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

Consultations du Conseil national du numérique : contributions de l'ADBU

L'ADBU est l'association professionnelle des cadres exerçant dans les services documentaires des établissements d'enseignement supérieur et de recherche.

1. Contribution relative à la fouille de contenus (*text and data mining*)

Qu'est-ce que la fouille de contenus (*text and data mining* - TDM) ?

Confrontés à l'inflation croissante, depuis la fin de la deuxième guerre mondiale, de l'information scientifique et technique[1] (IST), les chercheurs pratiquent depuis longtemps une lecture d'écrémage, réservant la pratique de la lecture intégrale au petit nombre d'articles ou d'ouvrages qu'ils identifient comme essentiels pour leur objet de recherche.
Avec l'apparition du Web et des nouveaux modes de communication scientifique qui en sont issus[2], l'inflation documentaire est telle qu'elle excède les capacités de veille même des équipes les mieux dotées.

Face à ce nouveau défi, la fouille de contenus (*text and data mining*) propose d'assister l'homme au moyen d'algorithmes de fouille, élaborés nécessairement à façon par les chercheurs eux-mêmes en fonction de leurs hypothèses de lecture et de veille. Le corpus concerné n'est rien moins d'autre que le Web, dans toute son étendue :

- Web visible, celui que parviennent à moissonner les moteurs de recherche ;
- Web invisible, qui échappe aux moteurs de recherche, pour diverses raisons : images fixes ou animées dépourvues d'indexation textuelle, objets dynamiques (en *FlashPlayer* par exemple) ou intégrant des éléments applicatifs (bases de données, jeux vidéos, etc.).

On imagine sans peine les possibilités ouvertes pour la recherche par ce mode de lecture algorithmique. Les États-Unis, la Grande-Bretagne, l'Irlande et le Japon ne s'y sont pas trompés, qui bénéficient aujourd'hui d'une législation autorisant la pratique de la fouille de contenus (*text and data mining*). Dans la course mondiale à l'innovation qui anime aujourd'hui les pays développés, sur fond de massification des données de tous types, l'accès régulé à cette technologie est désormais crucial : faute de pouvoir pratiquer le TDM en France, un nombre croissant de nos labora-

[1] Information scientifique et technique : information produite par et pour les chercheurs.

[2] Sites de chercheurs ou dédiés à une équipe / un projet de recherche, blogs, *microblogging*, réseaux sociaux spécialisés, publication en libre accès (*open access*) d'articles, de thèses, d'actes de colloques, de jeux de données issues de mesures, d'expériences, etc.

toires réalisent d'ores et déjà leurs opérations de fouille de contenus (*text and data mining*) à l'étranger, avec tous les risques associés à cette externalisation pour la compétitivité de notre recherche.

Indubitablement, une part importante du Web visible et invisible obéit à la réglementation relative au droit d'auteur (chartes graphiques de sites web, oeuvres de l'esprit, bases de données produites par les grands acteurs commerciaux de l'IST, etc.). Mais il est important de souligner que la fouille de contenus (*text and data mining*) n'a pas pour objectif la dissémination indue de ces contenus sous droits ou leur exploitation commerciale : c'est parce qu'en tant que lecture computationnelle, elle implique techniquement la création d'une copie du corpus à fouiller que la fouille de contenus (*text and data mining*) soulève un problème juridique. Une solution sécurisant les légitimes intérêts des divers ayants droit doit donc être trouvée. Il n'est pas anodin de souligner que parmi eux, se trouvent pour une bonne part précisément les chercheurs qui souhaitent accéder à cette technologie, et qui sont, eux aussi, des auteurs, au sens du Code de la propriété intellectuelle.

Pourquoi les solutions proposées par les grands acteurs de l'édition scientifique et technique sont-elles inadaptées ?

Aujourd'hui, seuls les deux principaux acteurs de l'édition en IST, Elsevier et Springer, proposent des solutions visant à répondre au besoin de TDM qu'ils ont eux-mêmes identifiés chez leurs clients (qui, encore une fois, sont aussi leurs auteurs) :

- Elsevier autorise ainsi contractuellement les chercheurs à télécharger, via une API, 10 000 articles par semaine issus de sa base de données ScienceDirect. Ce dispositif est profondément inadapté aux besoins de la science. La recherche chemine en effet par hypothèses et tests, essais et erreurs. Imaginons qu'une hypothèse de fouille de contenus 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.
En outre, le dispositif proposé contractuellement par Elsevier impose aux chercheurs de publier sous licence CC-BY-NC les extraits retenus comme pertinents à l'issue des opérations de TDM, en limitant la longueur de ces extraits à 350 mots. Or, la citation de travaux tiers est à la base même de tout travail académique, et l'exception de citation qui permet entre autres à la science d'exister n'impose pas de limitation aussi étroite que celle prévue par Elsevier : selon les termes de la directive 2001/29/CE, la longueur admise pour une citation doit être appréciée « dans la mesure justifiée par le but poursuivi ». Si dans les travaux académiques actuels, notamment en sciences humaines, les citations devaient se limiter à 350 mots, la plus grande part des publications académiques deviendrait tout simplement impossible.

- quant à Springer, qui vient d'acquérir le groupe Nature, il impose aujourd'hui que chaque projet de TDM portant sur ses contenus soit décrit et enregistré via un formulaire en ligne, se réservant par la suite le droit de décider si la demande lui semble ou non fondée. À rebours de tout principe décideur-payeur, ce système constitue à l'évidence une ingérence inacceptable du point de vue de l'indépendance de la recherche.

Rappelons enfin que dans les deux cas, Elsevier comme Springer, il s'agit de fouiller des corpus dont le droit d'accès, voire les contenus (dans le cadre du programme Investissements d'avenir ISTEX), ont été légalement acquis par les institutions académiques.

Pourquoi la voie contractuelle n'est-elle pas viable pour réguler la pratique du TDM ?

1. Les objets de recherche sont aujourd'hui de plus en plus interdisciplinaires (que l'on songe simplement à une question comme celle du développement durable) et les modalités du commerce entre savants excèdent largement le cadre de la publication dans des revues académiques commerciales.
C'est dire que les besoins en TDM concernent tous les corpus possibles : en un mot, l'ensemble du Web.

Or, aucun acteur commercial de l'IST, fût-ce Elsevier ou Springer, ne pourra jamais être en capacité de mettre en toute légalité, par la voie contractuelle, à disposition des chercheurs, l'ensemble du Web visible et invisible : pour s'en tenir aux seuls contenus éditoriaux en IST, cela supposerait la création d'une plateforme mondiale commune à tous les acteurs, dont on voit mal dans quels délais elle pourrait être mise en oeuvre, et surtout, selon quelles modalités, à la fois satisfaisantes pour les chercheurs et sécurisées pour les ayants droit. Mais par ailleurs, l'on n'aurait là qu'une infime partie du Web : en sciences humaines et sociales, c'est le Web visible qui intéresse bien davantage les chercheurs. Comment une offre contractuelle pourrait-elle émerger pour répondre à ce besoin ? Le nombre d'acteurs concernés est considérable, et progresse chaque jour de manière exponentielle. La réponse classique dans ce type de situation, celle de la gestion collective, obligatoire ou volontaire, s'avérerait ici inopérante : le Web étant un média mondialisé, il y faudrait une société de perception et de répartition des droits (SPRD) compétente à l'échelle planétaire, et organisée en conséquence, avec les coûts de gestion associés, incommensurables, et très probablement un nombre inacceptable de sommes irrépartissables.

Quelle que soit l'hypothèse envisagée, la voie contractuelle est inadaptée, car inapte à répondre aux défis posés par le TDM tout en sécurisant les intérêts des ayants droit.

2. En outre, sur un plan juridique, la voie contractuelle apparaît impraticable. En effet, le droit communautaire considère comme légal le monopole constitué par le droit de la propriété intellectuelle : pour que la création soit encouragée, il convient très justement que les oeuvres soient protégées, et que leurs créateurs puissent tirer un juste revenu du fruit de leur travail. Toutefois, si sur la base de ce monopole légal, les ayants droit entendent exploiter un marché dérivé dont la libre concurrence se verrait faussée précisément par leur situation de monopole légal, le droit communautaire estime qu'il y a alors, sur ledit marché dérivé, une situation d'abus de position dominante avéré, ou, pour utiliser la terminologie communautaire, un « pouvoir de marché » illégal.
C'est précisément ce qui se passerait si les ayants droit des contenus à fouiller entendaient contrôler seuls les activités de TDM portant sur leurs contenus : sauf à autoriser systématiquement toute demande de fouille (et toute SSII qui se serait fait, en qualité d'intermédiaire, une spécialité des opérations de TDM), ils useraient de leur monopole légal pour contrôler abusivement le marché dérivé du TDM.

3. Enfin, privilégier la voie contractuelle pour le TDM reviendrait à consolider les positions des grands « plateformeurs » du Web, dont on entend pourtant actuellement, notamment au nom du droit d'auteur, réguler l'activité. En effet, suite à divers études et rapports, un consensus s'est dégagé ces derniers mois en France pour interroger la pertinence que les « plateformeurs » bénéficient du régime particulièrement favorable accordé aux hébergeurs par la législation européenne (directive 2000/31/CE) :

- rapport du Conseil d'État sur le numérique et les droits fondamentaux ;
- rapport du Conseil national du numérique (CNNum) sur la neutralité des plateformes ;
- dans le cadre des travaux du CSPLA, rapport du Professeur Sirinelli sur l'avenir de la directive 2001/29/CE.

Notamment, suite aux analyses figurant dans le rapport du CNNum, il est désormais admis que l'activité des grandes plateformes du Web obéit à un modèle économique singulier, totalement différent de celui des hébergeurs proprement dits, consistant en l'exploitation d'un marché triface :

- sur une première face du marché, une audience est générée, généralement par la mise à disposition gracieuse d'un service (un moteur de recherche, par exemple, dans le cas de Google) ;
- sur une deuxième face du marché, cette audience est exploitée auprès d'annonceurs publicitaires ;
- sur une troisième face du marché, cette audience constitue une clientèle pour des services à valeur ajoutée (c'est le sens du rachat par Google de toute une série de *start-ups*).

Qui aujourd'hui, en toute légalité, dispose de l'ensemble des articles publiés en libre accès (*open access*) sur le Web ? Qui, par là même, est aujourd'hui le seul opérateur à pouvoir fournir une prestation de TDM sur ces contenus légalement collectés et pour lesquels les auteurs ont abandonné tous leurs droits ? Google Scholar.

Faut-il refuser aux chercheurs la possibilité de fouiller les contenus qu'ils ont eux-mêmes produits ? Faut-il réserver cette possibilité aux plateformes du Web, visible ou invisible, qui procèdent déjà à une captation abusive de la valeur produite par des ayants droit non rémunérés, financés par l'argent public ?

Pourquoi les exceptions au droit d'auteur déjà existantes sont inadaptées pour répondre au défi du TDM ?

Le TDM n'a pas pour objectif la dissémination indue de contenus sous droits, mais leur simple lecture algorithmique. Pour ce faire, une copie des corpus à fouiller est nécessaire, mais cette copie n'est pas la fin poursuivie par les opérations de TDM, non plus que l'exploitation commerciale de la copie.
C'est pourquoi l'hypothèse a pu être émise d'une adaptation de l'exception 5.1 de la directive 2001/29/CE pour régler juridiquement la question du TDM.

Cette option n'est pas adaptée : l'exception en question porte en effet seulement sur des copies provisoires. Ce qui emporterait que la copie d'un corpus soit systématiquement détruite une fois ce corpus fouillé. Et qu'une nouvelle copie soit générée à chaque nouvelle demande de fouille d'un corpus. Il y a fort à parier que les ayants droit des corpus concernés se verraient ainsi contraints d'engager des moyens très conséquents afin de faire face aux demandes de reproductions, et à la gestion de leur cycle de vie.
En outre, il est essentiel pour la validité et la nécessaire reproductibilité du processus de la recherche de garantir, dans la durée, l'accessibilité des équipes aux données qualifiées sur lesquelles elles ont fondé leurs hypothèses scientifiques.

En fait, si, selon les termes de la directive 2001/29/CE, pour le chercheur, la reproduction d'un corpus ne vise qu'un usage transitoire (le temps de la fouille) et accessoire (copier le corpus ne constitue jamais l'objectif de la recherche), dans l'intérêt des ayants droit, il semble raisonnable de concevoir un dispositif où la copie ne serait pas, comme l'impose l'exception 5.1 de la directive, provisoire. Ce qui implique logiquement le recours à un tiers de confiance entre usagers et ayants droits. Et partant, nécessairement, la création d'une exception spécifique au droit d'auteur, en faveur du TDM.

Que propose l'ADBU ?

1. L'instauration dans le cadre du projet de loi sur le numérique d'une nouvelle exception au droit d'auteur, autorisant la pratique du TDM pour les seuls acteurs de la recherche publique, et sans compensation financière.
2. La désignation par la loi d'un tiers de confiance, chargé d'héberger tous les corpus du Web visible et invisible, aux seules fins d'en permettre la lecture algorithmique par les technologies du TDM : cette immense base de contenus constituerait une base-maître qui éviterait aux ayants droit de gérer un nombre excessif de demandes de mises à disposition de corpus. Cette base-maître ne serait accessible et manipulable que par le tiers de confiance. Plusieurs acteurs publics sont aujourd'hui capables, en France, de jouer ce rôle.
3. À chaque demande émanant d'un acteur de la recherche publique, un « bac à sable » serait constitué à façon en copiant depuis la base-maître tous les corpus intéressant l'équipe de recherche. L'accès au « bac à sable » serait contrôlé selon les mêmes modalités que celles en usage dans le monde académique pour l'accès aux bases de données des éditeurs commerciaux (reconnaissance des adresses IP et *reverse proxy*) : ces modalités ont donné pleine satisfaction à toutes les parties depuis qu'elles existent, et renforcé la confiance mutuelle entre acteurs académiques et ayants droit. Elles permettent notamment de limiter les accès aux seuls acteurs publics, même dans le cas de projets de recherche partenariale, impliquant des entreprises privées.
4. Une fois conduites les opérations de TDM nécessaires au projet de recherche, la copie constituant le « bac à sable » serait détruite. Pour des raisons de validité et de reproductibilité des processus de recherche, ne seraient conservées que les occurrences pertinentes mises à jour par les opérations de fouille.

L'on répondrait ainsi pleinement aux besoins des chercheurs tout en assurant la protection légitime des intérêts des ayants droit. Et sans nullement remettre en question ou menacer les fondements du droit d'auteur, l'on garantirait la liberté de lire, sous toutes ses formes.

2. Contribution relative à la promotion du libre accès (*open access*)

Qu'entend-on par libre accès (*open access*) ?

Il existe aujourd'hui trois voies de mise en libre accès (*open access*) des productions scientifiques :
- la voie dite verte (*green open access*), qui consiste à déposer sur un site dédié (dit *open archive*), dès lors que l'éditeur l'y autorise, une contribution scientifique dont la version finale est par ailleurs publiée dans le circuit éditorial classique : la version en *open access* peut être la version avant évaluation par les pairs (*peer-reviewing*) ou celle revue par les pairs, très rarement celle telle que publiée par l'éditeur ;
- la voie dite dorée (*gold open access*) : initialement destinée à créer des revues en libre accès financées à prix coûtant sur d'autres modèles de financement que l'abonnement, la voie dorée a été très largement adoptée par les géants du secteur de l'édition en IST[3] dans le modèle de financement particulier dit auteur-payeur. Dans ce modèle, en échange du paiement d'APC (*article processing charges*) par son institution, le chercheur publie dans les revues des géants du secteur de l'édition en IST, lesquels mettent ensuite en libre accès les articles ou actes de congrès dont la publication a été ainsi financée. Cette voie dorée, qui connaît de nombreuses dérives, est aujourd'hui largement controversée (APC excessifs, ou non exclusifs du paiement par les bibliothèques d'un abonnement pour donner accès à des publications sur APC, bonus accordé aux institutions ou pays capables de débourser les crédits les plus importants pour publication, etc.) ;
- la voie dite platine (*platinum open access*) ou *freemium*, dans laquelle les publications sont mises en ligne, dès parution, de manière ouverte (par exemple, en *streaming*) et où seuls des services à valeur ajoutée sont payants (par exemple, version .pdf téléchargeable).

Les propositions qui suivent ne concernent que la voie dite verte (« libre-accès » vaudra systématiquement pour « *green open access* » dans les lignes qui suivent).

[3] Information scientifique et technique : information produite par et pour les chercheurs.

Pourquoi soutenir le libre accès (*green open access*) ?

Le secteur de l'édition en IST présente plusieurs particularités tout à fait spécifiques :
- le lectorat des publications scientifiques en constitue également l'auctorat ;
- hormis en sciences juridiques[4], les auteurs n'y sont pas rétribués pour leur production : ils cèdent l'ensemble de leurs droits patrimoniaux aux éditeurs ;
- le travail éditorial de relecture et d'amélioration du manuscrit initial est effectué gratuitement par les chercheurs eux-mêmes (*peer-reviewing*) ;
- les frais de publication sont souvent supportés pour partie par l'auteur ou son institution (insertion dans l'article d'une image scientifique, d'un schéma, d'un tableau, etc.), quand ce n'est pas entièrement via les APC.
L'on ne saurait mieux illustrer que le rapport des forces en présence dans une relation contractuelle n'est pas toujours équilibré : dans le cas présent, face à des multinationales de l'édition, les auteurs (les chercheurs), pressés par l'impératif du *publish or perish* sur lequel repose toute leur évolution de carrière et la gestion de leur capital symbolique, ne sont pas en position de négocier des termes satisfaisants pour la cession de leurs droits patrimoniaux.

Or l'essentiel du financement de la recherche est très essentiellement le fait, en France comme partout ailleurs dans le monde, de la puissance publique : cette dernière paie pour que des articles scientifiques soient produits.
Par ailleurs, via le financement dont bénéficient les bibliothèques universitaires, elle paie également environ 80 M€ par an en France afin que les chercheurs de l'Hexagone aient accès aux publications produites dans le monde entier… y compris celles dont ils sont les auteurs.
L'on assiste ainsi à une captation abusive de la valeur produite : les grands éditeurs commerciaux en IST contribuent peu dans la chaîne de création de valeur des publications scientifiques ; en revanche, ils en tirent un taux de retour sur investissement sans commune mesure avec leur valeur ajoutée, dégageant annuellement des bénéfices à deux chiffres (constituant de 15 à 30% de leur chiffre d'affaires…).
Face aux acteurs commerciaux, la puissance publique, qui a financé, et la recherche (salaires, infrastructures, fonctionnement), et l'accès des publiants aux publications (dépenses annuelles des bibliothèques académiques) ne peut librement mettre à la disposition du contribuable les produits de la recherche financés via l'impôt commun.

Le mouvement du libre accès (*open access*), né durant les années 90 en réaction à cette situation, ne vise rien d'autre que la correction de ce déséquilibre. Il est sans impact sur la réglementation relative au droit de la propriété intellectuelle, et un nombre croissant d'institutions (Union européenne notamment, à travers le programme Horizon 2020) ou d'états (Grande-Bretagne, Allemagne, Italie, États-Unis, etc.) ont adopté des dispositions législatives ou réglementaires pour favoriser les publications en libre accès (*open access*).

Que propose l'ADBU ?

- À la faveur du projet de loi sur le numérique, un rééquilibrage des forces en présence dans la relation contractuelle entre auteur (chercheur) et éditeur, par l'instauration d'une obligation pour ce dernier, dès lors que l'essentiel d'un programme de recherche est adossé à des financements publics, de libérer systématiquement les droits des publications scientifiques dans un délai contraint (6 mois maximum pour les STM, peut-être 12 pour les SHS[5], semblent des délais raisonnables, admis dans la plupart des pays comparables qui nous ont devancés sur cette voie, sous le terme d' « embargo »).
- Afin de préserver certains équilibres délicats et de garantir à l'éditeur son retour sur investissement, cette contrainte ne porterait que sur les articles et les communications dans des congrès (l'économie de l'édition d'ouvrages scientifiques, notamment en SHS, étant fragile, et nécessitant un temps long), et précisément, sur le dernier manuscrit produit par l'auteur et revu par les pairs avant mise en forme par l'éditeur : ainsi, ce dernier demeurerait le seul détenteur des droits sur la version de l'article intégrant sa plus-value propre, qu'il resterait pertinent de commercialiser même après la période d'embargo ; mais les résultats de la science produite sur fonds public seraient mis dans des délais raisonnables à la disposition de l'ensemble de la communauté de recherche et de son principal financeur, le contribuable.
- Un tel système serait en outre sans impact aucun sur le droit de la propriété intellectuelle, l'auteur en particulier conservant toutes ses prérogatives, notamment en termes de droits moraux : obliger

[4] Le caractère national du droit positif de chaque état emporte en effet une concentration du marché des lecteurs sur le pays concerné, situation plus favorable aux auteurs, et, symétriquement, plus défavorables aux multinationales de l'édition.

[5] Cette durée de 12 mois reste à préciser, vu l'importance de l'édition en SHS en France et la fragilité de ce secteur : à cette fin, une étude a été commandée auprès de l'IPP (Institut des politiques publiques) par le Ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche, qui devrait être remise sous peu.

l’éditeur à libérer la version pré-éditoriale des articles scientifiques au bout d’une période d’embargo n’oblige en effet en rien l’auteur à publier cette version en libre accès (*open access*) ; elle lui en ouvre seulement la possibilité.

L'on procéderait ainsi à un rééquilibrage entre intérêts publics et intérêts privés sur le marché très particulier des publications en IST, sans remettre en cause le principe de l'exclusivité des droits de la propriété intellectuelle.