couperin.org
Consortium Unifié des Etablissements Universitaires et de Recherche pour l'Accès aux Publications Numériques

Association des Directeurs & personnels de direction des Bibliothèques Universitaires et de la Documentation

CSPLA - Mission relative au *data mining* (exploration de données) : l'analyse de Couperin et de l'ADBU

1. La problématique des données pour la recherche

A. Les données numériques : enjeu majeur de la recherche d'aujourd'hui

L'exploitation de l'extraordinaire masse de données aujourd'hui produites, qu'elles soient nativement numériques ou obtenues par numérisation, constitue actuellement pour la recherche probablement la plus prometteuse des perspectives inaugurées par les révolutions digitale et des réseaux, au point que l'on parle même de plus en plus de « *data driven innovation* » ou de « *data driven science* ».

On connaît bien le versant économique de cette évolution, à travers la promesse du *Big data*, qui exploite les données personnelles des internautes, ce qui, même en anonymisant ces dernières, soulève des questions éthiques et de libertés publiques qui n'entrent pas de le champ des réflexions de cette note, même si les pratiques de recherche sont également concernées. De cette révolution des données, le grand public connaît moins en revanche le versant académique : il ouvre pourtant des possibilités inédites au travail scientifique, extrêmement prometteuses pour la recherche. En effet :

- l'instrumentation scientifique est entrée dans l'âge du numérique, avec pour corollaire la création de masses considérables de données : l'astronomie, la physique des hautes énergies et des particules en sont les illustrations les plus connues ;
- dans le domaine des sciences humaines, les humanités numériques (*digital humanities*) renouvellent l'approche des textes et corpus, tout comme en sciences sociales, la possibilité de fouiller de grandes masses de données, parfois issues de plusieurs études ou enquêtes statistiques (voire des archives du Web), en sociologie, en économie[1], permettent de réaliser des avancées

1 Un exemple parmi d'autres : « Par comparaison aux travaux antérieurs, la première nouveauté de la démarche développée ici est d'avoir cherché à rassembler des sources historiques aussi complètes et systématiques que possible afin d'étudier la dynamique de la répartition des richesses. Il faut souligner que j'ai bénéficié pour cela d'un double avantage par rapport aux auteurs précédents : nous disposons par définition d'un recul historique plus important […] **et nous avons pu, grâce aux possibilités nouvelles offertes par l'outil informatique, rassembler sans peine excessive des données historiques à une échelle beaucoup plus vaste que nos prédécesseurs.**
Sans chercher à faire jouer un rôle exagéré à la technologie dans l'histoire des idées, il me semble que ces questions purement techniques ne doivent pas être totalement négligées. Il était objectivement plus difficile de traiter des volumes importants de

considérables, dans une économie de moyens inenvisageable jusqu'alors. C'est aussi le cas en sciences dures, par exemple en génomique, avec la fouille d'immenses quantités d'articles scientifiques pour en extraire les données de séquençage qu'ils recèlent (projet Text2genome, exemple archétypal s'il en est) ;

- la science s'ouvre par ailleurs de manière croissante au grand public, autour de grands enjeux sociétaux tel l'environnement, et s'appuie sur l'intérêt du grand public pour ces problématiques, afin là encore de créer des corpus de données inenvisageables il y a seulement 10 ans : les collectes par *crowdsourcing* autour de la biodiversité se multiplient par exemple dans le monde ;
- enfin le TDM (*text and data mining*) offre également des possibilités que l'on commence à peine à exploiter, de fouille des corpus et données scientifiques. Les applications de cette lecture computationnelle sont infinies, et loin d'être toutes explorées. Citons malgré tout les progrès permis par le TDM dans l'analyse linguistique, la recherche de « signaux faibles » difficiles à appréhender par la lecture cursive, le repérage et l'analyse de comptes rendus d'expérimentations ratées (« les résultats négatifs de la science »), ce afin d'éviter d'engager des moyens dans des voies sans issues (enjeu majeur, aux implications financières lourdes, par exemple aujourd'hui dans les domaines de la recherche médicale et pharmaceutique).

B. Le contexte du Web de données

Au-delà de ce renouvellement des pratiques, les promesses portées par ce qu'on appelle parfois le Web 3.0, ou Web sémantique, ou plus communément aujourd'hui le Web de données, sont considérables : là où à ses débuts le Web reliait entre elles des machines, puis, avec l'apparition du langage HTML et de l'hypertexte, des documents, il est aujourd'hui possible de lier entre elles, via un formalisme de niveau supérieur, les unités d'information les plus discrètes, à savoir les données elles-mêmes, et de s'affranchir des questions de formats, de langues, d'environnements professionnels et techniques, par l'alignement des référentiels existants ou en création. Bref, de rendre toute donnée structurée actionnable et interopérable hors même son contexte de production : c'est ouvrir la possibilité à des réutilisations infinies et non-programmées de l'immense masses des données existantes, avec toutes les applications commerciales, scientifiques, citoyennes imaginables... sous réserve que ces données soient rendues disponibles, c'est-à-dire qu'elles soient autant que possibles ouvertes. Sans cette ouverture maximale (*modulo* la protection des données personnelles, du secret industriel, etc.), pas de Web de données, pas de levier de compétitivité.

Le Web de données fait entrer l'informatique dans une nouvelle étape de son histoire : il change absolument la donne quant à la manière d'agréger des données. Il y a peu, deux méthodes permettaient de procéder à l'interrogation satisfaisantes de jeux de données hétérogènes :

- soit l'on recourait, pour une interrogation multibases, à un métamoteur, qui mimait en quelque sorte l'interrogation manuelle de chaque base considérée, et agrégeait au mieux les résultats qui en étaient issus dans une liste unique. Mais ce au prix d'un appauvrissement de la requête : seules les fonctionnalités de recherche communes à chacun des moteurs des bases interrogées étaient mobilisables par le métamoteur ;

---

données historiques à l'époque de Kuznets, et dans une large mesure jusqu'aux années 1980-1990, qu'il ne l'est aujourd'hui. Quand Alice Hanson Jones rassemble dans les années 1970 des inventaires au décès américains de l'époque coloniale, ou quand Adeline Daumard fait de même avec les archives successorales françaises du XIX[e] siècle, **il est important de réaliser que ce travail s'effectue pour une large part à la main, avec des fiches cartonnées. Quand on relit aujourd'hui ces travaux remarquables,** ou bien ceux consacrés par François Simiand à l'évolution des salaires au XIX[e] siècle, par Ernest Labrousse à l'histoire des prix et des revenus au XVIII[e] siècle, ou encore par Jean Bouvier et François Furet aux mouvements du profit au XIX[e] siècle, **il apparaît clairement que ces chercheurs ont dû faire face à d'importantes difficultés matérielles pour collecter et traiter leurs données. Ces complications d'ordre technique absorbent souvent une bonne part de leur énergie et semblent parfois prendre le pas sur l'analyse et l'interprétation, d'autant plus que ces difficultés limitent considérablement les comparaisons internationales et temporelles envisageables. Dans une large mesure, il est beaucoup plus facile d'étudier l'histoire de la répartition des richesses aujourd'hui que par le passé. Le présent livre reflète en grande partie cette évolution des conditions de travail du chercheur.** », PIKETTY Thomas. Le capital au XXI[e] siècle. Paris : Éditions du Seuil, 2013, p. 45-46.

- soit l'on décidait de procéder, préalablement à l'interrogation des bases, à un retraitement des données afin de les homogénéiser, par exemple à travers un format pivot commun. Mais là encore, outre des temps de transaction (normalisation du format pivot) et de retraitement très longs, le résultat obtenu ne pouvait jamais prétendre qu'à un appauvrissement du matériau original, le format pivot étant contraint de retenir le plus grand dénominateur commun (souvent bien mince).

L'on se trouvait face à un véritable dilemme : on l'on disposait de données riches, mais impossibles à agréger à des jeux de données hétérogènes, ou l'on agrégeait mais en appauvrissant le tout. Le Web de données, par le recours à un formalisme de niveau supérieur dans la description de l'information, met fin à ce dilemme : il est désormais possible, en procédant préalablement à des alignements de référentiels, d'agréger entre elles des données hétérogènes en en conservant la totalité de la richesse, voire même, le jeu des alignements opérant souvent par transitivité, en établissant des relations inédites entre un jeu de données A et un jeu de données C, via les liens établis entre ces deux entités et un jeu de données B.

C'est dans ce contexte que prend place aujourd'hui la question du TDM et c'est ce que manquent dans leurs propositions les éditeurs avec leurs offres contractuelles (ce qui est un premier élément de réponse quant à l'inadaptation de ces offres aux besoins de la recherche) : la plupart des projets scientifiques s'appuyant sur la technologie du TDM ont besoin d'analyser de manière **croisée et simultanée (et non pas successive)** des jeux de données en provenance de plusieurs bases, de plusieurs éditeurs, voire du Web lui-même. Et de retraiter ces données pour procéder à des alignements permettant d'en assurer la fouille homogène sans que rien ne soit perdu de leur richesse initiale : c'est par exemple un aspect essentiel du projet ISTEX (http://www.istex.fr/) et l'une des raisons pour lesquelles il fédère autant d'acteurs et, côté opérateurs, autant de compétences diverses.

2. TDM : que proposent les éditeurs ?

Le point de vue des éditeurs est aujourd'hui que les pratiques de TDM constituent une nouvelle exploitation de contenus sous droits et nécessitent de ce fait la mise en œuvre de licences spécifiques, donnant lieu à paiement, en sus des licences déjà acquittées par les clients pour accéder aux contenus des bases de données. L'approche contractuelle présente un intérêt évident pour tous les acteurs : en tentant de réguler l'offre et la demande, elle évite des transactions coûteuses en temps et en énergie. Pour autant, malgré cet avantage, elles ne constituent pas la bonne solution en la matière.

En effet, l'examen des licences proposées à ce jour permet sans nul doute d'avancer qu'en l'espèce, la voie contractuelle constitue pour le TDM une impasse davantage qu'une solution.

A. La voie contractuelle est inadaptée aux besoins et aux pratiques de la recherche

Il n'existe pas de licence globale ou de proposition d'une API mutualisée entre les acteurs de l'édition et du Web, et il n'y en aura pas avant longtemps. Or, ce dont ont besoin les chercheurs, c'est de pouvoir fouiller librement tout corpus de documents ou de données, y compris le Web (dont bon nombre des contenus sont sous droits, et impliquent une multitude d'ayants-droits avec lesquels il faudrait s'accorder : coûts de transaction inouïs). Loin de permettre ce type de travail, une licence comme celle proposée par Elsevier pour ScienceDirect est encore plus restrictive : elle permet via une API fournie par l'éditeur de charger 10 000 articles de *ScienceDirect* par semaine, **là où les besoins de la recherche, répétons-le, sont d'une fouille simultanée (et non successive) de jeux de données en provenance de plusieurs bases, de plusieurs éditeurs, du Web, ce qui exclut des solutions praticables la voie contractuelle, même en imaginant une amélioration des licences existantes, et leur harmonisation[2].**

2 Ainsi de la licence-type proposée par les éditeurs académiques en STM au niveau européen : http://www.stm-assoc.org/text-and-data-mining-stm-statement-sample-licence/

Le consortium Couperin, qui vient de négocier avec Elsevier une licence nationale permettant la pratique du TDM sous les conditions que l'on vient de décrire, souhaite à travers cette disposition identifier les usages que les chercheurs feront de cette fonctionnalité. Il estime que les possibilités de TDM proposées par Elsevier se relèveront foncièrement inadaptées aux pratiques de recherche.

En effet, d'une part l'innovation se construit le plus souvent aux marges et croisements des disciplines, dans des approches de plus en plus interdisciplinaires : le besoin de croiser des bases de données hétérogènes, issues d'éditeurs différents, est fondamental, et par ailleurs il sera certainement nécessaire d'associer les contenus issues des publications avec des données primaires issues de l'activité scientifique (données d'instruments notamment). Par ailleurs, la recherche chemine par hypothèses et tests, essais et erreurs. Imaginons qu'une hypothèse soit examinée sur X jeux de 10 000 articles de ScienceDirect (à supposer que le seul corpus de ScienceDirect suffise à une étude), imaginons que sur cette base de X fois 10 000 articles, cette hypothèse ait été invalidée. Une nouvelle hypothèse est formulée. Elle devra être à nouveau éprouvée sur les X fois 10 000 premiers articles. Imaginons qu'à l'issue de cette étape cette nouvelle hypothèse soit validée : est-on certain que l'examen du reste du corpus de ScienceDirect (ou de la part jugée significative de ce corpus) permettra de valider elle aussi la nouvelle hypothèse ? Il faudra télécharger et fouiller les Y jeux de 10 000 articles nécessaires pour se faire une idée, et sans jamais pouvoir fouiller en une fois l'ensemble du corpus, ce qui limite les possibilités de découverte de signaux particulièrement faibles, et ralentit considérablement le processus de recherche. Le tout n'est jamais égal à la somme des parties.

Un projet comme Text2genome, impliquant la fouille de millions d'articles issues de bases éditoriales hétérogènes, aurait été impossible à conduire avec la solution proposée par Elsevier. C'est d'ailleurs pourquoi il a été conduit sur d'autres bases, en négociant au cas par cas, avec chaque éditeur, l'accès à leurs données pour TDM. Cela a pris trois ans. Quand on connaît les contraintes en temps auxquelles est soumise la recherche sur projets, un tel coût de transaction apparaît inconcevable dans la plupart des travaux de recherche.

Imaginons par ailleurs que des chercheurs aient besoin d'accéder à deux bases de données selon les modalités définies par Elsevier : il leur faudra d'abord fouiller une première tranche de 10 000 articles issus d'une base A avec une première tranche de 10 000 articles issus d'une base B, puis avec une deuxième tranche de cette base B, etc., puis prendre une deuxième tranche de la base A et recommencer le même processus avec toutes les tranches intéressantes de la base B, etc. Impraticable. Surtout lorsqu'on sait qu'un million et demie d'articles revus par les pairs sont produits chaque année dans le monde par le secteur académique. Et ne parlons pas de la croissance exponentielle du Web, et des données primaires nativement numériques issues de l'activité scientifique, en croissance exponentielle..

Certains éditeurs ont tenté de justifier ces restrictions par des arguments techniques : il s'agirait moins de limiter l'accès à leurs contenus que de ménager la capacité de traitement de leurs serveurs. L'éditeur *open access* PloS (*Public Library of Science*), qui autorise très libéralement la fouille de ses données, atteste n'avoir jamais rencontré de problèmes avec ses serveurs du fait de l'activité de TDM des chercheurs.

B. La voie contractuelle introduit des dispositifs menaçant l'indépendance de la recherche

La voie contractuelle soulève des interrogations de fond quant à la légitimité des procédures entourant la signature des contrats : l'exemple de Text2genome montre la nécessité actuelle, lors des négociations avec les éditeurs, d'exposer son projet de recherche et sa méthodologie afin d'emporter l'accord des ayants droits. Cela revient à donner aux éditeurs de contenus, qui ne sont pour rien dans le financement des projets de recherche, le droit de décider quelle recherche pourra ou non voir le jour : cette situation est **inacceptable du point de vue de l'indépendance de la science**.

Cette analyse ne relève pas d'un scénario théorique : Springer par exemple[3] impose que chaque projet de TDM portant sur ses contenus soit décrit et enregistré via un formulaire en ligne. L'éditeur se réserve ensuite le droit de décider si la demande lui semble ou non fondée. Au-delà du poids inconsidéré qui serait alors donné aux éditeurs dans l'orientation des politiques de recherche, à rebours de tout principe décideur-payeur, ce système permettrait également à des acteurs privés de connaître des travaux de recherche en cours en sus des recherches passées décrites dans les articles qu'il publie. Or ce type d'informations est aujourd'hui éminemment monnayables dans le cadre des systèmes qui émergent de pilotage de la recherche (repérage, pour une unité de recherche donnée, et au niveau mondial, de ses concurrents, de ses alliés potentiels, des coopérations à développer, etc.) dont on sait bien qu'il est le nouvel horizon commercial des grands éditeurs académiques tels Thomson Reuters ou Elsevier, sur un marché de la publication scientifique aujourd'hui saturé et dont on a tiré tout ce qu'il était possible dans le cadre d'une situation oligopolistique dénoncée depuis des décennies par tous les acteurs de la recherche.

C. La voie contractuelle proposée par les éditeurs limite les droits prévus par la loi

En matière de TDM, la voie contractuelle constitue également le Cheval de Troie d'offensives à couvert contre les exceptions actuelles au droit d'auteur : ainsi la licence TDM d'Elsevier impose-t-elle de publier les extraits retenus par les chercheurs suite à leur fouille de donnée sous licence CC-BY-NC, comme si des données pouvaient désormais relever du domaine couvert par la propriété intellectuelle. Sans même parler d'opérations de valorisation de la recherche, une telle clause contraint de façon inacceptable là encore le simple processus de validation scientifique, qui repose dans bien des disciplines sur la reproductibilité des résultats obtenus tels que décrits dans la publication initiale. Par ailleurs ces mêmes extraits se voient limités contractuellement à 350 mots, là où la longueur d'une citation, selon la directive communautaire 2001/29/CE et la Convention de Berne, doit être apprécié « dans la mesure justifiée par le but poursuivi »[4]. **Ces manœuvres sont également inacceptables : la citation est à la base même de tout travail académique et l'exception qui l'autorise ne saurait être aussi étroitement bornée, et certainement pas au-delà de ce que prévoit la loi**[5].

D. La voie contractuelle porte la menace d'abus de position dominante

Il peut être tentant, pour une industrie de l'édition académique qui atteint aujourd'hui les limites de son modèle de rentabilité, d'envisager de développer une offre de service de fouille de données à façon sur les corpus qu'elle commercialise. Cette virtualité est grosse de menaces pour la liberté du commerce : jouissant du monopole légal emporté par le droit de la propriété intellectuelle, les éditeurs académiques seraient en effet en position de barrer l'entrée d'acteurs extérieurs sur ce marché dérivé de services, ou de leur imposer un ticket d'entrée exorbitant. Or, c'est précisément cette capacité d'un pouvoir de marché de bloquer ou par trop contraindre l'entrée de nouveaux acteurs sur un marché dérivé qui constitue le **risque d'abus de position dominante.**

3. Les questions juridiques soulevées par le TDM

Les interrogations suscitée par la pratique du TDM reviennent en fait à une seule et même question : comment garantir aux éditeurs la légitime protection de leurs contenus sans empêcher ou excessivement contraindre l'activité de recherche d'aujourd'hui, pour laquelle le TDM est une pratique déjà devenue irremplaçable ?

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

4 Est-il utile de souligner que la note 1 du présent document excède 350 mots mais que son ampleur (toute relative : l'ouvrage de Thomas Piketty fait près de 1 000 pages) est justifiée par le but poursuivi ?

5 La licence proposée par Elsevier pose de nombreux autres problèmes dont une remarquable analyse a été conduite par la Ligue européenne des bibliothèques de recherche (LIBER) : http://www.libereurope.eu/sites/default/files/TDMdiscussionpaper-final.pdf

On l'a vu, la voie contractuelle est inapte à répondre à ce défi. Les questions juridiques soulevées par le TDM doivent donc être reprises à nouveaux frais, afin de dégager une solution efficace et opératoire pour toutes les parties prenantes.

Les arguments susceptibles d'être invoqués par les éditeurs pour défendre juridiquement leur position sont au nombre de trois.

A. Un nouvel usage implique de nouveaux droits

Extraire des données, pour les synthétiser ou les réutiliser dans le cadre d'un travail intellectuel, est une activité à la base même de l'activité scientifique, et qui préexistait à la création des bases de données informatiques.

Cette pratique est en effet largement attestée dans le contexte analogique : le relevé manuel d'informations est à l'origine de travaux académiques dès les progrès des mathématiques dans le domaine des grands nombres et de la statistique, au XVIII^{e} siècle : la citation de la note 1 du présent document le rappelle. Antérieurement à cette période, la tradition philologique et codicologique connaît bien la pratique des tables de concordance, qui n'est rien d'autre que du *text mining* manuel. Enfin, toute grande bibliothèque dispose encore d'une salle de référence, où exercent des bibliothécaires spécialisés, dits bibliographes, capables du fait de la connaissance approfondi qu'ils ont de leur fonds de tracer entre les documents d'orientation qu'ils proposent des liens non directement perceptibles par les lecteurs fréquentants : de la mise en relation de signaux faibles avant la lettre.

Le TDM permet seulement d'automatiser et donc de rendre plus efficient ce type de travail. Certes, la technologie peut introduire des différences de nature : c'est la base même qui a par exemple présidé en France à la Loi informatique et libertés. Le législateur a considéré que la possibilité de croiser aisément, grâce à l'informatique, des fichiers contenant des données personnelles facilitait l'atteinte aux libertés publiques et justifiait des dispositions législatives contraignantes. Mais dans le cas présent, la technologie n'introduit qu'une différence de degré dans l'activité, non une différence de nature : on montrera que le TDM n'est à l'origine d'aucun préjudice à l'égard des ayants droits.

La pratique du TDM ne consiste donc pas pour les chercheurs à exercer un nouveau droit, mais à poursuivre par des moyens technologiques modernes une activité très ancienne, intrinsèquement liée à l'activité de lecture savante. **Le TDM n'est rien d'autre qu'une manière de lire et d'exploiter l'information, caractéristique des pratiques de lecture intensive propres au monde académique.** Comme le souligne avec beaucoup de bon sens LIBER dans l'analyse déjà évoquée en note 5 de ce document : « *We believe that universities should be able to employ computers to read and analyse content they have purchased and to which they have legal access. An e-subscription fee is paid so that universities can appropriately and proportionately use the content they subscribe to. For what other purpose is a university buying access to information ?* ».

B. Les bases de données sont protégées par le droit de la propriété intellectuelle

C'est en effet le cas, et à double titre :

- d'une part sont protégés au sein des bases de données les contenus sous droits, c'est-à-dire des œuvres, comme en contiennent un certain nombre de bases de données textuelles (dans les bases de données constituées de seules données, ces données ne sont pas protégées en tant que contenus : données, faits, idées sont de libre parcours) dont les éléments relèvent de la propriété littéraire et artistique (corpus de textes littéraires, base d'articles scientifiques, etc.) ;
- d'autre part, au terme de l'article L112-3 du Code de la propriété intellectuelle, les bases de données en tant que telles, indépendamment de la nature de leurs éléments, constituent des créations

intellectuelles « par le choix ou la disposition des matières » : comme le précise l'arrêt C-444/02 - *Fixtures Marketing* de la Cour de Justice de l'Union européenne, c'est la structure de la base de données et l'extension atteinte par le travail de collecte des données qui fondent le caractère original d'une base de données, et constitue l'ensemble qu'elle représente en œuvre, et non en elles-mêmes les données qui la composent.

Pour autant il importe de noter que l**e processus même du TDM revient à dissoudre ce qui fait l'originalité même d'une œuvre, au sens de la propriété intellectuelle, à savoir l'expression originale d'un contenu** : ne reste suite au processus de TDM qu'un ensemble d'éléments à plat, qu'il s'agisse de tables ou de listes, dont la structure originelle, fondatrice du caractère original de l'œuvre, est perdue. On ne saurait par exemple considérer qu'une liste de mots issus d'un texte littéraire constitue une œuvre : ce sont les mots de la langue, ils appartiennent à tout le monde. C'est la manière de les agencer de façon originale qui fait œuvre. Or, cet agencement, suite à la procédure d'extraction qui constitue la première étape d'une fouille de textes, est perdu. Et quand de surcroît les données obtenues sont retraitées avant analyse...

Dans tous les cas, ce qui fait l'originalité des œuvres n'est pas copié : on extrait des œuvres en tant que telle une partie plus ou moins significative de leurs éléments constitutifs, non protégés. Il n'est donc pas possible aux ayants droits d'invoquer un préjudice au titre du droit de reproduction.

C. Les producteurs de bases de données sont protégés par un droit sui generis

Ce droit *sui generis* est une importation du droit de la concurrence et prend place à l'article L342-1 du Code de la propriété intellectuelle : il ne protège pas la base de données en tant qu'œuvre mais son producteur, au titre de l' « investissement substantiel » réalisé pour la constitution de sa base.

Cette protection n'interdit pas la pratique de l'extraction en tant que telle, mais en limite l'abus, dès lors qu'une partie substantielle de la base de données serait extraite (articles L342-2 et L342-3 du Code de la propriété intellectuelle). Ce qui est indubitablement le cas dans les opérations de TDM.

Ces notions peuvent introduire à une analyse de la chaîne de création de valeur, et les éditeurs en tirent argument pour estimer que leur investissement substantiel, dès lors qu'il est exploité par des opérations de TDM, justifie à tout le moins une contrepartie (que viendrait fixer un accord contractuel). Cet argumentaire est en fait :

- inadapté : l'analyse de la chaîne de création de valeur entraîne inévitablement un raisonnement de type *regressus ad infinitum* : pour créer leurs produits, les éditeurs académiques ont eu besoin des universités pour former des chercheurs, des informaticiens, les travailleurs intellectuels des maisons d'édition, et ainsi de suite. Ce type de raisonnement, qui peut sembler incongru en France, étonnerait beaucoup moins le philanthrope américain effectuant un don en faveur de l'université où il a fait ses études. La question à se poser en l'espèce n'est pas qui crée de la valeur ni comment elle est étagée, mais y a-t-il ou non préjudice causé par la pratique du TDM au producteur de base de données ;
- en l'espèce, inopérant : si l'on tient à analyser la chaîne de création de valeur, il convient de ne pas limiter son analyse à l'aval de la mise à disposition de la base de données, mais également à son amont. Or une telle analyse n'est pas favorable aux éditeurs : le secteur des publications académiques qui constituent par exemple les éléments d'une base de données comme ScienceDirect se caractérise par une captation souvent jugée abusive de la création de valeur par les éditeurs. Ainsi, un chercheur qui publie un article abandonne généralement tous ses droits à l'éditeur, sans aucune contrepartie financière (à certaines exceptions près : édition juridique par exemple). En échange, son article bénéficie de la renommée de la revue dans laquelle il est publié. Mais cette renommée est elle-même indexée sur le prestige du comité de lecture de la revue, constitué de chercheurs exerçant eux-même cette activité à titre gracieux. En bout de chaîne, les

bibliothèques universitaires achètent pour des sommes importantes des licences d'accès aux articles produits. Entre le producteur et l'acheteur, l'éditeur a en effet apporté une valeur ajoutée, mais en a bien plus sûrement capté la plus grande part : Elsevier effectue par exemple un bénéfice annuel de 40 %. Et ce pendant que la puissance publique paie plusieurs fois pour un contenu dont la valeur ajoutée du fait de l'éditeur n'est pas si importante : elle paie le chercheur pour qu'il produise, elle paie la bibliothèque pour qu'elle acquière la licence d'accès à l'article revu et publié, voire elle paie encore une fois s'il s'agit d'acquérir les droits à un accès pérenne aux données, comme dans le cadre des licences nationales du projet ISTEX. Aujourd'hui, les éditeurs voudraient générer encore un profit supplémentaire, sans rien apporter dans la chaîne de création de valeur, en commercialisant une licence pour TDM. Cela semble incontestablement abusif : le fait de mettre à disposition une base de données pour permettre d'en fouiller computationnellement le contenu ne coûte aucun investissement à l'éditeur. Quand comme Elsevier, l'éditeur entend de surcroît poser un droit de propriété intellectuelle, via une licence CC-BY-NC sur le résultat final de la fouille, l'on est non seulement au-delà de ce que permet la loi, mais aussi au-delà de ce que permet d'argumenter le raisonnement économique : dans un travail de TDM, toute la valeur ajoutée est créée par l'équipe de recherche, qu'il s'agisse de la définition des hypothèses scientifiques et du corpus pertinent, du retraitement de ce dernier (probablement l'opération la plus lourde : par exemple, dans le cadre du Web de données, travail d'alignement des données), de la conception et de la réalisation des algorithmes de fouille, de l'analyse des résultats.

4. Quelle réponse juridique ?

Si l'analyse économique de la chaîne de création de valeur ne saurait justifier le paiement de licences pour la pratique du TDM, au titre du retour sur investissement des éditeurs, est-il du moins possible de considérer que le paiement de telles licences viendrait justement dédommager, sur le mode de la compensation, le producteur de base de données d'un préjudice, réel ou éventuel, dont il serait victime ?

Le problème est qu'on voit mal de quel préjudice il serait question : le respect des licences qui permettent d'accéder aujourd'hui aux contenus sous droit des éditeurs repose sur un dispositif technique éprouvé et sûr (recours à un authentifiant + mot de passe individuel pour chaque chercheur). On ne voit pas que ce qui a fonctionné pour ce type de licence serait insuffisant à garantir la sécurité des contenus sous droits dans un contexte de TDM. Et ce d'autant que les pratiques de TDM n'ont jamais pour finalité :

- l'exposition à tout un chacun des données extraites des bases sous droits : tout au plus, le temps du travail, la base est-elle partagée sur des serveurs sécurisés entre les différents membres d'une équipe de recherche, et accessible seulement sur authentifiant + mot de passe individuel. **Le risque de contrefaçon est donc quasiment nul** ;
- ni même l'extraction, même substantielle, de la base de données : cette extraction n'est qu'une **finalité accessoire** du processus de TDM : la finalité principale reste l'extraction non-substantielle des éléments pertinents dans le cadre du travail de recherche conduit. **L'on a donc bien contravention aux dispositions du droit *sui generis* des bases de données, mais du fait des seules manipulations techniques impliquées par le TDM**. Il s'agit donc en somme d'un cas de figure proche de celui prévu à l'article 5.1 de la directive communautaire 2001/29/CE sur les actes de reproduction transitoires ou accessoires. Et comme le soulignait Michel Vivant devant la Commission Lescure, au sujet des actes de reproduction (mais l'on peut transposer le raisonnement aux opérations d'extraction impliquées techniquement et accessoirement par le TDM) : « 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 termes de régulation sociale ? ».

Enfin et surtout, les pratiques liées au TDM répondent au test en 3 étapes de la Convention de Berne, ainsi que le soulignent les travaux effectués dans le cadre du projet de loi visant à réformer le *copyright* au Royaume-Uni : à ce titre, les pratiques liées au TDM n'ont pas à ouvrir droit à compensation.

Les fondements économiques ou juridiques d'une rétribution des éditeurs au titre des activités de TDM des chercheurs apparaissent donc bien faibles. On verrait du reste mal comment mettre en place une telle rétribution :

- le modèle de la licence, on l'a vu, est profondément inadapté (notamment quand un projet de TDM implique de croiser plusieurs bases et/ou le Web) ;
- le modèle de la gestion collective obligatoire semble profondément impraticable en l'espèce : les masses de contenus concernés impliqueraient un travail de recherche et de reversement des droits considérable, générant des frais de gestion sans commune mesure avec les sommes en jeu et les actes à l'origine d'un droit à compensation, sans compter que dans bien des cas, par exemple pour les archives du Web, il serait sans nul doute impossible d'identifier les ayants droits, d'où le risque certain de générer d'importants montants de sommes irrépartissables.

L'exception sans compensation apparaît donc comme la seule voie praticable pour régler juridiquement le problème soulevé par le TDM au regard du droit *sui generis* des bases de données.

C'est du reste la voie choisie tout récemment par la Grande-Bretagne, qui semble faire de cette exception une extension à son exception au titre de l'enseignement et de la recherche, ce qui évite d'allonger la liste limitative des exceptions prévues par la directive communautaire 2001/29/CE. Cette voie britannique serait tout aussi praticable en France (du reste, l'article L342-3 du Code de la propriété intellectuelle articule déjà le droit *sui generis* des bases de données et la timide exception française au titre de l'enseignement et de la recherche). Pour offrir à chacun les meilleures garantie de régulation sociale, cette exception devrait :

- être limitée aux usages non-commerciaux, ou n'entraînant pas de préjudice commercial direct ou indirect aux éditeurs de bases de données (en contrepartie de quoi, il n'y a pas à prévoir de compensation financière au titre du TDM) ;
- être limitée au domaine académique, afin de préserver les équilibres économiques de la presse professionnelle vis-à-vis des grands moteurs de recherche du Web (l'articulation de l'exception TDM avec l'exception enseignement-recherche garantirait ce point) ;
- être étendue aux archives du Web (actuellement collectées depuis 2006 dans le cadre du Dépôt légal mais dont la consultation est limitée à l'emprise de la Bibliothèque nationale de France et des bibliothèques du dépôt légal imprimeur - BDLI : ces ressources gagneraient à être mises à disposition des établissements d'enseignement supérieur et de recherche) et aux bases appartenant au domaine public (par exemple Gallica) ou pour lesquelles les droits ont déjà fait l'objet d'un accord contractuel (par exemple Persée, ou les réservoirs d'articles en *open access*) ;
- ne pas voir son exercice pouvoir être gêné ou empêché par des mesures de protection techniques (DRM, API) ou autres (limitation de durée d'accès, de volume extractible, de longueur des citations, obligation d'autorisation ou d'information préalable, etc.) : les mesures techniques doivent pouvoir être légalement contournées et les clauses limitatives réputées nulles et non-écrites

L'adoption rapide d'une telle exception en France est cruciale pour la compétitivité de notre recherche : la pratique du TDM est déjà admise aux USA (jurisprudence HathiTrust), gravée dans la loi en Irlande et bientôt en Grande-Bretagne. Comme le montrent différentes études[6], les bénéfices pour l'ensemble de la société, qu'il s'agisse du secteur public ou commercial, sont très nettement supérieurs au peu probable préjudice que pourraient encourir les titulaires de droits du fait des usages attachés au TDM.

Si l'on considère souvent que le droit de la propriété intellectuelle est une incitation à créer, il convient de prendre garde qu'il ne devienne pas un obstacle pour innover.

Grégory COLCANAP	Christophe PERALES
Coordonnateur de COUPERIN.ORG	Président de l'ADBU

6 http://www.ipo.gov.uk/consult-ia-bis0312.pdf et http://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining