Savoirscom1, collectif pour les biens communs de la connaissance

Quel statut légal pour le content-mining ?

Synthèse de Savoirscom1,
consécutive à l'audition du 15 janvier 2013
par le Conseil Supérieur de la Propriété Littéraire et Artistique

Pierre-Carl Langlais et Lionel Maurel[1]

[1] La synthèse incorpore également des éléments d'un billet de Pierre-Carl Langlais, « Quel statut légal pour le data-mining ? », *soms.hypotheses.org*, 11 janvier 2014, http://scoms.hypotheses.org/56 / Licence CC0

Table des matières

Contextes et enjeux

L'exploration automatisée des textes et des données (ou *content-mining*[2]) est une activité en plein essor. Des outils d'extraction et d'analyse élaborés sont accessibles à un coût faible voire nul. Les savoir-faire se démocratisent. Les réseaux numériques permettent de récupérer et de croiser quantité d'informations.

L'amélioration et la diffusion de ces techniques d'exploration facilitent considérablement le travail de recherche. La délégation de nombreuses tâches de récupération et d'identification des informations à des algorithmes permet d'envisager des projets d'une ampleur inédite. Text2genome cartographie le génome humain en compilant automatiquement trois millions de publications[3]. La portée de cette métamorphose dépasse le seul cadre de la recherche académique. Le content-mining irrigue un écosystème de la connaissance en pleine recomposition, marqué notamment par l'essor de communautés ouvertes sur le modèle de Wikipédia ou de Wikidata.

Le content-mining pourrait ainsi susciter une véritable révolution des usages scientifiques. Toutes les disciplines sont potentiellement concernées : le mouvement des humanités numériques s'appuie sur ces techniques pour envisager sous un angle inédit des pans entiers de notre patrimoine culturel[4]. Les bienfaits sociaux, sanitaires et intellectuels de cette mutation majeure semblent innombrables[5]. « Pour le dire simplement, l'extraction automatisée des textes et des données sauve des vies humaines. »[6]

Si les limites techniques du content-mining ne cessent d'être repoussées, les limites légales demeurent. Chercheurs, contributeurs et professionnels sont confrontés à des insécurités juridiques récurrentes. En l'absence d'un cadre légal univoque et uniforme, il

---

[2] Nous utiliserons cette expression anglophone en raison de sa brièveté et en l'absence d'un équivalent français couvrant le même champ sémantique (l'expression assez courante « exploration des données », ne couvre que le champ du « data-mining » et non également celui du « text-mining »)

[3] Présentation du projet à l'adresse http://text2genome.smith.man.ac.uk/

[4] Cf. l'un des projets emblématiques de l'application du content-mining aux humanités numérique, pegasusdata : http://pegasusdata.com/a-propos/

[5] Pour un compte-rendu évocateur, cf. Andres Guadamuz and Diane Cabell "Data mining White Paper: Analysis of UK/EU Law on Data mining in Higher Education Institutions" p4; available at http://www.technollama.co.uk/wp-content/uploads/2013/04/Data-Mining-Paper.pdf

[6] Compte-rendu de l'atelier organisé par la ligue des bibliothèques européennes et la British Library le 27 septembre 2013, « The Perfect swell: defining the ideal conditions for the growth of text and data mining in Europe », p. 2, http://www.libereurope.eu/sites/default/files/TDM%20Workshop%20Report%5B1%5D.pdf

est nécessaire de recourir à des accords contractuels au cas par cas. Or les universités paient déjà, parfois assez chèrement, un droit d'accès aux revues universitaires. Les multinationales de l'édition scientifiques imposent ainsi de fait une nouvelle couche de droit : au droit de lire s'ajouterait un droit à extraire de l'information.

Un article de la revue Nature, publié en mars 2013, fait état de l'une des incidences majeures de ces pratiques contractuelles : les négociations s'étalent en longueur[7]. La réalisation du projet text2genome a nécessité 3 ans de tractations tortueuses. Ces complications contribuent à ralentir inutilement le processus de recherche, voire à décourager les initiatives.

L'avancée américaine

À cet égard, les États-Unis ont une longueur d'avance sur les États européens. La législation américaine prévoit en effet une exception suffisamment souple et générale pour s'adapter à ces enjeux émergents : le « fair use » ou « usage loyal »[8]. Une jurisprudence récente montre que le content-mining à des fins de recherche entre aisément dans le cadre de ce fair use.

En 2011, plusieurs associations de défense des droits d'auteur intentent un procès à la bibliothèque numérique HathiTrust, constituée en grande partie à partir de copies remise par Google aux bibliothèques partenaires du projet Google Books. Ce service, soutenu par plusieurs universités américaines, permet d'effectuer des recherches en plein texte sur des textes toujours protégés. Le jugement rendu en novembre 2012 a reconnu le bien fondé de la démarche de HathiTrust[9]. La motivation de la publication est « juste » (*fair*) puisque ce service a une nette vocation scientifique. Elle ne pénalise pas les œuvres originelles, dans la mesure où il s'agit d'un « usage transformatif ». Le juge reconnaît en effet une interprétation élargie de la transformation, qui ne concernerait pas seulement sur la métamorphose du contenu préexistant, mais aussi sur la métamorphose de ses conditions d'usage :

> Un usage peut être dit transformatif lorsque l'œuvre originale est changée. Néanmoins, un usage transformatif peut également être perceptible lorsque cette

---

[7] RICHARD VAN NOODEN, « Text-mining spat heats up », *Nature*, http://www.nature.com/news/text-mining-spat-heats-up-1.12636

[8] Section 107 du titre 17 du code des États-Unis, consultable à l'adresse http://www.law.cornell.edu/uscode/text/17/107

[9] Décision du juge Harold Baer à l'issue du procès Authors Guild v. HathiTrust, consultable à l'adresse http://www.tc.umn.edu/~nasims/HathivAG10_10_12.pdf

> œuvre a une fonction entièrement distincte (…) : des capacités de recherches supérieures plutôt que l'accès effectif au contenu protégé[10].

La saga judiciaire de Google Books a débouché sur une conclusion semblable en novembre dernier. La dernière décision du juge Denny Chin reprend clairement les termes de la jurisprudence créée par l'affaire HathiTrust[11]. Elle fait d'ailleurs explicitement référence au *content-mining* :

> Google Books est transformatif dans le sens où il a transformé le texte des livres en données à des fins de recherche, y compris pour de la fouille de données ou de texte (data mining et text mining), ouvrant ainsi de nouveaux champs à la recherche. Les mots dans les livres ont ainsi pu être utilisés d'une manière complètement différente par rapport à ce qui existait avant. Google a créé quelque chose de nouveau dans la manière d'utiliser le texte des livres, la fréquence des mots et les tendances dans leur utilisation fournissant des informations substantielles[12].

En somme les outils de content-mining mis en place par Hathitrust ou Google sont tout à fait légitimes, dans la mesure où ils combinent deux exceptions distinctes. Ils transforment le texte original : ils ne donnent pas à lire un texte expressif mais un simple index, éventuellement complété par quelques citations. Ils jouent un rôle social et scientifique bénéfique, délié, qui plus est, de toute logique de monétarisation immédiate.

Le « fair use » n'autorise pas tous les usages possibles de content-mining. Si l'on s'en tient à la jurisprudence existante (qui peut encore évoluer), il ne serait pas possible de donner à lire l'ensemble des bases de données ou des corpus textuels. La copie est autorisée, dans la mesure où le service ne saurait exister sans elle, mais elle ne peut pas être rendue publique. Cette distinction entre *copie* et *publicité* est importante : elle permet de fonder un compromis viable entre les praticiens du content-mining et les ayants-droit.

Cette possibilité pour les universités de réaliser des copies de textes protégés a été confortée par la décision rendue en 2012 dans l'affaire opposant la Georgia State University à plusieurs éditeurs à propos de la constitution de "e-reserves" d'articles

---

[10] *Ibid,* p. 16

[11] Décision du juge Denny Chin à l'issue du procès The Authors Guild v. Google Books http://fr.scribd.com/doc/184176014/Judge-Denny-Chin-Google-Books-opinion-2013-11-14-pdf

[12] Texte de la décision précédente, traduit par Lionel Maurel, « Verdict dans l'affaire Google Books : une grande leçon de démocratie ? », *S. I. Lex*, http://scinfolex.com/2013/11/15/verdict-dans-laffaire-google-books-une-grande-lecon-de-democratie/

scientifiques constitués à partir de scans de documentation papier et réutilisés à des fins éducatives[13].

Ce cadre légal émergent ne conforte pas que des institutions universitaires ou des entreprises. Il favorise également des initiatives bénévoles. La « Virtual Reading Room » de Internet Archive permet également d'effectuer des recherches sur des textes toujours couverts par une forme de propriété intellectuelle[14]. Une telle initiative serait impossible en France et dans la majorité des pays européens.

Les blocages européens

La législation en vigueur dans la plupart des pays européens, dont la France, n'augurent pas d'une reconnaissance comparable du content-mining.

Le « fair use » reste globalement étranger aux traditions juridiques européennes. La directive sur la société de l'information de 2001 a certes tenté d'introduire une forme d'exception pédagogique et scientifique. Elle prend la forme suivante :

> lorsqu'il s'agit d'une utilisation à des fins exclusives d'illustration dans le cadre de l'enseignement ou de la recherche scientifique, sous réserve d'indiquer, à moins que cela ne s'avère impossible, la source, y compris le nom de l'auteur, dans la mesure justifiée par le but non commercial poursuivi[15].

Cette exception inclut une contrainte forte : la fin exclusive « d'illustration ». Selon Lucie Guibault, le data-mining ne saurait s'inscrire dans ce cadre[16]. Il n'apporte pas une « illustration » du texte original, mais plutôt un nouveau mode de lecture. C'est en ce sens que la jurisprudence américaine évoque un « usage transformatif » : une métamorphose du contenu original en un index d'éléments et de données à rechercher et combiner.

En l'absence de fair use, le content-mining reste soumis à de nombreuses insécurités juridiques. Plusieurs couches de droit se superposent. Textes et bases de données sont

---

[13] Cf. la synthèse du procès sur *en.wikipedia.org,* http://en.wikipedia.org/wiki/Cambridge_University_Press_v._Becker

[14] Présentation de la « Virtual Reading Room » de InternetArchive sur *knightfoundation.org,* http://www.knightfoundation.org/blogs/knightblog/2014/1/7/internet-archives-virtual-reading-room-empowers-data-mining-societal-scale/

[15] Directive 2001/29/CE du parlement européen et du conseil du 22 mai 2001, sur l'harmonisation de certains aspects du droit d'auteur et des droits voisins dans la société de l'information, art. 5, 3, a

[16] Lucie Guibault, *Intellectual property rights' obstructions to text and data mining, http://www.youtube.com/watch?v=hfpkJs6GJUg*

soumis au droit d'auteur dès lors qu'ils répondent à un critère d'originalité[17]. Il existe également, depuis 1996, un droit sui generis des bases de données qui repose, lui, sur un critère d'investissement : toute personne ou organisation consacrant du temps ou des moyens financiers à la constitution et à l'entretien d'une base de donnée peut légitimement réclamer sa protection sous ce titre[18]. L'ajustement de ces divers paramètres est délicat et varie grandement d'une situation à l'autre. Il est en particulier très difficile de déterminer à partir de quand l'emprunt d'une base de donnée est « substantiel »[19]. Ces incertitudes concourent à démotiver les projets de content-mining à des fins scientifiques : une institution universitaire peut difficilement prendre le risque d'être assignée en justice.

Plusieurs États-membres envisagent de mettre en place des dispositions plus libérales. Le Royaume-Uni inclut ainsi une exception spécifique pour le content-mining dans le cadre de son projet de modernisation du copyright : il serait autorisé à des fins non-commerciales[20]. Un rapport similaire publié en Irlande en 2013 préconise la mise en place d'un « fair dealing », soit une forme plus restrictive du « fair use » américain : les projets de content-mining doivent nécessairement se soumettre à plusieurs contraintes prédéterminées[21].

Ces ouvertures ne peuvent aboutir à l'échelon national. La liste des exceptions apportée par la directive sur la société de l'information est limitative et n'inclut aucun principe de subsidiarité. Le texte de présentation souligne ainsi explicitement que « Les États membres ont la faculté de prévoir des exceptions ou limitations aux droits prévus aux articles 2 et 3 dans les cas suivants »[22]. La reconnaissance d'un statut légal spécifique au content-mining ne pourrait ainsi s'imposer qu'au niveau européen.

---

[17] En France, ce droit d'auteur des données est couvert par l'article L112-3 du code de la propriété intellectuelle.

[18] Directive 96/6/CE du Parlement européen et du conseil du 11 mars 1996, concernant la protection juriduqe des bases de données, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:fr:HTML

[19] Lucie Guibault, *Intellectual property rights' obstructions to text and data mining, http://www.youtube.com/watch?v=hfpkJs6GJUg*

[20] *Modernising Copyright : A modern, robust and flexible framework* , Government response to consultation on coppyright exceptions and clarifying Copyright Law, Londres, 2012, pp. 36-37, http://www.ipo.gov.uk/response-2011-copyright-final.pdf

[21] *Modernising Copyright* , The Report of the Copyright Review Committee, Dublin, 2013, pp. 85-88, http://www.enterprise.gov.ie/en/Publications/CRC-Report.pdf

[22] Directive 2001/29/CE du parlement européen et du conseil du 22 mai 2001, sur l'harmonisation de certains aspects du droit d'auteur et des droits voisins dans la société de l'information, art. 5, 3

Au début de l'année 2013, la Commission Européenne a lancé un groupe de travail sur le statut légal de l'exploration automatisée des données : le *Text and Data mining working group*[23]. Rapidement des dissensions émergent.

Les représentants des éditeurs envisagent la création de licences ad hoc : les chercheurs obtiendraient un droit à extraire moyennant le versement d'un abonnement supplémentaire. Une présentation de Springer dévoile ainsi un système d'accès complexe : tout chercheur désirant accéder à la base doit remplir un formulaire où il détaille son projet de recherche. C'est seulement au terme de ce processus qu'il reçoit une clé d'accès[24]. Ce processus ne permet pas seulement de maintenir un droit de regard sur le content-mining : il facilite la constitution de vastes corpus de méta-données de la recherche. Springer sait désormais exactement qui étudie quoi avec son corpus.

Inversement, bibliothécaires et défenseurs de la culture libre rejettent le principe des licences. De leur point de vue, le content-mining devrait simplement faire l'objet d'une exception, au même titre que le droit de citation.

En avril, la Commission a finalement arbitré le conflit au profit des multinationales de l'édition : le groupe de travail ne travaillera que sur les licences. De nombreux participants du *Text and Data mining working group* décident alors de se retirer une initiative non consensuelle. La Ligue des bibliothèques européennes de recherche a ainsi publié un communiqué d'une grande clarté :

> [nous quittons] un processus de décision dont le résultat est déjà pré-déterminé : l'ajout de nouvelles licences serait la seule solution envisageable aux multiples écueils rencontrés dès lors que l'on souhaite extraire automatiquement les données d'un contenu auquel l'on a déjà accès[25].

Depuis lors, le processus européen est relativement bloqué : il n'y pas eu d'actualisation depuis novembre 2013. Tout semble indiquer que le travail ne reprendra pas avant les prochaines élections européennes — peut-être sur des bases plus consensuelles. Dans le même temps, la Commission européenne a lancé une consultation sur la révision du droit d'auteur qui comporte une partie sur le Content-Mining, dans

---

[23] Présentation du « Text and data working group » sur le site de l'Union Européenne, http://ec.europa.eu/licences-for-europe-dialogue/en/content/text-and-data-mining-working-group-wg4

[24] Cette présentation de Springer a été mise en ligne sur le site de l'Union Européenne à l'adresse suivante : https://ec.europa.eu/licences-for-europe-dialogue/sites/licences-for-europe-dialogue/files/Publishers-Perspective-Initiatives_0.pdf

[25] Communiqué de la Ligue des bibliothèques européennes, « Licences for Europe - A Stakeholder Dialogue », publié le 26 février 2013 sur le site *libereurope.eu*, http://www.libereurope.eu/news/licences-for-europe-a-stakeholder-dialogue-text-and-data-mining-for-scientific-research-purpose

laquelle elle demande si la question doit être traitée sur une base contractuelle ou par le biais d'une exception[26].

Le droit de lire comme droit d'extraire

La réflexion sur le statut légal du content-mining reste assez formelle. Elle se focalise sur les exceptions déjà existantes, soit sur un éventail de dispositions très spécifiques.

Pourtant, l'invocation d'un principe fondamental de la législation sur la propriété intellectuelle pourrait régler en partie la question : la distinction entre information et expression. Par définition, le droit d'auteur ou le copyright ne portent pas sur des informations « brutes », mais sur leur expression originale. Ce principe se retrouve dans toutes les législations. Il est particulièrement bien explicité aux États-Unis où la dichotomie idée/expression fait autorité depuis le procès Baker v. Selden[27].

En France, on le retrouve plutôt in absentia. Le code de la propriété intellectuelle porte sur des œuvres. Les bases de données ne sont concernées que dans la mesure où « par le choix ou la disposition des matières, [elles] constituent des créations intellectuelles. »[28] Les données individuelles et les informations sont clairement exclues de cette définition.

Le *content-mining* avec un crayon et un papier

À cet égard, il convient de démystifier la pratique du content-mining. Il s'agit avant tout d'une technologie intellectuelle (au même titre, par exemple, que le boulier ou la comptabilité à partie double) : un outil permettant de faciliter ou d'automatiser certaines opérations mentales. Au lieu de récupérer et de recouper manuellement des milliers, voire des millions de données, il est possible de confier ces tâches peu gratifiantes à des algorithmes et à des systèmes de gestion de bases de données.

Si le content-mining marque un changement d'échelle, il ne fonde pas une activité nouvelle. Extraire et synthétiser des informations préexistantes constituent le labeur quotidien du chercheur depuis que la recherche scientifique existe. Notes de bas de page, graphiques, tableaux : tous ces usages structurants de l'écriture académique reposent sur une forme d'extraction manuelle d'informations préexistantes. La notice Wikipédia assez

---

[26] « Consultation publique sur la révision des règles de l'Union européenne en matière de droit d'auteur », *ec.europa.eu*, http://ec.europa.eu/internal_market/consultations/2013/copyright-rules/index_fr.htm

[27] « Baker v. Selden », *en.wikipedia.org,* http://en.wikipedia.org/wiki/Baker_v._Selden

[28] Code de la propriété intellectuelle, art. L112-3

complète sur l'exploration des données met en évidence que l'histoire de cette pratique est bien antérieure à l'informatisation :

> La généralisation de modèles à partir d'un grand nombre de données n'est pas un phénomène récent (…) Legendre publie en 1805 un essai sur la méthode des moindres carrés qui permet de comparer un ensemble de données à un modèle mathématique. Les calculs manuels coûteux ne permettent cependant pas d'utiliser ces méthodes hors d'un petit nombre de cas simples et éclairants[29].

Reconsidérer le content-mining dans cette perspective historique longue a des conséquences légales non négligeables. Le 27 septembre 2013, la Ligue des bibliothèques de recherche européennes organise un atelier de réflexion sur le statut légal du content-mining à la British Library. L'atelier vise à fédérer plusieurs institutions opposés à l'orientation pro-licences de la commission européennes. La synthèse de l'atelier relève qu'il n'y a pas de différence fondamentale entre le relevé manuel d'information et le content-mining informatisé :

> Il est intéressant de noter que le content-mining effectué avec un crayon et un papier n'est pas régulé par le code de la propriété intellectuelle car, il n'est pas nécessaire de faire des copies de l'œuvre entière, c'est-à-dire que la personne peut simplement faire une copie des faits et des données qu'il souhaite conserver avec un crayon, car les faits et les données en sont pas régulés par le droit d'auteur ou le droit des bases de données[30].

L'utilisation ou la non-utilisation d'un outil ne peut à elle seule légitimer un statut légal à part. Il est en effet tout-à-fait possible (mais impraticable) de réaliser le projet text2genome avec des papiers et des crayons. Au lieu de prendre quelques années, le travail s'étalerait sur plusieurs décennies voire plusieurs siècles : à raison de 20 minutes de lecture et d'extraction manuelle des informations par publication, il serait nécessaire d'y consacrer plus de cent milles journées de huit heures. La synthèse de l'atelier du 27 septembre souligne ainsi, qu'en soi, le content-mining se situe en dehors du régime de la propriété intellectuelle :

> En toute logique, l'exploration automatisée des textes et des données ne devrait pas être concernée par la législation sur le copyright, dans la mesure où elle ne porte pas sur l'expression d'idées régulée par le copyright, mais sur l'extraction et l'analyse de de

---

[29] « Exploration des données », *fr.wikipedia.org*, https://fr.wikipedia.org/wiki/Exploration_de_donn%C3%A9es

[30] Compte-rendu de l'atelier organisé par la ligue des bibliothèques européennes et la British Library le 27 septembre 2013, « The Perfect swell: defining the ideal conditions for the growth of text and data mining in Europe », p. 2, http://www.libereurope.eu/sites/default/files/TDM%20Workshop%20Report%5B1%5D.pdf

> faits et de données (…) Nous ne pouvons dresser une distinction absolue entre les humains et les robots à l'ère du numérique. L'extraction du texte et des données est une forme de lecture, qu'elle soit ou non réalisée par un humain ou par une machine. Le droit de lire devrait entraîner le droit d'extraire[31].

Dès lors que l'on abolit la distinction artificielle entre le content-mining et le relevé manuel des informations, l'exploration des données semble indissociable du droit de lire. L'accès à un texte permet la reprise de ses informations. En payant déjà un droit d'accès aux éditeurs de revues scientifiques, les universités obtiennent *ipso facto* un droit d'extraction. Nul ne saurait leur contester ce droit en tant que tel. Il s'inscrit en dehors du cadre de la propriété intellectuelle : dans le domaine public de l'information.

Le domaine public de l'information

Comme nous l'avons souligné plus haut, la plupart des législations sur la propriété intellectuelles reconnaissent la distinction, fondamentale, entre idée et expression. Pour autant, le statut exact de l'information n'est jamais explicitement précisé.

L'UNESCO a récemment développé l'idée d'un domaine public informationnel ou « indivis mondial de l'information ». Ce domaine public comprendrait toutes les informations « publiquement accessibles ». Ce constat se double d'une prescription : toutes les informations « intéressant le citoyen » devraient circuler librement et sans contrainte[32].

Le professeur Michel Vivant reprend une conception similaire en proposant d'inclure les informations dans un « fonds commun » qui comporterait à la fois les œuvres entrées dans le domaine public à l'issue de la période de protection des droits patrimoniaux, ainsi que les créations ou éléments ne pouvant accéder à la protection (informations, faits, données, etc) :

> Le domaine public évoque souvent, à tort, les seules œuvres qui chutent dans le fonds commun au terme de la protection. Or notre fonds commun est beaucoup plus large puisqu'il intègrent à la fois les oeuvres tombées dans le domaine public, les idées, les créations ne pouvant accéder à la protection (…) en un mot tous les "matériaux" ayant vocation à être utilisés dans le cadre d'un processus créatif[33].

---

[31] *Ibid*, p. 3

[32] Définition du « Domaine public de l'information » sur le site *Unesco*.org, http://portal.unesco.org/ci/fr/ev.php-URL_ID=1535&URL_DO=DO_TOPIC&URL_SECTION=201.html

[33] Précis Dalloz. Droit d'auteur. 2009, p. 79

L'extraction des bases de données des éditeurs scientifiques correspond à un double titre à ce domaine public informationnel :

Elle porte sur une information publiquement accessible. Il ne s'agit pas d'un corpus privé, limité à un nombre restreint et prédéfini de lecteurs. Il suffit de s'acquitter des droits d'accès pour consulter le contenu publié par les éditeurs scientifiques. Rien ne peut s'opposer légitimement à l'extraction automatisée d'une information déjà lisible.

Elle « intéresse » le citoyen et plus largement la société dans son ensemble. La jurisprudence américaine a ainsi accepté de placer plusieurs initiatives de content-mining sous le régime du « fair use », dans la mesure où elles servent un intérêt public.

Une illégalité collatérale

Reconnaître l'indissociabilité du droit de lire et du droit d'extraire ne résout pas tout. Si l'extraction automatisée des textes et des données est située, dans l'absolu, en dehors du champ de la propriété intellectuelle, certaines pratiques associées à cette activité sont potentiellement illégales au regard de la législation actuelle. Trois éléments sont régulièrement invoqués pour justifier un aménagement spécifique de la loi existante.

(1) Le content-mining nécessite fréquemment de créer une copie plus ou moins temporaire, visible pour un petit groupe de personnes. Des projets de grande ampleur ne peuvent se contenter d'aller constamment chercher l'information en ligne : les API ou le *Web scraping* ne permettent pas de requérir ou de combiner rapidement des millions de données. Seule une reproduction des corpus analysés permet d'effectuer une investigation approfondie. La directive sur la société de l'information de 2001 autorise certes la réalisation de reproductions transitoires à des fins purement techniques[34]. Transitoire n'est cependant pas un vain mot : seuls sont concernées des reprises véritablement éphémères[35]. Les reproductions issues du content-mining ne rentrent pas dans ce cadre : elles doivent demeurer accessible au groupe de recherche pendant plusieurs mois, voire plusieurs années. C'est ainsi essentiellement à travers la copie des contenus nécessaire à sa réalisation que le content-mining peut être saisi par le droit d'auteur. Le professeur Vivant estime cependant que la notion de reproduction devrait être redéfinie dans l'environnement numérique pour que toute copie ne donne pas lieu à l'application du droit d'auteur. L'exception sur les copies

---

[34] Directive 2001/29/CE du parlement européen et du conseil du 22 mai 2001, sur l'harmonisation de certains aspects du droit d'auteur et des droits voisins dans la société de l'information, art. 5, 1

[35] Lucie Guibault, *Intellectual property rights' obstructions to text and data mining,* http://www.youtube.com/watch?v=hfpkJs6GJUg

transitoires de la directive européenne de 2001 devrait donner lieu à une application plus large. Le professeur Vivant invite à aller au-delà, en se posant la question de savoir s'il est justifié de chercher à réguler toutes les copies :

> Il y a un piège : dès que l'on aborde des questions qui sont marquées par la technicité, on veut avoir un décryptage technique. Le droit est un instrument de régulation sociale. Qu'il y ait quelque part une copie, qui signifie reproduction, c'est une chose. Mais est-ce cela que nous devons appréhender en terme de régulation sociale ?[36].

(2) Dans certains cas, il s'avère profitable de maintenir d'une copie accessible, quoique non consultable. Sans avoir accès à une reproduction intégrale, des chercheurs ou de simples lecteurs peuvent ainsi effectuer des requêtes de manière à récupérer des données et des informations utiles. Grâce au « fair use » des bibliothèques numériques, comme Google Books, Hathitrust ou Internet Archive autorisent de telles requêtes sur des contenus protégés. En Europe, aucune exception ne garantit l'applicabilité de cette activité. Le droit de citation ne constitue pas une exception valide, dans la mesure où il ne s'applique pas aux bases de données. En France, elles doivent être « justifiées par le caractère critique, polémique, pédagogique, scientifique ou d'information de l'œuvre à laquelle elles sont incorporées. »[37] Or, même si elle possède un caractère scientifique, une base de donnée n'est pas une œuvre.

(3) Un risque accru de contrefaçons. Il s'agit du principal argument des éditeurs : en autorisant les lecteurs à extraire un contenu protégé, on faciliterait la réalisation et la publication de copies pirates. Le rapport du Royaume-Uni sur la modernisation du copyright souligne que les résistances des éditeurs portaient essentiellement sur la sécurité du contenu ainsi mis à disposition : des acteurs mal intentionnés pourraient profiter d'un droit d'extraire non encadré pour diffuser une reproduction substantielle de publications ou de bases de données scientifiques[38]. Or, la législation actuelle sur la propriété intellectuelle apporte des précautions suffisantes : si un lecteur republie un contenu sans autorisation, son détenteur sera habilité à l'attaquer en justice. Il convient également de souligner que les avancées actuelles en faveur du libre accès

---

[36] Audition devant la mission Lescure du 21 décembre 2012, mise en ligne sur *culture-acte2*, http://www.culture-acte2.fr/topic/audition-de-michel-vivant/

[37] Code de la propriété intellectuelle, art. L122-5

[38] *Modernising Copyright : A modern, robust and flexible framework* , Government response to consultation on coppyright exceptions and clarifying Copyright Law, Londres, 2012, p. 36, http://www.ipo.gov.uk/response-2011-copyright-final.pdf

contribuent à diminuer fortement ce risque. Actuellement, plus de 50% des articles scientifiques européens sont publiés sous une forme de libre accès[39]. Cette proportion va forcément s'accroître au cours des années à venir avec la généralisation du modèle auteur-payeur. Les multinationales de l'édition, qui se sont très bien adaptées à cette nouvelle donne, peuvent de moins en moins arguer que la contrefaçon représente un manque à gagner.

Les cadres légaux actuellement envisagées

Dans l'absolu, ces cadres légaux ne visent qu'à réguler et sécuriser plusieurs effets collatéraux du *content-mining*, et plus particulièrement la possibilité de réaliser des copies substantielles. Cependant, certaines solutions envisagées vont plus loin. Le système de licence, encouragé par les multinationales de l'édition scientifique vise ainsi à encadrer tous les aspects du content-mining, y compris la simple collecte de données individuelles.

Un système de licences *ad hoc*

Pour l'essentiel, ce système acte la situation actuelle : les éditeurs imposent un droit d'extraire qui se se superpose au droit de lire. Le principal argument en sa faveur est d'ordre financier. Il permettrait de financer directement la fourniture de procédures d'accès aux données (API, bases de données à télécharger).

En réalité, ce type de fourniture est peu onéreux. Wikipédia fournit par exemple un service d'une très grande qualité qui combine une API et un publication mensuelle de l'intégralité des bases de données wikipédiennes[40]. Ce service a été largement exploité par les communautés scientifiques : sur les 7000 publications académiques consacrées à Wikpédia, une large partie d'entre elles s'appuient sur des méthodologies quantitatives[41]. La mise en place d'un service similaire ne devrait poser aucune difficulté particulière pour les multinationales de l'édition (comme Springer ou Elsevier). Les institutions universitaires et scientifiques pourraient également apporter une aide au cas par cas : la création d'une API pour telle ou telle collection éditoriale pourrait sans doute facilement rentrer dans le

---

[39] Communiqué de presse de l'Union européenne, « Open access to research publications reaching 'tipping point' », 21 août 2013, http://europa.eu/rapid/press-release_IP-13-786_en.htm

[40] Présentation des multiples options d'extractions sur la page « Research:Data » de *meta.wikimedia.org,* https://meta.wikimedia.org/wiki/Research:Data

[41] NICOLAS JULLIEN, *What we know about Wikipedia. A review of the literature analyzing the project(s)*, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2053597 p. 2

cadre du financement d'un projet de recherche. Ce type de défraiement serait tout à la fois moins contraignant et moins onéreux qu'un système de licence.

Ce système de licences présente par ailleurs de nombreux inconvénients. Les chercheurs doivent effectuer des procédures chronophages au détriment du temps effectivement dévolu à la recherche. La procédure d'accès prévue par Springer implique ainsi une succession d'étapes et de formulaires à remplir, avant de pouvoir récupérer une clé autorisant l'extraction automatisée d'information[42]. Les projets de content-mining ne portent pas nécessairement sur une seule collection éditoriale : ils requièrent bien souvent le croisement de plusieurs corpus. La récupération de l'ensemble des publications dans un certain domaine scientifique implique ainsi d'effectuer de nombreuses démarches.

Les formulaires mis en œuvre incluent un autre biais : ils permettent de récupérer des méta-données de recherche. Chaque projet de content-mining doit préciser l'identité de ses acteurs, ses objectifs, sa portée. Les éditeurs exercent de fait un droit de contrôle et de surveillance sur les projets menés. Il savent qui consulte quoi et peuvent ensuite exploiter ces méta-données à des fins de marketing ou les revendre à des services tiers.

Par ailleurs, l'obligation de se soumettre à des autorisations préalables soulève des interrogations quant à l'indépendance de la recherche. Peut-on se satisfaire d'une science "autorisée" qui ne pourrait se saisir de certaines thématiques de recherche qu'avec l'aval des titulaires de droits ?

Une solution à base de licences ne pourrait en outre couvrir qu'une partie des contenus susceptibles de faire du content-mining. Des licences pourraient être demandées pour les publications scientifiques auprès des éditeurs, mais qu'en est-il par exemple de la masse des textes et des données sur Internet pour lesquels il peut être difficile, voir impossible de trouver des titulaires de droits auprès de qui solliciter une licence ? L'analyse par content-mining d'un forum sur Internet, nécessiterait ainsi d'obtenir l'autorisation de tous les contributeur).

Enfin, seuls les universitaires sont pris en compte dans le cadre de ce « deal ». Les membres des communautés en ligne tels que Wikipédia, qui participent manifestement à l'entretien d'un vaste écosystème de la connaissance en sont délibérément exclus.

---

[42] https://ec.europa.eu/licences-for-europe-dialogue/sites/licences-for-europe-dialogue/files/Publishers-Perspective-Initiatives_0.pdf

Un système de gestion collective.

Cette solution permettrait apparemment de contrecarrer plusieurs écueils du système de licence, par le biais d'une licence légale ou d'une gestion collective obligatoire, à l'image par exemple de ce qui existe en France en matière de reprographie.. Les initiateurs d'un projet de content-mining achèteraient un droit à extraire à une sorte d'équivalent scientifique de la SACEM qui redistribuerait ensuite les sommes aux ayant-droit. Un tiers désintéressé ferait ainsi « tampon » entre les éditeurs et les utilisateurs. L'achat d'un droit unique à extraire permettrait de croiser sans difficulté plusieurs corpus distincts. Les méta-données de recherche demeureraient inaccessibles aux éditeurs, ce qui prévient toute exploitation à des fins de marketing et de contrôle.

La mise en œuvre de ce système paraît peu réaliste. Même si les sociétés de gestion collective française ont une expérience assez reconnue en matière de gestion de contenus « complexes », nous sommes ici dans un tout autre ordre d'idée. Le projet text2genome porte à lui seul sur 3 millions d'articles de biologie, sachant que dans cette discipline, les articles sont fréquemment écrits par plusieurs auteurs et qu'ils peuvent reprendre du contenu lui-même protégé par ailleurs. Pour l'instant, des projets de cette ampleur sont encore assez rares, mais tout porte à croire qu'ils vont se multiplier au cours des prochaines années.

Une SACEM du content-mining aboutirait rapidement à une situation ingérable. Les frais d'administration de cette instance risquent d'être considérables. Le droit d'extraire serait ainsi potentiellement surévalué, au risque d'avoir un effet négatif sur le développement de ce type d'outils en France — et, de fait, d'accroître le retard pris sur les États-Unis en la matière. Par ailleurs, pour les contenus en ligne, il sera très difficile d'identifier les titulaires de droits à qui reverser les sommes collectées. Ce système risque donc de générer des montants importants de sommes irrépartissables, ce qui n'est pas de bonne politique.

Une exception

L'exception apparaît nettement comme la solution la plus souple et la plus satisfaisante.

Il s'agit de la voie majoritairement retenue par la réflexion juridique en cours dans plusieurs pays anglo-saxon. La jurisprudence américaine considère désormais que le content-mining intègre le régime du « fair use ». Le gouvernement anglais envisage une exception assez large, qui inclut plusieurs aménagements pour les éditeurs (notamment la

possibilité de limiter l'accès pour garantir la sécurité de la base)[43]. Un rapport irlandais prévoie un cadre assez codifié, dit *fair dealing*[44].

Il existe une amorce de jurisprudence similaire dans le droit français. Bien que déjà ancienne, la décision Microfore c. Le Monde de la Cour de Cassation rendue en matière d'indexation de contenus de presse anticipe assez nettement sur les évolutions du droit américain : les reproductions intégrales sont envisageables sous réserve qu'il ne soit pas possible de reconstituer le contenu d'origine. Sans employer explicitement le terme, la cour a retenu finalement l'hypothèse d'un usage transformatif. Elle autorise « l'analyse purement signalétique réalisée dans un but documentaire, exclusive d'un exposé substantiel du contenu de l'œuvre, et ne permettant pas au lecteur de se dispenser de recourir à cette œuvre elle-même »[45]

En s'appuyant sur la notion « d'œuvre d'information » ou « d'œuvre documentaire », la Cour de Cassation est allée en réalité plus loin qu'une exception. Elle a consacré selon certains une véritable *liberté documentaire*, consacrant que les pratiques d'indexation ou d'extraction de mots-clés ne sont pas soumises au droit d'auteur[46]. Elles ne donnent pas prises aux droits exclusifs des auteurs et de leurs ayants droit, ce qui les situe précisément dans la sphère du *domaine public de l'information* évoquée plus haut.

La notion d'oeuvre d'information de la jurisprudence Microfor pourrait être appliquée aux pratiques de content mining. D'une certaine manière, la Cour de Cassation a raisonné dans cette affaire comme l'ont fait juges américains sur la base du fair use. Pourrait-on s'appuyer sur cette jurisprudence pour introduire des éléments de fair use dans la loi française ?

Ce n'est pas si probable. Le fair use demeure relativement étranger à notre tradition juridique. Les quelques tentatives allant en ce sens sont demeurées assez limitées. L'exception pédagogique fait ainsi l'objet de quantité d'exceptions à l'exception qui limitent considérablement sa portée.

La voie irlandaise semble peut-être plus probante. Par contraste avec le fair use, le « fair dealing » ne vise pas à énoncer des principes généraux valables pour tous les

---

[43] *Modernising Copyright : A modern, robust and flexible framework* , Government response to consultation on coppyright exceptions and clarifying Copyright Law, Londres, 2012, pp. 36-37, http://www.ipo.gov.uk/response-2011-copyright-final.pdf

[44] *Modernising Copyright* , The Report of the Copyright Review Committee, Dublin, 2013, pp. 85-88, http://www.enterprise.gov.ie/en/Publications/CRC-Report.pdf

[45] Arrêt de la cour de cassation du 9 novembre 1983

[46] Didier Frochot, « L'affaire Microfor / Le Monde », article de 1988 republié sur le site *infostrateges.com,* http://www.les-infostrateges.com/article/880432/affaire-microfor-le-monde

usages, mais un nombre limité d'usages strictement explicités dans la loi. Les projets scientifiques, pédagogiques et non-commerciaux pourraient ainsi bénéficier d'une sécurité juridique qui couvrirait tous les aspects du content-mining. Il serait évidemment possible d'aménager ultérieurement cette liste limitative au gré des évolutions techniques et et des intérêts collectifs. Les provisions prévues par le projet irlandais[47] incluent :

(1) La nécessité de ne pas attenter aux intérêts du détenteurs des textes ou des données : le projet ou la pratique de content-mining doit être non-commercial.

(2) L'association du droit d'extraction au droit de lire : le lecteur ne peut pas extraire un contenu qu'il ne peut pas lire.

(3) La présentation de portions du contenu protégé dans la limite du droit de courte citation.

(4) L'autorisation de réaliser des copies temporaires. Le terme « temporaire » est à entendre dans un sens plus généreux que celui utilisé par la directive européenne sur la société de l'information : les reproductions peuvent être conservées tant que dure le projet de content-mining

Ce « fair dealing » pourrait servir d'inspiration pour un éventuel cadre européen.

Il offre un compromis acceptable pour l'ensemble des acteurs : les éditeurs ne perdent en rien le contrôle de leurs contenus qui ne peuvent pas être republiées, ni réutilisées à des fins qui leur seraient préjudiciable commercialement. Le rapport du gouvernement anglais sur la modernisation du copyright soulignent qu'ils pourraient même profiter de la situation : une libéralisation du content-mining contribuerait à revaloriser des publications scientifiques en leur conférant une nouvelle utilité[48].

Tout en encadrant strictement les usages autorisé, le « fair dealing » irlandais ne trace pas de limites franches entre une utilisation académique et une utilisation non-académique. Une distinction de ce type a de moins en moins de sens aujourd'hui. La connaissance s'élabore également en dehors des enceintes universitaires, au sein de communautés associatives et participatives. Rien ne justifierait de les exclure de ce cadre.

Par ailleurs, en application du domaine public de l'information, un content-mining qui n'outre-passerait pas les limites autorisées par les protections juridiques en vigueur des bases de données serait explicitement garanti.

---

[47] *Modernising Copyright* , The Report of the Copyright Review Committee, Dublin, 2013, pp. 86-87, http://www.enterprise.gov.ie/en/Publications/CRC-Report.pdf

[48] *Modernising Copyright : A modern, robust and flexible framework* , Government response to consultation on coppyright exceptions and clarifying Copyright Law, Londres, 2012, p. 37, http://www.ipo.gov.uk/response-2011-copyright-final.pdf

Enfin, compte tenu de l'importance des activités de recherche entreprise sur la base du content-mining pur la société toute entière, il est fondamental que cette exception ne fasse pas l'objet d'une compensation financière au profit des titulaires de droits. Le content-mining ne constitue pas un "préjudice" dont pourrait se prévaloir les titulaires de droits pour exiger un paiement, de la même manière que les citations ne font pas l'objet de compensation.

Recommandations

Cette section ne fait qu'exposer brièvement les principales conclusions de cette synthèse.

D'une manière générale, les pays européens offrent un cadre légal beaucoup moins favorable au content-mining que les États-Unis. Ces incertitudes actuelles risquent de pénaliser une activité émergente, appelée à recevoir de nombreuses applications sociales, intellectuelles et sanitaires au cours des années à venir.

L'action du législateur peut être envisagée dans deux directions : formaliser le domaine public de l'information ; mettre en place un « fair dealing » encadrant certaines pratiques de content-mining apparemment incompatibles avec le code de la propriété intellectuelle. La première disposition peut être réalisée à la seule échelle française. La seconde nécessiterait une coordination européenne.

Enfin plusieurs recommandations complémentaires contribuerait à encourager le développement du content-mining à des fins de recherche en France.

Une définition explicite du domaine public de l'information

Au cours de l'année passée, une réflexion importante a été engagée en France autour du domaine public. Bien que constitutive de notre conception de la propriété intellectuelle, cette notion n'a jamais été explicitement définie dans la loi. Cette définition négative se traduit par de nombreuses pratiques abusives, qualifiée de « copyfraud » : des entreprises, des organisations ou des institutions publiques revendiquent des droits sur des contenus placés dans le domaine public. Il s'agit en quelque sorte d'un piratage inversé : au lieu de mettre en circulation un bien protégé, on s'approprie un bien commun[49].

---

[49] Pierre-Carl Langlais, « L'inverse du piratage c'est le copyfraud et l'on n'en parle pas », *Rue89*, 14 octobre 2012, http://blogs.rue89.nouvelobs.com/les-coulisses-de-wikipedia/2012/10/14/linverse-du-piratage-cest-le-copyfraud-et-personne-nen-parle

L'une des 80 propositions du rapport Lescure vise à consolider la définition du domaine public afin de mettre un terme à ces abus. Elle souligne que « la consécration d'une définition positive du domaine public ne relève pas d'une logique purement symbolique : elle permettrait de renforcer la protection du domaine public face aux menaces que différentes pratiques, notamment dans le champ numérique, font peser sur lui »[50]. En octobre dernier, la députée Isabelle Attard et le collectif Savoirscom1 ont consacré une journée d'étude au domaine public à l'Assemblée nationale[51]. Cette réflexion a débouché un mois plus tard sur un projet de loi ambitieux, qui vise non seulement à formaliser une définition positive du domaine public mais inclut plusieurs mesures destinées à élargir sa portée et à garantir sa préservation[52].

Ces dispositions pourraient aisément intégrer l'affirmation d'un domaine public informationnel, selon la définition proposée par l'UNESCO : « le domaine public informationnel inclut toutes les informations, les faits et les données publiquement accessibles ». Cette précision ne changerait rien à la philosophie générale du code de la propriété intellectuelle, qui intègre de facto son existence en se focalisant sur des œuvres et des créations de l'esprit marqués par une expression originale. Elle apporterait cependant une sécurité juridique importante en liant clairement le droit de lire au droit d'extraire.

Introduire une définition positive du domaine public dans la loi permettrait de faciliter l'accès aux corpus numérisés par les institutions culturelles françaises pour des recherches de type data ou text mining. A titre d'exemple, les contenus de la bibliothèque numérique Gallica ne peuvent aujourd'hui être pleinement utilisés de cette manière, car les conditions d'utilisation du site revendiquent l'application du droit des bases de données qui peut s'avérer problématique[53]. Pour que le téléchargement du contenu soit possible, il est nécessaire d'approuver manuellement une clause de diffusion non-commerciale, ce qui complique considérablement le recours à des outils d'extraction automatisés.

---

[50] Pierre Lescure, *Contribution aux politiques culturelles à l'ère numérique,* t. I, p. 452

[51] http://www.savoirscom1.info/2013/10/02/31-octobre-une-journee-detude-sur-le-domaine-public-a-lassemblee-nationale/

[52] Projet de loi visant à consacrer le domaine public, à élargir son périmètre et à garantir son intégrité, enregistrée le 21 novembre 2013, http://www.assemblee-nationale.fr/14/propositions/pion1573.asp

[53] http://gallica.bnf.fr/html/conditions-dutilisation-des-contenus-de-gallica

Un « fair dealing »

L'affirmation du domaine public de l'information ne résout pas tout. Elle contribue à reconnaître et clarifier une situation fondamentale : dans la mesure où il ne porte que sur des faits le content-mining se situe hors du champ de la propriété intellectuelle. Cependant, sa mise en œuvre contredit potentiellement plusieurs droits. En particulier, la reproduction intégrale ou substantielle du contenu à extraire s'avère bien souvent indispensable. Cette reproduction n'est pas publiée mais elle est techniquement accessible à un groupe plus ou moins large de chercheurs ou de contributeurs. Dans cette même logique, certains services permettent d'effectuer des requêtes sur une reproduction d'un contenu protégé, sans la donner à lire.

Trois options sont actuellement envisagés pour combler les nombreuses incertitudes juridiques auxquelles font face les pratiques de content-mining. Les deux premières paraissent problématiques à plusieurs niveaux. Un système de licences contractuelle risque de se traduire par une perte de temps notable pour les chercheurs, astreints à remplir de nombreux formulaires ; il contribue à affermir un pouvoir de contrôle et de surveillance via la collecte de nombreuses méta-données ; il ne concerne que les membres d'une institutions universitaires et exclut de fait les nombreux participants au processus de recherche scientifique issus de la société civile. Le système de gestion collective corrige la plupart des limitations du système de licences contractuelles. Il s'avère cependant ingérable en raison de la vaste ampleur de certains projets de content-mining, qui portent potentiellement sur des millions de publications scientifiques.

La troisième option, celle de l'exception, apparaît clairement comme la solution la plus souhaitable. Sans aller jusqu'à mettre en œuvre un « fair use », il paraît concevable de développer un « fair dealing », soit un régime d'exception valable si quelques conditions précises sont respectées (et non pas juste un ensemble de principes généraux). Le fair dealing irlandais apporte plusieurs pistes judicieuses : autorisation des copies temporaires et de la présentation de courts extraits dans le respect du droit de courte citation pour tous les projets ne s'opposant pas aux intérêts commerciaux des propriétaires du contenu.

La réalisation de ce fair dealing ne peut être décidée en France : elle excèderait la marge de manœuvre accordée par la directive sur la société de l'information. On observe cependant une convergence de plusieurs pays européens sur cette question. L'Irlande et le Royaume-Uni préparent une exception spécifique sur le content-mining. L'Allemagne a commencé à constituer un cadre légal assez ambitieux pour les publications en libre accès (en particulier, toute publication financée en majorité par l'argent public, peut être

publiée sous une licence non commerciale après un embargo de 12 mois)[54]. Une exception sur le content-mining se situerait dans le prolongement directe de cette politique.

Le gouvernement français ne devrait ainsi pas manquer d'interlocuteurs sur cette question. La concrétisation d'un « fair dealing » européen sur le content-mining et même, plus largement, sur l'utilisation générale des contenus protégés à des fins scientifiques et pédagogiques pourrait s'avérer plus rapide que prévu. Le gouvernement français devrait saisir l'opportunité de la consultation lancée par la Commission européenne sur la révision du droit d'auteur pour demander à ce qu'une telle exception en faveur du data mining soit introduite dans la directive. L'Union européenne disposerait alors de tous les outils légaux pour accueillir et favoriser une révolution scientifique de grande ampleur.

Autres recommandations

Même sans introduire de nouvelle exception, plusieurs mesures permettraient d'élargir rapidement les corpus de contenus mobilisables par les chercheurs pour faire du content-mining :

— Elargir les conditions d'accès aux Archives du web, rassemblées par le biais du dépôt légal de l'Internet par la BnF et l'INA[55]. Les robots utilisés par ces établissements dans le cadre du dépôt légal du web rassemblent des contenus qui par définition se prêtent à des recherche de type content-mining. Cependant, les conditions de consultation de ces corpus sont aujourd'hui trop restrictives (uniquement sur place dans les emprises de ces établissements). La BnF a profité de la parution du décret d'application de la loi pour élargir la consultation à des BDLI (Bibliothèque de dépôt légal imprimeur). Mais cette extension est encore insuffisante Il faudrait que les Archives du web puissent être consultées dans les grandes bibliothèques universitaires et de recherche du pays pour faciliter l'accès aux chercheurs et rentabiliser les investissements considérables que représentent l'archivage du web aujourd'hui.

— Favoriser le développement du libre accès et clarifier les conditions de réutilisations des archives ouvertes. Il est évident que plus les articles déposés en Open Access dans des archives ouvertes seront nombreux, plus les possibilités d'étudier ces contenus par le biais du content-mining seront facilités. Néanmoins actuellement, les conditions

---

[54] Herbert Gruttemeier, « Point sur le Libre Accès en Allemagne », *openaccess.inist.fr*, 18 novembre 2013, http://openaccess.inist.fr/?Point-sur-le-Libre-Acces-en

[55] Présentation des Archives du web sur le site de la BNF, http://www.bnf.fr/fr/collections_et_services/livre_presse_medias/a.archives_internet.html

d'utilisation des archives ouvertes en France sont loin d'être claires. Une plateforme comme HAL par exemple est placée par défaut sous le régime du droit d'auteur et du droit des bases de données[56]. Les archives ouvertes devraient être placées sous licence libre pour permettre des usages comme le content-mining dans de bonnes conditions. La question se pose pour HAL, mais aussi sur d'autres sites comme Persée, Revues.org, Hypothèses.org.

— Ouvrir le corpus des livres indisponibles du XXe siècle au content-mining. La loi du 1er mars 2012 sur l'exploitation des livres indisponibles du XXe siècle a prévu la mise en place d'un système de gestion collective pour favoriser la numérisation et la réédition d'un nombre important d'ouvrages publiés au XXe siècle[57]. Le corpus ainsi constitué aurait pu devenir une plateforme intéressante pour la recherche, notamment pour des pratiques de content-mining. Or la loi n'a prévu que la recommercialisation de ces usages et pas des usages en faveur de la recherche. Le texte pourrait être modifié afin d'offrir des possibilités de content-mining sur le corpus. On notera qu'à défaut, un hiatus important apparaîtra entre la France et les Etats-Unis puisque Google Books contient essentiellement des ouvrages indisponibles, qui pourront faire l'objet de data mining par les chercheurs sur la base du fair use.

— Interdire l'utilisation de droits connexes pour empêcher la réutilisation du domaine public numérisé. La France est l'un des pays où la numérisation du patrimoine est la plus avancée, que ce soit dans les bibliothèques, les archives ou les musées. Or comme expliqué plus haut, l'acte de numérisation est souvent pris pour prétexte pour revendiquer de nouveaux droits sur le domaine public numérisé. De telles pratiques devraient être interdites pour garantir l'intégrité du domaine public sous forme numérique. L'un des bénéfices qui pourrait être retiré de cette démarche serait d'ouvrir au content-mining de grands ensembles de contenus, particulièrement utiles par exemple au développement des humanités numériques en France.

---

[56] Lionel Maurel, « Un open access sans licence libre a-t-il un sens ? », *S. I. Lex*, 4 novembre 2013, http://scinfolex.com/2013/11/04/un-open-access-sans-licence-libre-a-t-il-un-sens/

[57] Loi n°2012-287 du 1er mars 2012, http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000025422700&categorieLien=id