CNNum
Conseil National du Numérique

PROJET DE LOI
POUR UNE RÉPUBLIQUE NUMÉRIQUE

LA FOUILLE DE TEXTES ET DE DONNEES

DE QUOI PARLE-T-ON ?

La fouille de textes et de données (ou *text and data mining - TDM*) désigne un ensemble de **traitements automatisés consistant à extraire des connaissances dans un ensemble de contenus numériques**, qui peuvent inclure des textes, des données, des sons, des images ou d'autres éléments, ou une combinaison de ceux-ci. Elle permet d'analyser parallèlement de vastes quantités de données selon un critère de nouveauté ou de similarité, et ainsi de dégager des conclusions difficiles à appréhender par la simple lecture cursive.

Les grands éditeurs, qui détiennent la majeure partie des publications scientifiques, peuvent aujourd'hui proscrire la fouille de textes et de données aux chercheurs – notamment la copie provisoire, techniquement nécessaire afin de la réaliser. **Ces derniers se sont pourtant acquittés des droits d'accès aux publications scientifiques qui constituent les bases de données qu'ils souhaitent traiter**. Cette interdiction s'appuie notamment sur le droit *sui generis* des bases de données.

Un exemple

Le projet Text2genome a permis de cartographier le génome humain par la compilation automatique de trois millions de publications.

POURQUOI EST-CE IMPORTANT ?

POUR LA RECHERCHE ET L'INNOVATION

Les activités de *text and data mining* sont porteuses de nombreux potentiels pour la découverte scientifique et le développement de nouvelles connaissances. Elles doivent permettre au monde de la recherche de bénéficier des progrès rendus possibles par l'analyse des mégadonnées (*big data)* en autorisant les chercheurs à opérer des fouilles automatisées dans l'immensité des documents scientifiques disponibles.

POUR LES ENTREPRISES

De nombreux pays ont d'ores et déjà mis en place cette exception. Au-delà des apports intellectuels, des bienfaits sociaux ou des progrès en matière de santé publique, il s'agit d'un **enjeu crucial pour la compétitivité de la recherche en France, moteur de l'innovation et de la transformation de l'ensemble de notre tissu économique, créatrice de richesse et d'emplois**.

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, 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.

Grégory COLCANAP, Coordonnateur de COUPERIN.ORG et **Christophe PERALES**, Président de l'ADBU (CSPLA - Mission relative au data mining (exploration de données) : l'analyse de Couperin et de l'ADBU, 2014)

IDÉES REÇUES & CONTRE-ARGUMENTS

"Le risque de diffusion numérique de contrefaçons est trop grand"

Certains ayants droit s'inquiètent de voir le nombre de contrefaçons réalisées à partir des copies nécessaires à la fouille automatisée de textes augmenter considérablement avec l'introduction d'une exception. Ce risque, basé sur la présomption que les chercheurs seraient enclins à commettre des actes aujourd'hui déjà fortement réprimés par la loi, peut toutefois être aisément contourné par **la mise en place d'une plateforme jugée légitime par les chercheurs et les éditeurs, qui jouerait le rôle d'un tiers de confiance.** La Bibliothèque nationale de France (BnF) s'est d'ores et déjà portée candidate pour jouer un tel rôle.

"Les solutions contractuelles proposées par les éditeurs sont satisfaisantes"

Les systèmes de licences *ad hoc* sont largement décriées par les chercheurs. En effet, l'expérience du projet ISTEX a démontré que l'analyse de base de données en provenance d'éditeurs multiples, est rendue impossible par la superposition des licences spécifiques d'exploitation à chaque éditeur**. Outre la lenteur et les limites du nombre de textes à fouiller, la voie contractuelle soulève des questions quant à la légitimité des procédures encadrant la recherche publique et de l'indépendance de la science.** De telles licences imposent discrètement l'idée que les informations pourraient être protégées, et que leur exploration pourrait nécessiter la création de nouvelles rémunérations, alors même que le droit d'auteur, qui protège la forme d'expression et non les idées, permet aujourd'hui de lire et de réutiliser des informations ou données incluses dans un texte sur lequel on a obtenu un droit d'accès. Le CNNum estime qu'il n'y a pas de raison légitime à restreindre ce droit dans le cadre d'un traitement automatisé.

"Cela ne relève pas du cadre législatif national"

Le Royaume Uni a d'ores et déjà adopté une exception au droit d'auteur en 2014 en faveur du *text and data mining* à des fins de recherche, sur la base d'une extension de son exception au titre de l'enseignement et de la recherche. L'exception est ainsi en accord avec la liste limitative des exceptions prévues par la directive communautaire 2001/29/CE, et n'a pas entraînée de recours en manquement de la Commission européenne. Cette articulation avec l'exception pour la recherche est tout à fait envisageable en France.

CNNum
Conseil National du Numérique

Contact : Yann Bonnet, Secrétaire général - yann.bonnet@cnnumerique.fr

QUE CHANGER DANS LA LOI ?

Le CNNum recommande d'instaurer une exception au droit d'auteur autorisant la fouille de textes et de données, limitée aux usages non-commerciaux et au domaine académique, sur la base d'une extension de l'exception au droit d'auteur pour des fins de recherche. Elle pourrait être étendue aux archives du Web, aux bases appartenant au domaine public et les bases pour lesquelles les droits ont déjà fait l'objet d'un accord contractuel. L'exercice de cette exception ne devrait pouvoir être gênée par des mesures de protection techniques ou des clauses limitatives

Elle pourrait entraîner les modifications législatives suivantes.

I. L'article L. 1225 du code de la propriété intellectuelle est ainsi modifié :

Après le neuvième alinéa, il est inséré un alinéa ainsi rédigé :

« 10° Les copies ou reproductions numériques réalisées à partir d'une source licite, en vue de l'exploration de textes et de données pour les besoins de la recherche publique, à l'exclusion de toute finalité commerciale. Un décret fixe les conditions dans lesquelles l'exploration des textes et des données est mise en œuvre, ainsi que les modalités de conservation et communication des fichiers produits au terme des activités de recherche pour lesquelles elles ont été produites »

II. Après le cinquième alinéa de l'article L. 3423 du même code, il est inséré un alinéa ainsi rédigé :

« 5° Les copies ou reproductions numériques de la base réalisées par une personne qui y a licitement accès, en vue de fouilles de textes et de données dans un cadre de recherche, à l'exclusion de toute finalité commerciale. La conservation et la communication des copies techniques issues des traitements, au terme des activités de recherche pour lesquelles elles ont été produites, sont assurées par des organismes désignés par décret. »

Contact : Yann Bonnet, Secrétaire général - yann.bonnet@cnnumerique.fr