d'origine. **Et ce, sans exclusivité.** Une forte incitation à la coopération d'autres bibliothèques nationales européennes en la matière pourrait résulter des larges fonds étrangers conservés à la BnF. À ce titre, la numérisation des collections ne doit peut-être plus se limiter aux seuls

ouvrages initialement publiés en France.

D'autres plates-formes existantes ou à créer pourraient également jouer ce rôle en Europe. Ces plates-formes échangeraient entre elles leurs fichiers respectifs et établiraient les liens susceptibles de faciliter la consultation et le moissonnage de leurs fichiers.

III. 3. 2. Faire évoluer Europeana

L'accélération de la numérisation et l'extension corrélative des contenus de Gallica ne pourront que renforcer Europeana, dont l'alimentation dépend des Etats et des institutions qui en sont à l'origine. Outre cet aspect « quantitatif », des réflexions qualitatives pourraient être menées.

Europeana devrait devenir un lieu d'échanges. Trois axes de coopération pourraient ainsi être développés, à travers ce lieu d'échanges :

- coopération **sur les métadonnées, les formats et les modes d'indexation**, afin d'actualiser en permanence les options communes aux partenaires du réseau mentionné ci-dessus. Ce travail d'harmonisation serait directement utile à l'alimentation du portail ;
- aide aux bibliothèques pour qu'elles puissent passer entre elles des accords d'échanges de fichiers ;
- coordination de la réflexion sur les questions des œuvres orphelines et des œuvres épuisées. L'intégration du programme Arrow va d'ailleurs dans ce sens.

S'agissant des fonctionnalités actuelles d'Europeana, le passage à des modes de recherche plein texte, au moins sur les contenus écrits, améliorerait sensiblement la qualité de la navigation sur le portail : l'utilisateur verrait les résultats de recherche devenir plus pertinents. Il convient en outre de trouver les moyens de l'indexation du plein texte par les moteurs de recherche, indexation garantissant une visibilité accrue du portail et de ses contenus sur la toile.

Enfin, devrait être poursuivie une réflexion sur les modes de financements d'Europeana et la possibilité pour certains projets de numérisation, notamment ceux impliquant des États aux ressources limitées, de recevoir des financements du budget de l'Union européenne plus importants que ce que prévoit actuellement le modèle communautaire.

III. 3. 3. Une charte commune des partenariats publics/privés

Les bibliothèques membres de la Conference of European National Librarians (CENL) réfléchissent actuellement à un projet de charte européenne en matière de partenariats de numérisation avec des sociétés privées. La mission se félicite de cette initiative, qui doit être soutenue et poursuivie.

Elle suggère, à ce titre, qu'un certain nombre de points soient abordés dans la charte : conditions préalables, limitation des clauses d'exclusivité et de durée, clauses permettant la visibilité et la mise çà jour des fichiers.

■ Conditions préalables

Les préalables à la mise en place du partenariat devraient être de deux ordres. D'une part, la reconnaissance de la nécessité d'un accord sur les droits d'auteur et sur les modes de commercialisation

des livres sous droits. Autrement dit, le ou les partenaires privés devront respecter les législations européennes en matière de droit d'auteur.

D'autre part, le respect d'un certain nombre de règles relatives à l'utilisation des données personnelles des utilisateurs, afin de protéger ces derniers d'utilisations abusives.

■ Clauses d'exclusivité et de durée

Si le partenaire privé prend à sa charge la numérisation des livres, les clauses d'exclusivités doivent être les plus limitées possibles et ne devraient pas interdire l'accès aux fichiers par des entreprises concurrentes, dès lors qu'elles seraient disposées à payer un droit d'accès ou une redevance[32]. Par ailleurs, les mesures techniques interdisant l'indexation du texte par les moteurs de recherche doivent être catégoriquement refusées.

L'accord devra par ailleurs prévoir la possibilité d'accéder librement aux fichiers depuis le portail Europeana, voire depuis les sites des autres institutions européennes partenaires.

Afin de ne pas brider les usages potentiels, il conviendra par ailleurs d'être vigilant quant à la qualité des fichiers remis par le partenaire, en s'efforçant de définir cette qualité *ex ante* mais aussi les modalités de vérification *ex post*.

Enfin, les clauses d'exclusivité devraient pouvoir être rediscutées périodiquement, et les accords devraient permettre une éventuelle sortie anticipée dans un certain nombre de cas.

■ Clauses permettant la visibilité et la mise à jour des fichiers

La visibilité des contenus est une condition indispensable de leur existence sur l'internet et de leur accessibilité. Aussi de telles clauses sont-elles particulièrement importantes. L'objectif est d'aboutir à un traitement équitable, par le partenaire, de ses propres fichiers et de ceux numérisés par les bibliothèques elles-mêmes, dans les mécanismes de recherche et d'affichage.

S'agissant de la qualité des fichiers livrés à la bibliothèque, il conviendra de prévoir que des fichiers image et des fichiers texte soient remis, dans une qualité comparable à celle utilisée par le partenaire privé. Les références permettant de faire le lien entre ces deux types de fichiers doivent également être fournies afin d'en permettre une bonne exploitation par la bibliothèque.

L'accord doit également permettre la mise à jour réciproque des fichiers numérisés, au moins pendant la durée de l'accord. Si le partenaire fait évoluer ses techniques d'océrisation, il devra en faire bénéficier les fichiers qu'il a déjà numérisés par la bibliothèque. Réciproquement, si la bibliothèque modifie les bases techniques de ses fichiers, elle devra en donner le bénéfice au partenaire.

*
* *

L'accélération du processus de numérisation des livres est un impératif. Il exige de trouver de nouveaux modes d'organisation, en faisant de Gallica une véritable plate-forme de coopération avec les bibliothèques et les éditeurs partenaires. À cette première strate, qui constitue une forme aboutie de partenariat public-privé, pourraient s'ajouter de nouvelles formes d'accords équilibrés avec d'autres acteurs privés comme les moteurs de recherche ou les plates-formes d'ouvrages numérisés comme Google Livres. Enfin, une réflexion approfondie se développe au niveau européen, qu'il convient

[32] On peut relever à cet égard que le Settlement américain entre Google et les éditeurs prévoit des dispositions de ce type.

d’appuyer et de prolonger.

Synthèse des conclusions / Résumé exécutif

La perspective de numérisation de masse des ouvrages imprimés - quel que soit leur statut, hors droits ou sous droits - constitue une chance pour le rayonnement de la culture française. Elle conduira cependant à une révision en profondeur des politiques publiques dans le domaine de l'écrit, notamment vis-à-vis du grand public qui aura accès aux ouvrages, selon des modalités totalement nouvelles, pour ses recherches personnelles, comme pour ses activités quotidiennes. Tel est le principal enjeu, politique et culturel, du processus qui s'engage.

La numérisation de masse a ses contraintes et ses limites :

- contraintes, liées d'une part au mode d'organisation industriel qu'elle suppose, d'autre part aux exigences de qualité à définir, aussi bien au niveau des normes retenues qu'à celui de la fiabilité des fichiers numériques, de leur indexation et de leur conservation ;

- limites, dès lors qu'elle ne doit pas occulter les autres modes de numérisation, d'indexation et de recherche requis par les publics spécialisés (chercheurs, universitaires, professionnels…) ou les publics empêchés (déficients visuels…).

Le vaste programme engagé au niveau mondial par la société Google a donné une impulsion décisive. Toute politique publique, aussi ambitieuse soit-elle, ne peut cependant ignorer l'avance prise par cet opérateur privé, tant au niveau mondial qu'en Europe. Certains procédés utilisés par cette entreprise, notamment dans ses rapports avec les bibliothèques publiques et les éditeurs, prêtent néanmoins à critique. À cet égard, les limites imposées à la diffusion des fichiers reçus par les bibliothèques, la durée des clauses d'exclusivité commerciale, l'imprécision des choix techniques retenus et la confidentialité des contrats passés avec les bibliothèques sont des conditions difficilement acceptables en l'état, notamment pour une bibliothèque nationale. Il importe donc pour la France de conserver la maîtrise du processus de numérisation et surtout de diffusion des contenus numériques.

La décision du Président de la République sur le financement des politiques de numérisation crée les bases d'une politique nationale, autonome et respectueuse des droits de toutes les parties concernées. Encore faut-il que chaque institution, qu'elle soit publique ou privée, fasse l'effort nécessaire pour contribuer à l'émergence d'un pôle francophone susceptible de se comparer aux plates-formes de recherche et de commercialisation mises en place par les groupes mondiaux, tels que Google, Microsoft, Amazon… C'est bien toute l'organisation de la filière de numérisation qui doit être conçue à l'aune de cette ambition. Ce changement d'échelle aura en effet des implications importantes, non seulement quantitatives mais également qualitatives : il imposera de repenser la puissance des infrastructures, du fait des volumes à traiter, mais également l'évolution des métiers et l'acquisition de compétences nouvelles.

*
* *

La mission propose trois pistes d'action, qui ne sont pas exclusives les unes des autres :
- partir de l'outil existant, Gallica, mis en place par la BnF, en réformant profondément son pilotage et ses fonctionnalités ;
- proposer à la société Google une autre forme de partenariat, fondé sur l'échange équilibré de fichiers numérisés, sans clause d'exclusivité ;

- relancer l'impulsion européenne, à la fois au niveau des autres bibliothèques européennes engagées et du portail culturel commun Europeana.

*

* *

La mission tient à souligner que la priorité doit être donnée à la valorisation du patrimoine écrit sur les réseaux numériques et à sa visibilité pour les internautes du monde entier. Les autres étapes du processus (chaîne de numérisation, stockage et maintenance des fichiers numériques), également importantes et, en tout cas, les plus onéreuses, peuvent rester gérées par les institutions existantes, et notamment par la BnF. Il serait même souhaitable à cet égard que l'on puisse tirer profit de la concentration physique, dans les locaux de la BnF, d'ouvrages réunis au titre du dépôt légal, pour engager un processus de numérisation de masse, non seulement pour le patrimoine hors droits, mais aussi pour les livres épuisés ou orphelins ; sous réserve, bien entendu, de la signature préalable d'une convention cadre entre les éditeurs, les ayant-droits et les pouvoirs publics sur les conditions de cette numérisation et, bien sûr, l'exploitation ultérieure.

L'essentiel pour la mission est la mise en place d'une entité coopérative réunissant les bibliothèques publiques patrimoniales et les éditeurs, dans une logique de partenariat public-privé. Elle devra aussi faire place aux ayants droit et aux autres acteurs de la chaîne du livre.

Cette entité aurait la responsabilité de concevoir, mettre en place et exploiter une plate-forme commune où l'ensemble des ouvrages pourraient être accessibles aux recherches des internautes et, si souhaité, pourraient être feuilletés. Cette entité coopérative aurait la responsabilité d'organiser l'accès aux ouvrages et de concevoir les interfaces avec d'autres plates-formes, telles que les sites communautaires, les moteurs de recherche, les sites de commercialisation… Enfin, par son positionnement à la frontière du champ patrimonial et du secteur sous droits, elle serait appelée à jouer un rôle de plate-forme de diffusion et de valorisation des œuvres orphelines lorsque le code de la propriété intellectuelle permettra les utilisations numériques de ces documents.

À cet effet, la mission est d'avis de mettre en chantier une réforme profonde, prolongeant le travail de qualité qui a permis à la BnF, via Gallica [33], de prendre position sur le réseau internet, mais en changeant radicalement les perspectives actuelles. L'ambition doit être de constituer une base d'ouvrages numérisés de langue française de qualité comparable à celle de Google Book pour la langue anglaise. C'est à cette condition que la France pourra éviter un face-à-face trop déséquilibré avec les entreprises de dimension mondiale et jouer d'un effet d'exemplarité en Europe.

La mission n'avait pas vocation à définir dans le détail le statut juridique d'une telle entité coopérative et son positionnement vis-à-vis de la BnF. Elle souligne toutefois que plusieurs conditions doivent être réunies : collégialité du processus de décision et transparence des modalités de commercialisation qui doivent être autorisées par chaque partenaire concerné.

Le monde de l'internet est organisé sur une base multipolaire. C'est pourquoi chaque partenaire doit pouvoir conserver son propre site, voire organiser directement la commercialisation des ouvrages dont il est titulaire, en fonction de leur statut. Cependant, pour avoir accès aux concours financiers publics à la numérisation, chaque partenaire devra se soumettre à plusieurs obligations :

- déposer ses fichiers sur la plate-forme coopérative,

- adopter des formats et des normes techniques compatibles avec ceux définis par cette même plate-forme.

33 La mission préconise un changement de dénomination.

- déléguer à la plate-forme les droits permettant l'indexation et le feuilletage des fichiers par le grand public, voire, le cas échéant, autoriser l'exploitation commerciale de certains de ces fichiers sur la plate-forme, sans préjudice d'exploitations commerciales par les éditeurs eux-mêmes, *via* les portails de vente de leurs choix ;

*
* *

Les perspectives offertes par les partenariats public/privé s'en trouveront considérablement améliorées. D'accords par nature souvent déséquilibrés, puisque financés par la seule partie privée, on passera dorénavant à des formules « d'échanges réciproques », enrichissant les bases documentaires des deux parties, évitant les doublons dans le processus de numérisation et favorisant la visibilité du corpus francophone. À titre d'illustration, les ouvrages français seraient ainsi largement référencés dans Google Livres, tandis que la plate-forme nationale serait enrichie par l'inclusion d'ouvrages déjà numérisés par Google, notamment ceux disponibles dans les fonds des bibliothèques étrangères partenaires.

La mission s'est efforcée de définir le cadre éventuel de tels accords réciproques et les conditions préalables aux mandats éventuels qui seraient donnés aux négociateurs, notamment en ce qui concerne les droits d'auteurs, la protection de la vie privée et la limitation, de la portée comme de la durée, des clauses d'exclusivité éventuelles.

Dans le cas d'une éventuelle négociation avec la société Google, la mission n'est pas en mesure d'affirmer la disponibilité de cette dernière de discuter dans les termes ainsi développés, en dépit de l'intérêt évident pour cette entreprise d'élargir sa base dans le corpus francophone. Une offre de négociation faite au nom de la France aurait toutefois un double mérite :

- mettre en évidence aux yeux de nos partenaires européens notre volonté de rechercher des partenariats équilibrés, avec cette entreprise dont la présence en Europe est forte,

- en cas de refus, mettre l'autre partie dans la position d'avoir à mieux préciser sa stratégie et ses objectifs concernant le patrimoine écrit européen.

*
* *

Les ressources financières engagées conforteront la position avancée de la France en Europe en matière de numérisation du patrimoine écrit. Il est essentiel de tirer profit de cette situation pour créer un effet d'entraînement, soit au niveau de l'Union et des États membres, soit à celui des Institutions publiques ou privées, en préconisant, pour ces dernières, une mise en commun des moyens techniques et de la plate-forme développée par la France. La société Google est en mesure aujourd'hui de proposer un cadre éprouvé à des conditions économiques sans équivalent. Les initiatives prises en France doivent permettre de combler cet écart et de créer une alternative au profit de l'ensemble des partenaires européens.

La mission propose de retenir trois axes :

- faire bénéficier les partenaires européens des économies d'échelle réalisées à l'occasion des investissements français, en proposant aux bibliothèques européennes intéressées de mutualiser, éventuellement via la plate-forme coopérative et la BnF, l'ensemble du processus de numérisation,

- poursuivre et sans doute approfondir Europeana (à terme),

- élaborer une charte encadrant les partenariats public-privé dans le prolongement des discussions déjà engagées entre plusieurs grandes bibliothèques nationales.

*
* *

La visibilité du corpus écrit francophone sur les réseaux est un objectif majeur qui nécessite de mettre en commun des compétences et des expériences extrêmement variées.

L'aval, c'est-à-dire les conditions d'un accès ouvert, et l'amont, à savoir la qualité de la numérisation et la pertinence des modes d'indexation, doivent être mis sur le même plan. D'aucuns pensent que Google privilégie l'aval au détriment de la qualité, tandis que, au contraire, les bibliothèques publiques viseraient avant tout à valoriser leur savoir-faire en amont, au détriment des modes de consultation de masse. Comme souvent, ces analyses sont caricaturales. Elles n'en témoignent pas moins de l'ardente obligation qui est la nôtre de concevoir un mode d'organisation et de partenariat qui préserve cet équilibre. C'est ce que la mission s'est efforcée d'imaginer.

Liste des annexes

Annexe 1 : Lettre de mission

Liberté Égalité Fraternité
République Française

Ministère de la Culture et de la Communication

Le Ministre

21 OCT. 2009

Monsieur Marc TESSIER
Directeur général
Vidéo Futur
Entertainment Group
27 rue d'Orléans
92200 Neuilly-sur-Seine

CC/6434

Monsieur le Directeur général,

Depuis 2005, la Bibliothèque nationale de France a entrepris de faire de la Bibliothèque numérique *Gallica* un outil majeur de diffusion des contenus francophones et européens sur le Web, notamment à travers sa participation au projet *Europeana* soutenu par la Commission européenne et dont la France est aujourd'hui le premier contributeur.

Cet objectif a impliqué pour la BnF la mise en oeuvre d'une politique de numérisation de masse particulièrement ambitieuse, fortement soutenue par l'Etat à travers les financements accordés depuis 2007 par le Centre national du Livre. Aujourd'hui, la Bibliothèque numérique *Gallica* propose plus de 850 000 documents en ligne dont environ 140 000 livres.

Animés par le souci de proposer aux internautes une offre diversifiée et contemporaine, le Ministère de la culture et de la communication et la BnF ont également développé un partenariat avec plus d'une centaine d'éditeurs français permettant via *Gallica* la consultation d'œuvres sous droits. Ce partenariat est effectif depuis 2008 et plus de 10 000 livres numérisés sont accessibles dans ce cadre.

Cette politique de numérisation implique cependant des financements de l'Etat très importants (plus de 25 M€ affectés au projet depuis son origine) alors que de nouveaux chantiers de numérisation sont en train d'être lancés ou envisagés par la BnF : collections spécialisées, livres rares, presse, audiovisuel... Pour cette raison, un récent rapport de l'Inspection générale des finances sur la Bibliothèque nationale de France a ainsi conclu à ce sujet que l'établissement public « devrait également envisager de renouer un dialogue équilibré et exigeant avec Google -ou d'autres partenaires privés- et sortir de la logique de concurrence actuelle, dans la perspective d'un éventuel financement partagé de la numérisation de masse d'ouvrages hors droits ».

Comme vous le savez, plusieurs articles parus dans la presse ces dernières semaines, certains sur un ton polémique, ont fait état des contacts pris entre la Bibliothèque nationale de France et Google et critiqué l'idée d'un partenariat entre la BnF et cette société américaine.

3, rue de Valois 75033 Paris Cedex 01 France - Téléphone : 01 40 15 80 00

Afin de permettre au gouvernement de défendre au mieux l'intérêt général et d'arrêter la position la plus appropriée, je souhaiterais vous confier une mission d'analyse et d'évaluation de la pertinence d'un tel partenariat. A cet effet, je souhaiterais que vous puissiez préciser :

- les axes de collaboration à ce jour évoqués entre la BnF et Google,
- les points à considérer comme non négociables pour la partie française, qu'il s'agisse de questions techniques, juridiques ou économiques,
- l'opportunité d'un accord entre la BnF et Google, du triple point de vue de la visibilité de la culture et de l'accès aux contenus français sur Internet, de l'intérêt économique et financier pour l'Etat et le contribuable, du message politique à adresser à la communauté internationale,
- les propositions alternatives que vous jugerez pertinentes en matière de partenariat privé, à l'aune de l'ensemble des éléments de contexte que vous aurez pu analyser.

La mission que vous présiderez sera composée de :

- Emmanuel HOOG, président de l'Institut National de l'Audiovisuel,
- Olivier BOSC, conservateur en chef des bibliothèques à la BnF, maître de conférences à l'Institut d'Etudes politiques de Paris,
- Alban CERISIER, directeur des fonds patrimoniaux et du développement numérique aux éditions Gallimard,
- François-Xavier LABARRAQUE, directeur de la stratégie et du développement de Radio-France.

Sophie-Justine LIEBER, maître des requêtes au Conseil d'Etat, sera le rapporteur de vos travaux.

L'ensemble des services du ministère de la culture et de la communication, et notamment la direction du livre et de la lecture, seront à votre disposition pour l'accomplissement de cette mission. Je vous remercie de bien vouloir me remettre votre rapport et vos recommandations d'ici au 15 décembre prochain, avec un rapport d'étape au 25 novembre.

Je vous prie d'agréer, Monsieur le Directeur général, l'expression de mes sentiments les meilleurs.

F. Mitterrand

Frédéric MITTERRAND

Annexe 2 : Liste des personnes auditionnées

Jean-François **Aguesse**, Directeur du développement & Associé, Synodiance

Francis **Amand**, chef de service à la Direction générale de la concurrence, de la consommation et de la répression des fraudes

Patrick **Bazin**, directeur de la Bibliothèque municipale de Lyon

Arnaud **Beaufort**, directeur général adjoint de la Bibliothèque nationale de France, directeur des services et des réseaux

Bernard **Benhamou**, délégué aux usages de l'Internet, ministère de l'enseignement supérieur et de la recherche

Philippe **Colombet**, directeur du programme Recherche de Livres, Google France

Robert **Darnton**, directeur de la Bibliothèque de l'université de Harvard

Olivier **Daube**, Responsable du pôle veille des usages et innovation, Radio France

David **Drummond**, senior vice-président, directeur juridique de Google

Serge **Eyrolles**, Président du Syndicat national de l'Edition

Jeannette **Frey**, directrice de la Bibliothèque cantonale et universitaire de Lausanne

Antoine **Gallimard**, PDG des Editions Gallimard

Emmanuel **Gureghian**, société Bertin Technologies

Theo **Hoffenberg**, PDG de Softissimo

Hervé **Hugueny**, journaliste à Livres Hebdo

Jean-Noël **Jeanneney**, ancien président de la Bibliothèque nationale de France

Dominique **Lahary**, directeur de la bibliothèque départementale de prêt du Val-d'Oise, porte parole de l'Interassociation Archives, Bibliothèques, Documentation

Bernard **Lang**, directeur de recherche à l'Institut national de recherche en informatique et automatique

Pierre **Louette**, président-directeur général de l'AFP

Michel **Marian**, chef de la mission pour l'information scientifique et technique et les réseaux documentaires, ministère de l'enseignement supérieur et de la recherche

Nicolas **Masson**, société Bertin Technologies

Alexandre **Moatti**, ingénieur en chef des Mines, ancien secrétaire général du comité de pilotage de la Bibliothèque numérique européenne

Marc **Mossé**, directeur des affaires publiques et juridiques, Microsoft France

Elisabeth **Niggemann**, directrice générale de la Deutsche Bibliothek, présidente de la Conférence européenne des directeurs des bibliothèques nationales, présidente de la Fondation European Digital Library (Europeana)

Arnaud **Nourry**, PDG de Hachette Livre

Richard **Ovenden** conservateur à la Bibliothèque Bodleienne, Université d'Oxford

Bruno **Patino**, directeur de France Culture

Alain **Pierrot**, société I2S

Michael **Popham**, Université d'Oxford

Bruno **Racine**, Président de la Bibliothèque nationale de France

Stéphane **Ramezi**, Directeur du Multimedia, Radio France

Pascal **Rogard**, directeur général de la Société des auteurs et compositeurs dramatiques (SACD)

Claude **Rubinowicz**, directeur général de l'Agence du Patrimoine immatériel de l'Etat (APIE)

François **Stasse**, conseiller d'Etat, ancien directeur général de la Bibliothèque nationale de France

Sarah **Thomas**, directrice de la bibliothèque Bodleienne et des bibliothèques de l'Université d'Oxford

Jacques **Toubon**, ancien ministre, co-responsable de la Mission Création et Internet

Erreur! Signet non défini.

Annexe 3 : Les enjeux qualitatifs de la numérisation de masse. Réflexions autour de Gallica et de Google Livres

Annexe réalisée par Alban Cerisier

Gallica et Google Livres constituent l'une et l'autre des bibliothèques numériques. Elles conservent et mettent à disposition des publics, sur un site web dédié, des collections de livres numérisés, dont elles ont elles-mêmes commandé et dirigé les opérations de digitalisation.

Elles proposent une recherche sur le texte complet des ouvrages et/ou sur les références qui les décrivent (métadonnées) et permettent d'accéder aux textes eux-mêmes, soit intégralement soit par extrait, soit en ligne soit par voie de téléchargement, soit directement sur la plate-forme soit en renvoyant l'internaute sur d'autres plates-formes de lecture en ligne.

Google Livres affiche par ailleurs une ambition de portail bibliographique, en intégrant dans son corpus des références d'ouvrages dont elle n'a pas encore numérisé ni indexé le contenu (notamment des ouvrages récents).

Les deux applications proposent des fonctionnalités distinctes et évolutives, tant au plan de l'expression et du traitement des requêtes effectuées par l'internaute qu'à celui de la restitution et de la mise à disposition des ouvrages. Ces options techniques et fonctionnelles servent à ce jour des conceptions singulières de valorisation des fonds documentaires numérisés et déterminent des usages différenciés ; et plus généralement, attestent de stratégies de positionnement distinctes de la part des institutions concernées.

On peut en dire de même des modalités de numérisation, tant le scanning en masse d'ouvrages implique, en l'état de l'art, des choix qualitatifs problématiques. Il s'agit donc ici d'envisager les contours fonctionnels et techniques de ces bibliothèques numériques, tant au plan des éléments documentaires produits qu'à celui de leur accès et de leur valorisation.

I. Enjeux qualitatifs de la numérisation de masse
   a. Une numérisation de masse pour une partie du fonds.
   b. Le couple image/texte.
   c. La structuration du livre.
   d. La qualité de l'image scannée.
   e. La reconnaissance de caractères.
   f. Une correction a posteriori ?
   g. Une qualité doublement problématique (conservation et usages).
   h. Un élémentaire principe de précaution.

II. Enjeux qualitatifs de la valorisation
   a. Corpus.
   b. Les métadonnées.
   c. Les moteurs de recherche
   d. Les restitutions
   e. Le feuilletage
   f. Le référencement

I. Les enjeux qualitatifs de la numérisation de masse

• Une numérisation de masse pour une partie du fonds

On doit distinguer les chantiers de numérisation de masse des initiatives plus ponctuelles portant sur des corpus réduits, reposant sur des normes qualitatives élevées et modulables (traitement des images et des textes). Une institution comme la Bibliothèque nationale de France, à l'image des autres bibliothèques européennes, a mené, mène et continuera de mener des opérations ciblées, notamment pour des contenus atypiques (cartes et plans, manuscrits anciens étrangers…). Ces efforts doivent être poursuivis, notamment en raison de l'impossible traitement en masse d'un nombre important de documents présentant des attributs physiques particuliers (dimensions, états de conservation…).

Les chantiers de numérisation de masse tels que ceux engagés par Google Livres et la BNF ces dernières années, consistent, eux, à adresser à une chaîne de production industrielle un nombre très important d'ouvrages à des fins de numérisation sérialisée. Il s'agit de scanner chaque page de chaque exemplaire, sans destruction ni détérioration de l'ouvrage, celui-ci devant être restitué après traitement à l'institution qui le conserve pour réintégration dans les collections. Aucun ouvrage n'est massicoté : les pages sont tournées à la main par un opérateur (plusieurs milliers de pages par jour pour un individu).

Ces chantiers de numérisation de masse ne peuvent être sérieusement envisagés que dans la mesure où la Bibliothèque concernée dispose d'un catalogue informatisé de son fonds, sans quoi la traçabilité des ouvrages entre la Bibliothèque et la chaîne de production, et sur la chaîne elle-même, ne peut être correctement assurée.

Une sélection est effectuée par les équipes des bibliothèques afin de distinguer parmi les candidats à la numérisation ceux qui ne peuvent être retenus au regard de leur état de conservation, de leur caractère précieux et de leur conformité aux critères de normalisation de traitement industriel (Google exclut ainsi de ses chaînes tout ouvrage dont un feuillet excèderait, déplié, les dimensions du bloc imprimé). Au regard de l'expérience des chantiers de numérisation de masse menés jusqu'à ce jour à la BnF, il semble que la part d'ouvrages rejetés soit très élevée.

→ **DES TRAITEMENTS DIFFÉRENCIÉS POUR UNE MISE A DISPOSITION HOMOGENE.** La numérisation de masse, même manuelle, ne concerne donc qu'une partie des collections patrimoniales ; elle est une des composantes de **l'allotissement nécessaire à l'organisation de la conversion numérique des fonds patrimoniaux**. Il faut donc veiller à ce que ces types de traitement différencié en amont n'impliquent pas en aval, au sein de la bibliothèque numérique, des **régimes de mise à disposition hétérogènes aux publics**. C'est un des écueils identifiés dans les partenariats public/privé, et notamment dans les accords de partenariat connus entre Google Livres et les bibliothèques partenaires.

• Le couple image/texte

L'acte de numérisation présente deux volets : il s'agit d'une part de réaliser un scan image de chaque page, couverture y compris ; puis, d'autre part, de déduire de l'image réalisée, après retraitement et optimisation de celle-ci par le recours à des applications *ad hoc*, un fichier texte par la voie d'une reconnaissance automatisée des caractères d'imprimerie (***Optical Character Recognition, OCR***). On dit alors que l'ouvrage a été « océrisé ». Cette étape peut nécessiter le recours à une image intermédiaire, distincte de celle livrée au demandeur, spécialement adaptée à la réalisation de l'OCR.

La « couche texte » déduite de l'image peut être elle-même optimisée, avec des outils de correction spécifiques. Ces deux activités font encore l'objet de travaux de recherche et développement[34]. Mais ils sont couramment utilisés depuis des années, notamment dans le cadre des chaînes graphiques traditionnelles (réimpressions d'ouvrages des fonds éditoriaux).

À l'issue de cette opération, on dispose, pour chaque page du livre, d'un ou plusieurs fichiers

[34] C'est une des raisons de l'acquisition par Google de la société reCaptcha (annoncée en septembre 2009), spécialisée dans la conversion d'images en texte, notamment pour la sécurisation de sites et la protection antispam.

images (selon que l'on a choisi de scanner l'ouvrage en noir et blanc, en niveaux de gris ou en couleurs, voire en définitions plus ou moins grandes) et d'un fichier texte, généralement structuré au format XML, s'appuyant sur un modèle de description des données normalisé. **L'enjeu de la numérisation de masse est de rendre solidaires ces deux « états » du livre.**

Certains textes présentés sur Gallica ne présentent pas ce couple image / texte, mais simplement une image : il s'agit de fichiers appartenant à une génération antérieure de numérisation (avant 2007). Une importante conversion de ceux-ci a déjà été effectuée par la BnF (marché dit des « 60 000 »).

• La structuration du livre

Afin de pouvoir permettre un lien entre l'image de la page et le fichier texte qui en est déduit, les coordonnées graphiques de chaque mot sont mémorisées dans le fichier texte au moment de sa génération, sous forme de balises escamotables. Des balises permettent également de restituer la structuration logique de la page (repérage des zones de haut et de pied de page et des marges par exemple), d'identifier les espaces où se situe une illustration... L'ensemble de ces opérations est rendu possible par le zonage préalable du fichier image et l'utilisation d'un format de description normalisé de type XML. Gallica recommande l'adoption de la norme Alto, qui semble également avoir été retenue pour partie par Google.

Pour restituer le livre dans son intégrité logique, il faut également associer chaque couple de fichier (image/texte) à la page qui lui correspond dans l'ouvrage. Ainsi, le 7e fichier correspondant à un livre ne correspond pas nécessairement à la page portant le numéro de folio 7 dans l'ouvrage (il peut en effet y avoir des pages blanches, des pages non foliotées, des doubles foliotations en chiffres romains et arabes…). Il s'agit donc de constituer un troisième fichier, lui aussi structuré en XML, qui restitue la suite logique du livre et fait le lien entre celle-ci et l'ensemble des fichiers et répertoire qui composent le livre sous sa forme numérisé. C'est la **colonne vertébrale de l'ouvrage**, essentielle pour pouvoir informer l'internaute, par exemple, que telle citation a été trouvée à telle page du livre. On peut également utiliser ce fichier descriptif pour identifier, ou typer, des pages significatives du livre : page du même auteur, page de faux-titre, page de titre, page de table des matières… Cette démarche est décisive, dans la mesure où elle permettra d'exclure certaines pages soit de la restitution du livre, soit du champ d'indexation du moteur de recherche.

Ce **typage** ne semble pas avoir été fait de façon approfondie ni par Google ni par Gallica ; on peut le regretter ; son automatisme est, de fait, problématique. De même on peut s'étonner de ce que les potentialités de la norme Alto n'aient pas été mieux exploitées ; ainsi, par exemple, on aurait pu retirer du périmètre du moteur de recherche l'indexation des titres courants, qui se répètent à chaque page des livres qui en comportent. Il semble que cela n'est pas été toujours le cas dans ces chantiers de masse.

Certaines parties du livre peuvent faire l'objet d'un traitement particulier, à l'image des **tables des matières**. Le repérage de celles-ci dans le livre peut résulter d'un traitement automatique, de même que l'on peut envisager qu'un programme puisse en restituer la structure logique. Il reste que les résultats obtenus à ce jour par de tels traitements sont très imparfaits. L'exécrable qualité des tables de matières proposées par Google Livres le démontre ; elles sont lacunaires et erronées et, partant, le plus souvent inutilisables. Gallica a préféré à ce jour un traitement manuel de ces tables des matières qui, sans être parfait, est bien plus acceptable. Les tables des matières, océrisées, sont systématiquement reprises et corrigées par un opérateur. Cette opération est particulièrement sensible, la structuration des tables des matières pouvant être très complexe et problématique d'un point de vue logique et éditorial. On comprendra que cette étape est essentielle dans la mesure où l'on voudrait imaginer une exploitation d'ouvrages par partie, s'appuyant sur la structuration même du livre imprimé, dans le cadre par exemple d'une consultation ou de téléchargement par chapitres…

• La qualité de l'image scannée

Toutes les étapes de la chaîne doivent faire l'objet de **contrôles de qualité**, dans le cadre de protocoles détaillés définissant les types d'alertes et d'actions à mener en cas d'anomalies constatées dans la chaîne. Il s'agit ainsi, par exemple, de s'assurer que des pages n'ont pas été omises et de définir

une procédure de reprise si c'est le cas.

La qualité de l'image produite reste également un élément sensible. La qualité se joue à plusieurs niveaux : définition de l'image, adaptation de l'image à l'OCR, non détérioration de l'image par l'intervention manuelle… La médiocrité de la numérisation opérée par Google Livres a été maintes fois relevée[35]. Elle montre qu'il convient de renforcer les contrôles automatisés sur la chaîne ; et donc de renforcer les exigences des institutions publiques à l'égard de leur prestataire. C'est aujourd'hui un élément, parmi d'autres, de leur mission de conservation du patrimoine culturel, physique et numérique. À ce titre, le niveau de définition des fichiers produits peut être un élément important dans la campagne. Généralement produit en 300 Dpi, la BNF a récemment exigé de ses fournisseurs une numérisation en 400 Dpi, plus coûteuse et plus volumineuse, mais de nature à donner des résultats plus satisfaisants dans la perspective de déploiement d'un service d'impression à la demande. **Les enjeux de conservation pérenne et d'exploitation sont donc intimement liés.**

• La reconnaissance de caractères

L'autre grand enjeu qualitatif est lié à l'OCR. Celui-ci peut-être plus ou moins satisfaisant selon le mode d'impression de la page, les variations typographiques internes, les alphabets et graphies syllabaires, les langues… Des engagements qualitatifs doivent être pris par les prestataires de numérisation, après examen des corpus candidats à la numérisation ; le taux d'OCR (pourcentage d'erreurs dans l'ouvrage, calculé à partir d'échantillonnages effectués par le prestataire, exclusion faite des cas d'illisibilité), très variable d'un livre à l'autre dans ce type de chantier, doit être transmis à la bibliothèque. Même si le taux de reconnaissance annoncé peut être très élevé (99,9 % par exemple, soit deux fautes grossières possibles par page de 2.000 signes en moyenne), il faut rappeler que ce niveau de qualité serait peu acceptable pour un livre imprimé, dûment établi et corrigé.

Le lien entre l'image et le texte déduit sert à compenser les failles de l'OCR. Si le moteur de recherche indexe la couche texte, c'est bien l'image du livre qui est montrée en premier lieu à l'internaute en réponse à sa recherche. L'image, elle, n'est pas fautive. Grâce aux coordonnées du mot conservées dans le fichier texte, l'application de feuilletage peut surligner facilement l'espace de la page où se trouve la citation ou le mot recherchés. Mais l'index, lui, est toujours fautif. Aussi faut-il considérer qu'en l'état de l'art, toute recherche sur une bibliothèque numérique est, *stricto sensu,* approximative. Elle ne permet pas d'avoir l'assurance que le mot recherché ne se trouve qu'aux endroits signalés par le moteur de recherche. Une réponse fiable à 100 % imposerait une étape, coûteuse et laborieuse, de correction manuelle des fichiers textes obtenus par l'OCR. Ce type de traitement est à ce jour marginal.

• Les livres numériques proposés aux internautes

Les fichiers proposés en visualisation ou en téléchargement par Google Livres et Gallica sont issus de ce travail de préparation. Ils ne peuvent être que de simples images intégrées dans le **feuilleteur**, avec une couche texte invisible pour l'internaute ; ou bien la couche texte elle-même, non corrigée ; ou bien un **PDF image**, généré à partir des différents fichiers du scan (non interrogeable, non indexable : Google Livres) ; ou bien un **PDF texte/image**, généré à partir des scans et du fichier texte de l'OCR (interrogeable, indexable : Gallica) ; ou bien encore un **fichier texte, comme le fichier Epub**. Ce fichier Epub, qui s'est imposé ces dernières années comme un standard pour les livres numériques (et même s'il demeure très imparfait au regard du degré de perfectionnement éditorial du livre d'origine, voire totalement inadapté à celui-ci), se présente comme un répertoire compressé comprenant notamment l'ensemble du texte de l'ouvrage sous formes de différents fichiers. Il est lisible sur les PC (à l'aide du logiciel gratuit Adobe Digital Edition), mais aussi sur la plupart des terminaux mobiles de lecture et les iphones. Il faut préciser que la qualité de ce fichier (et donc sa lisibilité) est étroitement lié à la qualité du traitement de l'OCR. S'il n'a pas fait l'objet d'une révision, il contiendra les mêmes fautes

[35] Il n'est qu'à télécharger un Epub sur Google Livres pour s'en convaincre. Dans le sous-répertoire « Images » du répertoire OEBPS se trouve un florilège de clichés où figurent notamment la main gantée de l'opérateur. Ces images se retrouvent dans le livre électronique, au titre des éléments graphiques non analysables par l'OCR.

de transcription, qui peuvent être extrêmement **pénalisantes** pour le lecteur (c'est bien le cas des quelques Epub consultés dans le cadre de cette enquête depuis le site Google Livres)..

Le **service automatisé de lecture à voix haute**, disponible sur Gallica, s'appuie également sur cette couche texte issue de l'OCR. En conséquence, la qualité du résultat audible est très liée au traitement initial ; en cas d'OCR faible, le texte devient incompréhensible à la lecture. De tels problèmes ne se rencontrent pas si, comme pour les ouvrages récents, on s'appuie sur un fichier texte non pas déduit d'un scan, mais fourni par le compositeur de l'ouvrage imprimé, réputé vierge d'erreurs.

Gallica et Google Livres ne restituent jamais le fichier image source aux internautes, tant en visualisation qu'en téléchargement. Elles génèrent, en aval de la chaîne de numérisation, un fichier beaucoup plus léger, en 75 Dpi. La rapidité de la restitution l'impose ; mais c'est une façon également de **limiter la dissémination des fichiers** adaptés à l'impression à la demande ou, plus généralement, à la commercialisation.

Dans les contrats connus unissant Google Livres aux bibliothèques, le moteur de recherche américain communique une copie des fichiers source de la numérisation à ses partenaires[36], mais ne les autorise pas à délivrer aux internautes un fichier unissant l'image au texte (PDF image/texte, par exemple). La bibliothèque ne peut donc offrir en téléchargement que le PDF image[37], non indexable et non interrogeable (tel que Google Livres le fait sur son propre site actuellement), « à l'unité » et pour un « usage individuel ». C'est une limite importante[38], que peut justifier **la volonté manifeste du moteur américain de ne pas rendre accessible à des tiers concurrents les fichiers numérisés par ses soins**. Cette précaution semble relever de la même préoccupation que celle qui pousse Google à mettre en œuvre des mesures techniques de protection contre les téléchargements massifs sur son propre site et à exiger des bibliothèques partenaires, dans le cadre de leur propre site, des garanties du même ordre[39], voire peut-être plus importantes encore selon l'interprétation des clauses des contrats – et somme toute, très engageantes en termes de responsabilité pour les institutions. Il y a, à cet égard, un risque de déséquilibre entre l'usage que pourraient faire celles-ci de leurs fichiers et celui que s'est réservé Google Livres.

• Une correction *a posteriori* ?

La plupart des failles qualitatives exposées ci-dessus (absence de typage des pages, qualité médiocre de l'OCR, erreur sur les images, tables des matières dégradées) pourraient faire l'objet d'une correction *a posteriori*, en ayant recours à l'intelligence collective d'un réseau (wiki), en refaisant passer les scans d'origine dans une chaîne de traitement OCR fiabilisée ou encore en travaillant sur des solutions de corrections automatisées s'appuyant sur la comparaison des œuvres numérisées à plusieurs reprises.

L'appel à la correction collaborative de l'OCR représente cependant un certain nombre de risques relatifs à l'intégrité des textes, d'autant qu'elle peut impliquer une dissociation progressive du fichier texte par rapport à son référent image. Ce risque, rarement formulé, n'est pas à négliger.

L'évaluation de ces possibilités d'améliorations *a posteriori* reste à faire, même si des initiatives ont déjà été prises en ce domaine sur le Web.

• Une qualité doublement problématique (conservation et usages)

---

[36] Sans signifier toutefois de quel fichier il s'agit précisément, point sur lequel les bibliothèques devraient être bien plus vigilantes ; le CCTP de l'accord de Lyon indique que Google transmet à la Bibliothèque municipale le texte « au format brut et sans enrichissement typographique » - après que l'acte d'engagement prévoie, sans plus de précision, la fourniture d'une « copie des fichiers et des métadonnées numériques créés par le prestataire » (art. 6-1). Qu'est-ce à dire ? S'agit-il vraiment du fichier structuré et coordonné, permettant le lien entre l'image et le texte issus de l'OCR ? Google dispose-t-il en aval d'un fichier plus abouti que celui fourni à son partenaire ?

[37] Article 24 du CCTP de l'accord Google/BM de Lyon.

[38] Elle peut toutefois être relativisée par la mise à disposition en téléchargement des fichiers Epub.

[39] Clause figurant dans les contrats de l'Université de Michigan et de l'Université de Californie : « *University shall implement technological measures […] to restrict automated access to any portion of the University Digital Copy or the portions of the university website on which any portion of the University Digital Copy is available.* »

→ **Conservation.** Les processus actuels de numérisation de masse, malgré leurs progrès récents et les nouveaux usages qu'ils ont ouverts, se caractérisent par un niveau de qualité assez médiocre. L'idée d'une conservation pérenne numérique, au-delà des questions liées à la seule infrastructure technique de stockage et à l'évolutivité des formats informatiques, est remise en cause par l'irrégularité des résultats obtenus en termes d'image, et plus encore de conformité textuelle. Si la notion de copie de conservation est entendue dans le seul sens d'un fac-similé analogique des ouvrages, on peut en effet s'en satisfaire, dans la mesure où le travail de scanning est réalisé avec soin. Si cette même notion s'entend en terme de restitution d'un corpus textuel structuré, on est encore très loin du compte et la marge de progression de la chaîne semble considérable. En termes de traitement textuel, même simple (sans raffinement sémantique), les bibliothèques numériques, reposant essentiellement sur la numérisation de masse, consacrent des **pratiques approximatives**.

→ **Usages.** Ces éléments techniques impactent les usages. Pour des utilisations de type documentaire (chercheurs, étudiants, voire recherche ponctuelle de citations), les bibliothèques numériques sont devenues des outils incontournables, tant par l'indexation des textes que par leur restitution intégrale en mode image. Tout ce qui contribuera à renforcer la fiabilité de l'indexation et à perfectionner le moteur de recherche par un meilleur traitement amont des ouvrages (maîtrise des éléments structurels du livre, optimisation de l'OCR…) sera profitable, et répondra à la mission, non de conservation, mais de communication et de valorisation des bibliothèques patrimoniales. **Celles-ci doivent donc s'impliquer plus avant dans la définition des Engagements de Qualité de Service liés aux grands chantiers de numérisation à venir**, et non se laisser imposer des conditions peu transparentes par un prestataire-partenaire.

Pour un usage relevant d'une lecture usuelle (du type : « je cherche à accéder à une édition en libre accès du *Rouge et le noir* pour le lire sur mon *iphone* »), la technologie actuelle de numérisation en masse ne répond pas au besoin, tant les formats texte produits en sortie (Epub) sont fautifs. Il paraît difficile pour une Bibliothèque nationale, et peu compatible avec sa mission de valorisation des fonds culturels nationaux, de distribuer massivement des textes classiques de la littérature française entachés d'innombrables fautes. Pour ce type d'approche, il conviendrait de proposer un traitement sélectif des ouvrages patrimoniaux qui pourrait faire l'objet d'un traitement plus pointu, approchant les 100 % de taux de reconnaissance de caractère (voir les quelques ebooks exemplaires proposés à ce jour Gallica) ; ou envisager des partenariats avec des éditeurs actuels de ces textes, s'ils existent, pour en proposer des versions librement téléchargeables de qualité, par ailleurs déjà trouvables à ce jour sur internet sur d'autres sites.

• Un élémentaire principe de précaution

→ **Evolutivité des fichiers.** Il est cependant probable que les chaînes d'OCR et les possibilités de corrections a posteriori des textes, par traitement de masse, vont continuer de se perfectionner dans les années à venir. Il est étonnant que les contrats passés entre Google Livres et ses Bibliothèques partenaires ne prévoient jamais, dans la durée d'exclusivité commerciale conférée à Google Livres, l'obligation pour celui-ci de faire bénéficier la Bibliothèque partenaire de ses éventuelles avancées technologiques, par la livraison périodique des fichiers issus d'éventuels retraitements. Faute de quoi, un fossé se creusera entre l'offre de Google Livres et celle proposée par la Bibliothèque partenaire sur son propre site. Il y a, à cet égard, un risque majeur de marginalisation des établissements.

II. Les enjeux qualitatifs de la valorisation

• Corpus

Les deux plates-formes référencent des livres sous droit et des livres du domaine public.

Les modalités d'interrogation et de restitution de ceux-ci dépendent essentiellement de leur statut juridique.

Sur Google Livres comme sur Gallica, les ouvrages considérés comme relevant du domaine public sont intégralement interrogeables et peuvent être lus dans leur intégralité. Ils proviennent dans leur grande majorité de la numérisation d'exemplaires conservés en bibliothèques (bibliothèques associées au projet Gallica ; bibliothèques partenaires pour Google).

Le statut et le traitement des ouvrages sous droit relève de deux scénarios distincts, selon que l'institution a, ou non, passé des accords avec les ayants droit concernés.

En cas d'accord (partenariat Syndicat national de l'édition/BNF pour Gallica, avec le soutien du Centre national du Livre ; « Programme Editeurs » de Google ou encore diffusion de livre sous licences « *Creative Commons* »), la bibliothèque dispose des éléments pour indexer l'ouvrage dans son intégralité et restituer tout ou partie de celui-ci à l'internaute, soit dans son propre feuilleteur (Google) soit dans le feuilleteur indiqué par l'ayant droit (Gallica). C'est l'ayant droit qui fixe, dans l'un et l'autre cas, le nombre de pages librement feuilletables.

En l'absence d'accord, Google Livres s'autorise toutefois à numériser et indexer les ouvrages sous droits à partir d'un exemplaire scanné dans le cadre de son « Programme Bibliothèques ». Elle le rend interrogeable sur sa plate-forme, mais ne donne à voir que trois courts extraits correspondant à la recherche effectuée, sous la forme de *Snippets*, dans la mesure où une juridiction nationale ne s'est pas opposée au principe même d'une telle restitution issue d'une numérisation non préalablement autorisée par le titulaire de droits (Affaire Google/La Martinière/SNE/SGDL).

Notons enfin que Gallica recense également un grand nombre d'estampes et d'images. Si ce choix est parfaitement cohérent à l'égard des collections patrimoniales de la Bibliothèque nationale de France, on peut s'interroger sur l'opportunité d'une approche multi-support à l'égard des attentes des publics, notamment au titre du « bruit » qu'elle peut engendrer dans le cadre d'une recherche simple. Il est à noter que Google, à ce jour, n'a pas retenu ce type d'approche.

• Métadonnées

Les métadonnées sont les éléments de description de l'ouvrage. Google Livres et Gallica s'appuient sur la même norme internationale, le Dublin Core, schéma garantissant l'interopérabilité des métadonnées descriptives de ressources documentaires diverses. Ce modèle conceptuel, simple et efficace, est constitué de quinze éléments, optionnels ou non, répétables, permettant une description formelle, intellectuelle et juridique du document[40].

Google Livres utilise une norme complémentaire de description, permettant d'identifier le niveau de visualisation possible de l'ouvrage (*viewability*), leur portabilité sur des sites tiers (*embeddability*), les commentaires qui lui sont liés (*review*).

Mais l'usage d'une norme homogène ne peut garantir à elle seule la qualité des métadonnées.

Le caractère contraignant de la norme peut, d'une part, obliger les bibliothèques numériques à simplifier les métadonnées dont elles peuvent disposer par ailleurs en amont, ou à rassembler des métadonnées plus finement structurées dans une seule et même catégorie. C'est l'effet entonnoir. Ainsi trouvera-t-on décrits, dans l'élément « format », des éléments de description relevant tantôt de la description de l'ouvrage physique (in-8°, 16 p.), tantôt du fichier numérique (application/PDF), ce qui,

[40] Ainsi l'auteur est-il désigné par l'élément « Creator » ; l'éditeur, par l'élément « Publisher » ; les contributeurs, par l'élément répété « Contributor » ; l'origine du document, par l'élément « Source »…

de fait, ne renvoie pas au même ordre de réalité.

La norme ne peut pas, d'autre part, pallier les défauts des métadonnées sources. On voit ainsi répéter dans l'élément « Title » sur Gallica le nom de l'auteur, repris par ailleurs dans l'élément « Auteur », ainsi que des différents contributeurs de l'ouvrage. L'ensemble des contributeurs d'un titre ne sont pas repérés comme des contributeurs (élément « Contributor ») mais, eux-aussi, de façon indifférenciée, dans l'élément « Titre ». Cette inadéquation du contenant et du contenu contribue à rendre moins pertinents les résultats de recherche et à rendre inopérante la recherche par « Contributeur » et les tris liés à cette catégorie. De la même façon, il apparaît que Google Livres traite dans l'élément titre des informations relatives à la collection ou à la série dans laquelle s'insère l'ouvrage (« 13e volume des Classiques Vaubourdolle ») : ce choix est peu structurant, même si l'on comprend l'intérêt d'une telle souplesse en termes d'efficacité et de rapidité de déploiement.

Les sources d'informations des deux bibliothèques numériques étant multiples et hétérogènes, la normalisation des métadonnées reste problématique dans les chantiers de numérisation de masse[41]. Souhaitable, cette normalisation passe par l'usage de liste d'autorités unifiées, ce qui n'est pas le cas à ce jour. Ce défaut est source d'erreurs et nuit à la création de liens entre les livres (livres du même auteur, par exemple) et à la qualité des actions opérables sur les listes d'ouvrages (classement ou filtre par auteur, notamment). Cette hétérogénéité apparaît clairement dans l'offre sous droit de Gallica, les sources d'information provenant, sans contrôle, des éditeurs et diffuseurs eux-mêmes.

→ **AU CŒUR DU SYSTÈME : LA QUALITÉ DES METADONNÉES.** La mise en œuvre de bibliothèques numériques doit donc impérativement s'accompagner d'un redressement des notices bibliographiques issues de la conversion rétrospective des catalogues papier, afin de faire pleinement bénéficier l'internaute de leur valorisation documentaire. À la mutation des supports correspond nécessairement la mutation des éléments qui les décrivent, dans un triple dessein : l'interopérabilité des catalogues entre bibliothèques numériques, le gain de pertinence sémantique des moteurs de recherche, la mise en œuvre du web sémantique (OWL, RDF/XML du W3C). Ce triple enjeu relève pleinement du périmètre d'activité des institutions de conservation et de valorisation des fonds patrimoniaux. Elles doivent s'y impliquer significativement : c'est leur valeur ajoutée d'aujourd'hui et de demain.

La faiblesse des métadonnées de Google Livres s'explique notamment par le type d'imprécisions détaillées ci-dessus (hétérogénéité de traitement des catégories « rédacteur » ou « auteur » ; mentions d'origines absentes…). Mais les aberrations fréquemment citées par les observateurs, notamment au regard des datations et des catégorisations des ouvrages, sont d'un autre ordre. Elles ne se retrouvent que par exception dans Gallica, dont le référentiel est beaucoup plus fiable, parce que lié dans sa plus grande part au catalogue de la BnF. Gallica sait par exemple beaucoup mieux gérer les publications en série que ne le fait Google Livres.

Il est difficile de déterminer avec certitude si ces erreurs de Google Livres proviennent des métadonnées récoltées auprès de la bibliothèque préalablement à la numérisation, ou bien d'une activité complémentaire d'indexation appartenant au processus même de numérisation[42]. La fréquence de ces erreurs est aussi révélatrice de la fragilité de la chaîne d'indexation de Google Livres que de la faiblesse du contrôle qualité effectué par les bibliothèques partenaires.

De telles faiblesses sur les métadonnées, au-delà des incertitudes documentaires qu'elles provoquent dans le cadre des usages traditionnels des chercheurs, sont constitutives d'un risque juridique. S'agissant d'erreurs de datation, elles sont en effet à l'origine de la mise en accès libre d'ouvrages réputés du domaine public, qui s'avèrent encore être sous droit. La responsabilité de Google Livres et des bibliothèques partenaires pourrait, de ce fait, être engagée ; Gallica prend également de tels risques[43].

---

41 Voir, par exemple, le traitement de l'auteur Théodore de Banville dans Google Livres.

42 Quand, par exemple, la fiche associée à tel *Dictionnaire de la noblesse* de 1775 dans Google Livres décrit un ouvrage sur *Le Génie des eaux chez les Dogons* ; ou que tel exemplaire de la *Revue musicale* d'Henry Prunières et d'André Cœuroy, où il est abondamment question de Maurice Ravel et de Paul Dukas, se voit daté de 1827…

43 Lorsque, par exemple, elle considère à tort comme relevant du domaine public le *Pour Thaelmann* (Éditions universelles,

→ **Contrôle qualité.** On ne saurait trop mettre en avant l'importance du contrôle qualité sur les métadonnées associées aux fonds numérisés. Il relève tant de l'Engagement de Qualité de Service (EQS) fourni par l'opérateur de la numérisation que par de la révision des fonds numérisés assuré par les équipes des institutions. Il apparaît clairement aujourd'hui que les chantiers de numérisation de masse sous-estiment l'importance décisive de ce travail.

D'expérience, on sait que l'accès immédiat aux pages des ouvrages permet de corriger et de compléter *de visu* les métadonnées. L'usager pourrait ainsi se satisfaire d'une qualité moyenne de description des livres auxquels les bibliothèques numériques lui donnent accès. Mais certaines erreurs ne sont pas « compensables » par l'internaute, dans la mesure où elles impactent directement les modalités de traitement des requêtes formulées via le moteur de recherches : on ne peut corriger que ce qu'il est donné de voir.

Il convient de noter que les métadonnées apparaissant dans les notices associées à la présentation des ouvrages sont et seront également utilisés dans les propriétés mêmes des fichiers numérisés, comme éléments descripteurs internes au fichier (propriétés d'un PDF, par exemple). La dissémination des fichiers implique donc, dans le même courant, la dissémination des erreurs d'indexation.

• Les moteurs de recherche

À ce jour, Gallica utilise le moteur de recherches *open source* Lucene (Java, Fondation Apache), dédié à la recherche plein texte. Sur Gallica, Lucene opère ses recherches sur un index unique préconstitué ; il n'interagit avec aucune base de données.

Le choix de Lucene s'avère plutôt opératoire au regard des volumes actuellement traités et des performances visées. Mais il doit être repensé dans la mesure où Gallica veut faire évoluer ses outils documentaires, notamment en y adossant des outils d'analyse sémantique plus poussé. Cela peut passer soit par une extension du déploiement de Lucene, soit par l'adoption d'un autre moteur de recherche. Cet enjeu est aujourd'hui identifié comme prioritaire par les équipes techniques de Gallica. Ce chantier pose également la question de la performance et de sa mesure ; sur des recherches complexes ou portant sur des expressions longues, il apparaît que Gallica peine à remonter des résultats pertinents dans des délais comparables à ceux de Google Livres, bien que le corpus traité par le portail national soit très sensiblement inférieur. Au-delà de la robustesse de l'infrastructure et de la puissance du moteur de Google Livres, cette différenciation peut être aussi perçue comme l'expression de choix documentaires distincts : la mise en œuvre de l'intelligence documentaire (tris, facettes…) n'est pas nécessairement compatible avec la brièveté des temps de réponse. C'est la dialectique classique entre pertinence et performance. Elle deviendra particulièrement sensible avec l'augmentation sensible du corpus numérisé visé par Gallica.

En l'état, l'approche de Gallica n'est pas exempte d'ambigüités majeures, qui nuisent à la compréhension immédiate par l'usager des éléments de réponse fournis par la bibliothèque[44]. Il en ressort un sentiment d'efficacité médiocre de Gallica, alors qu'il ne s'agit que de choix de paramétrage et de développement amendables. Ces effets ont pu nuire à la réputation de Gallica, qui s'est efforcée et s'efforce encore de les corriger.

Google Livres et Gallica proposent deux niveaux de recherche, simple ou avancée, cette dernière

1935), regroupant notamment des discours d'André Gide et d'André Malraux.

[44] Ainsi, dans le cas d'une recherche sur un ensemble de mots (une citation, par exemple : « Au sein de l'infini, nous élançons notre être », issu du *Poème sur le désastre de Lisbonne* de Voltaire), Gallica va remonter les ouvrages pertinents, comme Google Livres peut le faire, mais en ne privilégiant pas la restitution immédiate de la seule phrase. Gallica indique les ouvrages où la citation se trouve ; c'est en choisissant l'un des ouvrages, puis en cliquant sur la page proposée, que l'on aboutit enfin à la citation. Soit au bout de trois clics, contre un seul sur Google Livres. La perception de pertinence par l'internaute est par ailleurs brouillée par le fait que chaque mot présent dans la citation est surligné lorsqu'il est trouvé par le moteur dans les métadonnées et dans l'ouvrage.

permettant classiquement de croiser des recherches sur le texte des livres eux-mêmes avec celles sur les métadonnées, et donc d'affiner en amont les recherches effectuées. La qualité des métadonnées devient dès lors un atout de premier ordre. C'est à ce niveau que l'on peut regretter, comme noté plus haut, qu'une analyse plus fine des structures de la page ou du livre n'ait pas été menée lors du processus de numérisation ou ne soit pas exploitée plus finement par le moteur ; on ne peut ainsi exclure certaines parties des livres qui, pourtant, sont créateurs de bruits à l'indexation. Une numérisation de masse peut permettre ce type d'approches, sans nuire à la productivité de la chaîne.

L'algorithme de recherche de l'une et l'autre des bibliothèques numériques, les lois de pondération qui président aux modalités de présentation des ouvrages dans les listes précédant leur feuilletage, demeurent peu explicites pour les usagers. Il n'est par exemple jamais indiqué clairement combien de fois le mot ou l'expression recherchés ont été trouvés dans les livres proposés. On sait que l'un des principes de base du moteur Google est de mettre en œuvre des coefficients de popularité (fréquence des accès et des liens) et de « confiance » qui, dans le domaine de la recherche sur les livres, peut se traduire par une évaluation du niveau de citation d'un livre par les autres livres conservés dans la bibliothèque numérique. On comprend dès lors que le niveau du corpus rassemblé joue en faveur de la performance du moteur. A ce titre, comme à d'autres, il est difficile d'établir des comparaisons significatives entre le moteur de Gallica et celui de Google Livres.

Pour des types d'usage experts qui ne peuvent se satisfaire des niveaux de pondération implicites, l'absence d'outils de traitement documentaire secondaire de la requête est particulièrement dommageable. C'est, à l'évidence, l'une des grandes faiblesses de Google Livres, partiellement palliée par les performances de ses temps de réponse et sa simplicité d'usage, qui facilite le repérage spontané de l'internaute qui sait ce qu'il cherche. Pour autant, Google Livres laisse parfois l'usager sceptique par ses choix de paramétrages. Une recherche rapide sur « Montesquieu » montre par exemple que Google fait porter principalement sa pondération sur la présence du mot dans la métadonnée « titre » de l'ouvrage et qu'il ne sait pas identifier « Montesquieu » prioritairement comme un auteur. De ce fait, les œuvres de Montesquieu se trouvent isolées parmi un grand nombre de textes divers sur Montesquieu, sans perception de ce qui sous-tend leur ordre de présentation. Nulle restriction par facettes ne permet de sélectionner les notices voulues sur le seul critère de « Montesquieu » comme auteur, sous une seule graphie qui fasse autorité ; ce que Gallica, à l'inverse, comme bien d'autres outils bibliographiques du marché (Electre, par exemple ; ou Cairn), sait faire.

L'opacité du paramétrage du moteur de recherche pose une difficulté majeure, pour des raisons qui, on le sait, ne ressortissent pas qu'à des ordres documentaire et catalographique et renvoient à des problématiques sociétales.

On notera enfin que Google Livres intègre des aides à la recherche similaires au moteur de recherches Web Google, même si celles-ci ne paraissent pas nécessairement très pertinentes (puisque, précisément, s'appuyant sur un corpus Web non structuré et non dédié à l'univers du livre). Les corrections orthographiques automatiques, l'auto-complétion des expressions saisies par l'internaute, semblent implémentées. La distinction entre les deux bibliothèques est donc flagrante sur ce point ; et pour une requête du type « alain rabbe rillet », Google Livres propose de faire une recherche sur « Alain Robbe-Grillet » ; ce que Gallica ne sait pas faire et qui est fort regrettable.

→ **QUEL MOTEUR POUR QUELLE MONTÉE EN CHARGE ?** Une réflexion sur le traitement des recherches doit continuer d'être menée, en l'état, autour de Gallica, dans le but de concilier le paramétrage du moteur avec les attentes perceptibles en termes d'usages.

La montée en charge très significative du corpus de Gallica est un enjeu majeur dans les choix qui présideront à l'amélioration des performances du moteur de recherche retenu dans l'avenir. Sa capacité à atteindre un point d'équilibre nouveau entre performance et pertinence sera l'un des critères qui retiendra l'attention des usagers. Cet aspect doit être pris en compte dans le dimensionnement du corpus cible de Gallica, notamment à l'égard de son éventuelle vocation à agréger de multiples fonds. Il y a là un enjeu de haute importance.

• Les restitutions

Le premier niveau de restitutions proposé par les bibliothèques numériques est constitué d'une liste d'ouvrages susceptibles de satisfaire la requête, simple ou complexe, formulée par l'internaute. Comme dit ci-dessus, la pertinence et la rapidité de la réponse est un des éléments clés de cette étape. Dans le cadre d'une recherche rapide, les modalités de calcul de la pertinence sont aussi opaques d'une bibliothèque à l'autre ; on ne connaît pas les clés de pondération mises en œuvre par les moteurs de recherche. Il en résulte un sentiment d'incertitude chez l'internaute, notamment lorsque la liste des ouvrages proposés est importante.

Les fonctionnalités mises à la disposition de l'usager pour affiner son choix à partir de la première liste générée peuvent pallier partiellement ce manque de transparence.

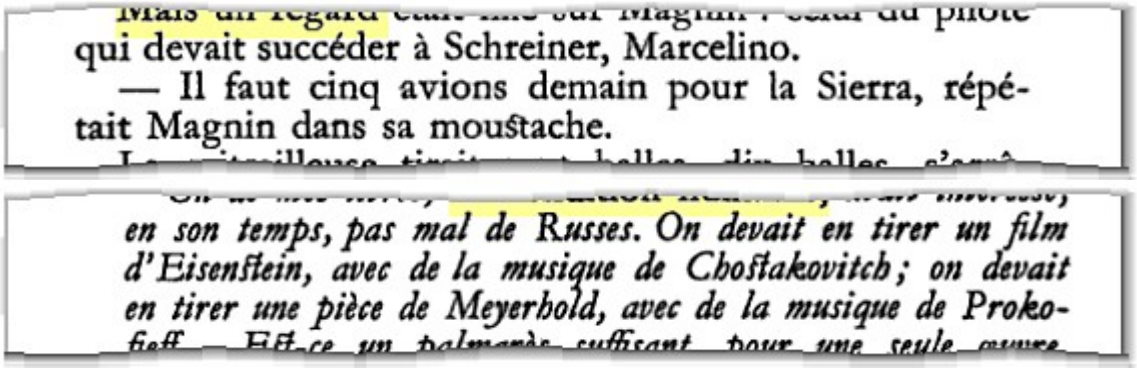
Ces fonctionnalités sont de deux ordres : les classements et les filtres. Google Livres est très en retrait sur cette offre de service documentaire à l'internaute. Aucun tri n'est possible sur les listes générées ; et les filtres proposés ne permettent que d'effectuer une sous-sélection sur le seul critère de l'accessibilité du fichier numérisé. Gallica propose à l'inverse une gamme étendue de tris et de filtres (ou facettes), techniquement opérationnels (malgré quelques bugs persistants comme, parfois, le doublonnage des œuvres) et s'appuyant sur des métadonnées plus homogènes et fiables.

Google propose cependant quelques outils qui permettent de se faire une idée du contenu du livre autrement que par l'usage des outils classiques. Un nuage de mots clés est ainsi lié à chaque livre (y compris ceux sous droit, non restitués), issus de l'analyse du contenu textuel des livres. Des mots clés sont ainsi mis en avant (noms de personnes et de personnages, noms de villes, notions…), qui permettent une appréhension thématique des ouvrages. Mais cette méthode est parfois d'un profit douteux, l'analyse automatique trouvant ses limites dans la mauvaise maîtrise du bruit ; quand, par exemple, les éléments d'une page regroupant les autres auteurs publiés par le même éditeur sont retenus comme des mots clés de l'ouvrage, à l'image d'un recueil des *Poèmes élégiaques* de Laurent Tailhade. Encore une fois, les choix de la numérisation en amont sont inséparables des modalités en aval de restitution, et donc des usages.

→ **L'AIDE À LA RECHERCHE, ELEMENT DE DIFFÉRENTIATION.** L'absence d'outils documentaires adaptés à la gestion des listes sur Google Livres surprend. Elle renforce l'impression de vrac et trahit une absence de maîtrise documentaire des fonds numériques proposés par le moteur de recherches, non compensés à l'heure actuelle par la pertinence et la puissance de son algorithme. En termes d'usages, elle peut indiquer un positionnement de Google Livres vers des publics moins familiers de ces outils documentaires, ou du moins plus tournés vers une recherche ponctuelle d'informations (recherche commode de citations…) plutôt qu'inscrits dans une démarche plus experte de recherche.

Gallica doit continuer à renforcer cette dimension ; elle va devenir un enjeu technologique majeur avec l'augmentation du nombre de références disponibles. Elle est un élément clé de différenciation entre les deux bibliothèques numériques.

Google Livres propose un accès limité aux œuvres réputées sous droits et pour la promotion desquelles elle n'a pas signé d'accord de partenariat avec les éditeurs, sous la forme de *snippets* (dans la limite de trois par expression recherchée), prenant l'allure de morceaux de pages déchirées. Il est à noter que ce type de restitution, plus ou moins appréciée des utilisateurs, s'appuyant sur l'analyse des coordonnées graphiques des mots trouvés dans la couche texte du fichier, privilégie une optique de sécurité plutôt qu'une logique d'usage. En effet, les *snippets* semblent être des extraits prédécoupés ; cela signifie qu'un mot ou une phrase se situeront toujours dans le même *snippet*, celui-ci n'étant pas recalculé, ou composé, à la volée. Il suffit donc qu'un mot se trouve sur une ligne se situant sur la bordure coupée pour qu'il n'apparaisse pas. Dans ces deux exemples, pour l'expression « Mais un regard » puis pour l'expression « *La Condition humaine* » trouvés dans l'exemplaire des *Romans* de Malraux (exemplaire de la « Bibliothèque de la Pléiade ») présenté à ce jour dans Google Livres :


qui devait succéder à Schreiner, Marcelino.
— Il faut cinq avions demain pour la Sierra, répétait Magnin dans sa moustache.
en son temps, pas mal de Russes. On devait en tirer un film d'Eisenstein, avec de la musique de Chostakovitch ; on devait en tirer une pièce de Meyerhold, avec de la musique de Proko-

Avec ce système, Google Livres s'assure qu'un livre ne peut être récupéré intégralement par l'accumulation de requêtes permettant, d'une ligne à l'autre, de reconstituer l'intégralité d'une page. Mais en termes d'usage, le résultat n'est guère satisfaisant.

Le choix d'une restitution sous forme de *snippets* s'explique probablement par la médiocre qualité du fichier texte qui était d'abord associé aux images ; l'image, elle, pour peu qu'elle soit bien positionnée, reste toujours présentable. Pour autant, on n'accède aux *snippets* qu'après être passé par une liste de résultats de livres qui peut, elle, au-dessous de chaque élément de notice, restituer une phrase du texte brut. C'est ce niveau de restitution qui avait alerté les éditeurs sur l'état réel des textes numérisés par Google et le risque d'atteinte à l'intégrité des œuvres.

Gallica se refuse pour sa part d'aller au-delà de la restitution d'un extrait au format texte pour les ouvrages sous droit déposés par les éditeurs. Elle pourrait cependant le faire puisque ces ouvrages sont, au même titre que les livres du domaine public, indexés par Lucene, à partir d'un fichier texte légèrement structuré mis à disposition de Gallica par les éditeurs partenaires. Pour le feuilletage des extraits, elle renvoie au feuilleteur indiqué par l'éditeur, hors du site Gallica, feuilleteur qui peut être distinct d'un éditeur à l'autre, et qui pourrait très bien être normalisé dans le futur, suivant l'évolution des partenariats.

• Le feuilletage des livres

Le feuilletage des livres permet d'accéder à tout ou partie des pages d'un ouvrage. L'application permettant ce feuilletage varie d'un site à l'autre. Google Livres favorise la succession des pages dans un défilement vertical, tout en proposant utilement d'autres types de mise en forme (damier, vignettes…). Pour les livres sous droits du partenariat éditeurs, certaines pages ne sont jamais restituées. Google Livres utilise la technologie javascript pour afficher les images des pages, ce qui évite le téléchargement d'un plug-in pour y accéder. Google Livres veille par ailleurs à ne jamais rendre accessible, dans cette modalité d'affichage, la couche texte des ouvrages, par un traitement en back office de la concordance entre le texte et l'image. Si on peut, par programmation, récupérer via le feuilleteur l'ensemble des images d'un livre, il n'est pas possible d'en aspirer la couche texte, si celui-ci n'est pas rendu disponible dans l'autre mode d'affichage par Google Livres (« texte brut », par groupe de pages), proposé pour une part des livres du domaine public, par ailleurs disponible en téléchargement au format ePub ou PDF (et donc potentiellement accessibles par un moteur de recherches tiers, dans les limites des protections mises en œuvre par Google Livres pour éviter les requêtes automatisées[45]). Les temps d'affichage du

[45] « Ces livres sont […] la propriété de tous et de toutes et nous sommes tout simplement les gardiens de ce patrimoine. Il s'agit toutefois d'un projet coûteux. Par conséquent et en vue de poursuivre la diffusion de ces ressources inépuisables, nous avons pris les dispositions nécessaires afin de prévenir les éventuels abus auxquels pourraient se livrer des sites marchands tiers, notamment en instaurant des contraintes techniques relatives aux requêtes automatisées. Nous vous demandons […] de [...] *ne pas procéder à des requêtes automatisées*. N'envoyez aucune requête automatisée quelle qu'elle soit au système Google. Si vous effectuez des recherches concernant les logiciels de traduction, la reconnaissance optique de caractères ou tout autre domaine nécessitant de disposer d'importantes quantités de texte, n'hésitez pas à nous contacter. Nous encourageons pour la réalisation de ce type de travaux l'utilisation des ouvrages et documents appartenant au domaine public et serions heureux

feuilleteur sont excellents.

Le feuilletage sur Gallica, fondé sur un feuilletage séquentiel page à page, est, sur le principe, assez similaire : accès aux images avec occurrences surlignées en amont de la restitution ; accès au texte brut, par page, avec, insistons-y, affichage du taux de reconnaissance de caractères ; accès au texte lu, par synthèse vocale. Un nouveau feuilletage a été mis en place plus récemment, utilisant la technologie Adobe Flex, proche des logiques de feuilleteur flash. Il répond à la promesse d'une plus grande fluidité de consultation (par pré-chargement), une amélioration des fonctionnalités de zoom… L'usage qu'en feront les internautes montrera s'il s'agit véritablement d'un progrès.

On sait que Google Livres veille à préserver une maîtrise de l'accessibilité à la couche texte des ouvrages de sa bibliothèque numérique au travers des obligations contractuelles figurant dans ses contrats avec ses partenaires. Il reste néanmoins que dans son feuilleteur, il permet un niveau d'accès au texte des ouvrages du domaine, au même titre que Gallica. Il semble donc se confirmer que la vigilance de Google Livres porte essentiellement sur l'appropriation massive, et non ponctuelle, de ses contenus textuels qui échapperait à sa maîtrise ; et qu'à ce titre, la bibliothèque numérique pose les jalons techniques nécessaires pour s'en préserver dans les conventions passées avec ses partenaires.

• Le téléchargement des livres

Au-delà du feuilletage en ligne, les bibliothèques numériques offrent la possibilité de télécharger les livres du domaine dans leur intégralité, au format PDF (dans une définition basse de 75 Dpi) ou Epub.

Pour le PDF, Google Livres ne propose qu'un fichier image, non interrogeable par l'internaute, précédé d'un avertissement sur les conditions d'exploitation dudit fichier, et filigrané à la marque du moteur de recherche. Son téléchargement est très aisé. Gallica fournit également un PDF, mais double couche (image/texte), qui permet une interrogation du texte. On retrouve donc à ce stade la prudence de Google Livres pour protéger ses contenus textuels. Pour autant, un OCR pourrait être passé sur ces mêmes fichiers, afin de restituer une couche texte.

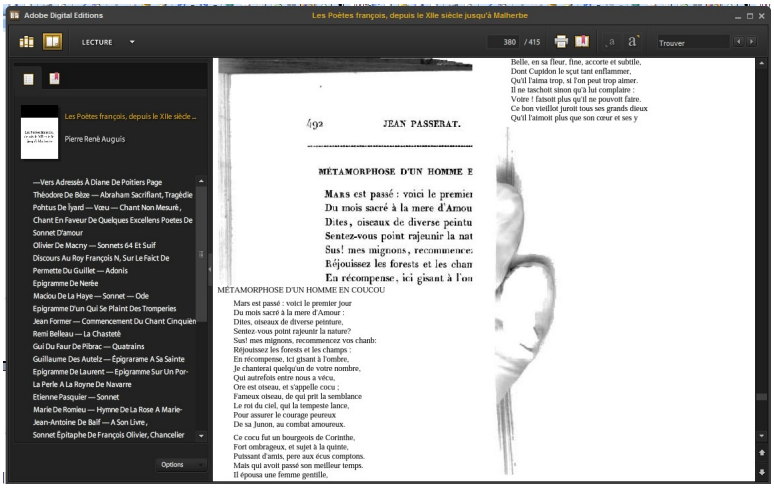
Google Livres propose pourtant le téléchargement d'un fichier Epub pour un grand nombre d'ouvrages du domaine public. Ces fichiers sont constitués à partir des mêmes sources de numérisation. Ils reproduisent donc intégralement les erreurs de l'OCR qui, pour toute une série d'ouvrages, et notamment pour les plus anciens, rend l'expérience de lecture des plus exotiques, bien que Google Livres s'en défende[46].

L'effondrement qualitatif que constitue de fait ce traitement interroge sur le devenir de ces bibliothèques numériques du début du XXI^{e} siècle et sur la nature du retraitement que nous ou nos successeurs devrons faire subir aux fichiers ou aux livres eux-mêmes. Elle relativise en tout état de cause le statut de pérennité que l'on veut bien accorder à ces sources.

de vous être utile. » (Extrait de l'avertissement figurant dans un PDF téléchargé de Google Livres).

[46] « *Despite our best efforts you may see spelling mistakes, garbage characters, extraneous images, or missing pages in this book. Based on our estimates, these errors should not prevent you from enjoying the content of the book. The technical challenges of automatically constructing a perfect book are daunting, but we continue to make enhancements to our OCR and book structure extraction technologies. We hope you'll enjoy these books as much as we do*. » (Extrait de l'avertissement figurant dans les Epub de Google Livres).

Pour l'exemple, une double page d'un fichier Epub d'un livre français du début du XIX[e] siècle issu de Google Livres[47].



• Compléments interactifs

L'un des éléments de différenciation de l'offre de Google Livres, outre le dimensionnement de son corpus multilingue, réside dans les liens qu'elle opère entre les livres (du type : ce livre est cité dans tels autres livres…) et avec des bases de données ou sources externes : Worldcat de l'OCLC, le Sudoc pour l'univers des bibliothèques (afin de permettre l'identification des exemplaires attestés dans d'autres bibliothèques), éventuellement Gallica ou encore les ressources du Web (sites d'association d'amis d'auteur…) ; libraires d'anciens et de nouveautés ; sites d'éditeurs. Pour autant, ces liens demeurent encore d'un usage assez douteux, tant la cohérence et la fiabilité des renvois d'une notice à l'autre restent imparfaites. Mais cela témoigne pour le moins d'une volonté d'ouverture indéniable de cette bibliothèque numérique.

Les deux bibliothèques proposent également des outils d'annotation sur les volumes, par le biais d'espaces personnels ou de dépôts de commentaires. Sur ce point encore, les fonctionnalités de Gallica sont plutôt avancées, comparativement au dispositif de Google Livres.

Gallica et Google Livres proposent par ailleurs des possibilités d'intégration de liens aux sites Web. Mais Google Livres va plus loin, proposant déjà des API permettant une interaction plus avancée de ses contenus avec d'autres sites de ressources (libraires, bibliothèques, blog…), par exemple via un système de vignette exportable, type *embed*. La fonction « sélectionner », paraissant dans les pages de présentation de livre, permet par exemple d'intégrer dans un site tiers un lien vers un extrait d'une page d'un livre. Google Livres peut ensuite faire un usage poussé de l'utilisation de ses services dans ses pratiques d'indexation et de pondération des contenus les plus populaires.

Les services de corrections collaboratives des métadonnées et des corpus textuels sont à ce jour des plus limités ; aucune fonctionnalité de type « wiki » n'est présente. Un service minimaliste de signalement des erreurs est proposé sur Google Livres. Gallica a engagé une expérimentation en ce sens avec Wikipédia et Wikisource sur une partie de ses corpus (1400 documents).

Parmi les liens interactifs, il faut intégrer les liens publicitaires proposés par Google Livres. Ils apparaissent tant sur les listes de résultats que sur les fiches de présentation et le feuilleteur des ouvrages. Ces liens sont contextuels, dans les limites de la définition de la contextualité selon le moteur américain. La requête sur la phrase déjà citée du poème de Voltaire sur la Providence, « Au sein de l'infini, nous élançons notre être », ramène des liens publicitaires vers les sites du libraire Amazon ou

47 *Les Poëtes françois depuis le XIIe siècle jusqu'à Malherbe* de Pierre-René Auguis, tome 4, Imprimerie de Crapelet, 1824, collection de l'Université de Michigan

des automobiles « Infiniti » (également présent sur le moteur Bing de Microsoft), ou encore vers un site proposant des solutions d'augmentation esthétique du volume mammaire (jugé pertinent par rapport au mot-clé : « sein »), et vers un site confessionnel de prières en ligne. On doit comprendre que tout accord de partenariat avec Google Livres est le support de telles campagnes ; et que les revenus escomptés, si des accords de partage de revenus étaient trouvés avec la Société Google, se feraient au titre de tels liens, peu en phase avec ce qu'il convient d'appeler la promotion du patrimoine culturel national. Il faut imaginer la réaction des lecteurs actuels de la BnF si on glissait, dans un volume communiqué en salle de lecture, un *flyer* pour de la chirurgie esthétique.

• Le référencement

La section Livres constitue un onglet spécifique sur la page du moteur de recherches Web Google. Des résultats issus de Google Livres peuvent cependant remonter via une recherche simple sur le moteur web, mais sans exhaustivité ; Google semble donc indexer les données de Google Livres, y compris le texte complet des ouvrages, dans le cadre de son activité d'indexation générale du Web.

Une fois une requête formulée sur le moteur Web, on peut relancer la même requête sur le périmètre « Livres », comme on peut le faire sur le périmètre « Images » (lien « Afficher les options », puis « Livres »). C'est à ce stade qu'apparaissent les liens publicitaires de Google Livres. Un accès direct à Google Livres propose enfin un autre type d'accès, mais aux résultats identiques avec l'accès précédent (mais des liens publicitaires différenciés).

S'il y a un rapport étroit entre Google et Google Livres, il reste que Google ne peut être considéré comme une modalité d'accès suffisante aux contenus de la bibliothèque numérique. Il est à noter également que les contenus de Google Livres ne semblent pas correctement indexés par les autres moteurs Web (Yahoo!, Bing). On rappellera à ce titre que les possibilités d'accéder au texte des ouvrages sur Google Livres ne sont que modérément limitées ; mais que Google annonce, dans ses avertissements, avoir mis en œuvre des solutions de protection de ses contenus, afin de limiter leur indexation massive par des tiers.

Les contenus de Gallica (listes de documents et livres) ne sont guère indexés par les moteurs de recherche du Web (Google, Yahoo!, Bing). Et c'est une des problématiques majeures qui se pose aujourd'hui à la Bibliothèque numérique de la BnF. Des stratégies multiples sont envisagées à ce jour : multiplier les accès à la base par la mise en place de liens fins éditorialisés, intégrer dans les pages web des documents des métadonnées structurées permettant une meilleure reconnaissance des éléments par les moteurs de recherche (utilisation de meta Dublin Core) dans le cadre du développement du web sémantique, diffuser ses métadonnées (via un serveur OAI-PMH ouvert, ce qui permet, par exemple, à WorldCat d'intégrer les descriptions des notices BnF). Cette dernière approche, limitée par l'usage du protocole utilisé (OAI-PMH), devrait être complétée par le déploiement massif et automatisé de pages HTML liées reprenant ces informations (projet dit du « pivot documentaire » par la BNF), et les rendant facilement indexables par les moteurs de recherche. L'étude préalable en a été confiée à Cap Gemini. Enfin, une optimisation de l'indexation des pages du site va être réalisée, à travers l'usage classique des instructions aux robots d'indexation (robot.txt et sitemap.xml). Cette refonte devrait permettre une meilleure indexation des pages de contenus textuels, même si le nombre de celles-ci (une page web par page de livre) constitue un frein notable, relevant de la conception même du système de consultation. Une phase de test est aujourd'hui en cours, dans le cadre du projet européen TELplus. Elle doit s'accompagner d'un effort de concertation renouvelé avec les grands moteurs de recherche du Web. Gallica travaille par ailleurs à des outils de bookmarking qui devraient favoriser la diffusion de ses contenus dans les réseaux sociaux du Web.

L'ensemble de ce dispositif est l'un des grands enjeux de performance de Gallica pour 2010. Il doit pouvoir pallier la grande insuffisance des résultats obtenus en la matière jusqu'à ce jour.

La numérisation de masse est une voie possible, mais non exclusive, de la numérisation ; ses contraintes et ses limites, tout comme ses indéniables apports, doivent être aujourd'hui intégrés à toute réflexion poussée sur ce qui constitue les missions historiques des bibliothèques patrimoniales, tant en termes de conservation que de valorisation. C'est une des conditions nécessaires pour ne pas perdre le fil du débat.

Annexe 4 : Liste des bibliothèques européennes partenaires du programme Google Recherche de livres

au 21 décembre 2009
(Source : http://www.google.fr/googlebooks/partners.html **[consulté le 21/12/2009]**

Allemagne

Bibliothèque d'Etat de Bavière (2007)

Belgique

Bibliothèque universitaire de Gand (2007)

Espagne

Bibliothèque nationale de Catalogne (2007)

Bibliothèque de l'université Complutense de Madrid (2006)

France

Bibliothèque municipale de Lyon (2008)

Royaume-Uni

Bibliothèque Bodleienne (Université d'Oxford) (2004)

Suisse

Bibliothèque cantonale et universitaire de Lausanne (2007)