#TDM : Fouille de textes et de données dans le contexte de la loi pour une République numérique - Journée d'étude ADBU du 13/12/16

Catherine Muller

Les billets d'EnssibLab
27 mars 2017

Le Text & Data Mining : pour quoi faire ?

La fouille de textes et de données, désignée sous le terme anglo-saxon de TDM pour Text & Data Mining désigne toute technique d'analyse automatisée visant à analyser des textes et des données sous forme numérique afin d'en dégager des informations stratégiques pour la recherche telles que des constantes, des tendances et des corrélations1. Avec une croissance sans précédent du volume des textes et des données disponibles sur Internet et partout ailleurs - dont plus de 2,4 millions d'articles scientifiques publiés chaque année, il n'est évidemment pas possible pour les chercheurs d'analyser et d'exploiter cet ensemble de connaissances sans disposer de dispositifs logiciels de pointe adaptés à l'ampleur titanesque de l'entreprise.
Enjeu majeur de la recherche, porteur de nombreux potentiels pour la découverte scientifique et le développement des connaissances, le TDM peut permettre à la recherche de profiter des avancées en matière d'analyse des "big data", dont l'enjeu de compétitivité internationale ne fait aucun doute. Actuellement, les pratiques du TDM sont autorisées en Irlande, au Royaume-Uni, aux Etats-Unis et au Japon.

Quel régime juridique pour la fouille de textes en France ?

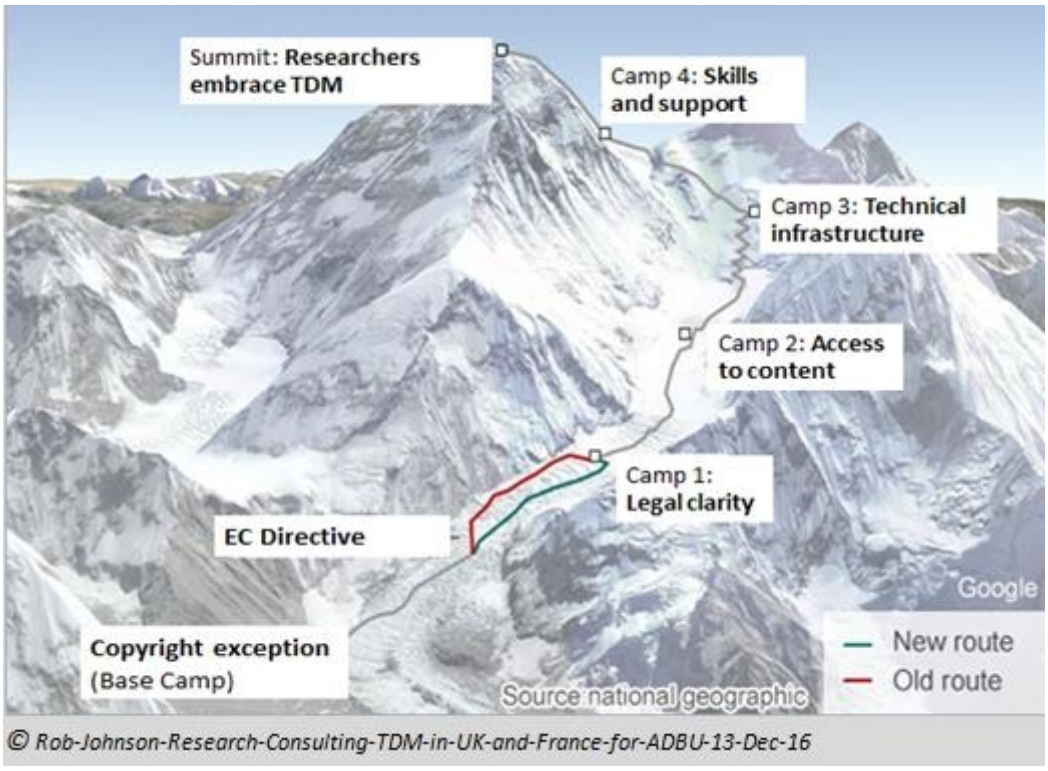
En France, les grands éditeurs qui détiennent la majeure partie des publications scientifiques, peuvent proscrire, par des solutions contractuelles, la fouille de textes et de données aux chercheurs, même si les abonnés disposent par ailleurs d'un accès légal à l'ensemble des publications scientifiques comprises dans les bases de données fouillées. Cette interdiction se réfère notamment au droit *sui generis* des bases de données. Cette pratique nécessitait donc de toute urgence la création d'une exception au droit d'auteur, à l'image du régime d'exception adopté en 2014 au Royaume-Uni sur la base d'une réinterprétation du droit pour la recherche.
C'est dans ce contexte que l'instauration d'une exception au droit d'auteur autorisant la fouille de textes et de données a été adoptée en France en 2016 par la loi pour une République numérique[2], dont les avancées en la matière ont repris en grandes parties les propositions soutenues durant la consultation publique par COUPERIN et l'ADBU en faveur du TDM et de la promotion du libre-accès.

Quels enjeux pour la recherche et les bibliothèques ?

L'ADBU, l'association des directeurs et personnels de direction des bibliothèques universitaires et de la documentation, a souhaité organiser à son tour, en écho à l'actualité législative, une journée d'étude sur la fouille de textes et de données dans l'enseignement supérieur et la recherche publique qui s'est tenue le 13 décembre dernier. À l'adresse des décideurs de l'ESR mais aussi des professionnels de l'IST, cette rencontre avait pour objectif de sensibiliser les communautés universitaires et de recherche aux enjeux du TDM pour l'avenir de la recherche scientifique.
L'analyse des enjeux s'appuyait sur une étude comparative inédite d'impact de la fouille de textes et de données sur la compétitivité de la recherche en France et en Grande Bretagne. Pilotée par l'ADBU et confiée à un cabinet de consulting anglais[3], l'intérêt premier de l'étude était de préconiser des recommandations à l'intention des acteurs de l'ESR. Les objectifs de la démarche adoptée visaient tout autant à mesurer l'impact économique considérable du TDM sur la recherche publique dont témoignait l'analyse comparée de cas français et britanniques, qu'à démontrer la pertinence de l'exception au droit d'auteur pour le TDM. En appui à ces conclusions, l'ISTEX et l'INRA pour la France et les Universités de Cambridge et de Manchester pour le Royaume-Uni, nous proposaient un retour d'expérience sur leurs propres pratiques en soulignant tout aussi bien les conditions de mises en œuvre du TDM que les freins auxquels ils étaient confrontés.

Les conclusions de l'étude comparative France/Royaume-Uni sur le TDM

L'état des lieux comparé sur les pratiques de TDM dans ces 2 pays engagé par le consultant Rob Johnson, directeur du cabinet Research Consulting, s'est appuyé sur l'analyse des enjeux pratiques, organisationnels et juridiques auxquels sont confrontés les acteurs du TDM dans l'enseignement supérieur. La démarche d'analyse comparative du cadre réglementaire en France, au Royaume-Uni et dans l'Union européenne couplée à des études de cas qui relatent les expériences des spécialistes du TDM dans ces aires géographiques, mais aussi aux États-Unis, a permis d'identifier les obstacles et les éléments favorables à l'usage de la fouille de textes et de données. Au chapitre des opportunités, Rob Johnson souligne que deux conditions au moins sont réunies pour favoriser le TDM dans la recherche universitaire, à commencer par la part d'investissement consacré en France à la R&D (Recherche et Développement) et l'innovation dans le secteur public et l'enseignement supérieur, qui est très élevée[4]. Par ailleurs, il souligne le seuil relativement bas de technicité demandé par l'outil de fouille, qui ne demande pas de savoir coder.



© Rob-Johnson-Research-Consulting-TDM-in-UK-and-France-for-ADBU-13-Dec-16

L'étude met à jour les différentes conditions et mesures nécessaires pour favoriser l'usage du TDM dans la recherche. Cinq stades de progression sont dénombrés, ainsi que l'illustre la métaphore de l'escalade dans la photo ci-contre. Organisée en 3 volets, l'argumentation s'attache à montrer pourquoi la fouille de données est cruciale pour la recherche universitaire, avant d'expliciter à la fois les freins et les leviers qui pourront les lever.

Fournir un cadre juridique clair et faciliter l'accès au contenu

Avec l'apparition du Web et des nouveaux modes de communication scientifique, l'IST est confrontée à une inflation documentaire inédite, avec une production de plus de 2,5 exaoctets (10^{18} octets) de données par jour, qui excède définitivement les capacités de veille des équipes de recherche les mieux dotées.

Face à ce nouveau défi, le TDM n'est rien de moins que la condition *sine qua non* de la recherche scientifique, de sa diffusion et de sa performance, sans laquelle les chercheurs européens ne

pourront pas rattraper leur retard sur le reste du monde. En effet, dorénavant, seule la fouille de contenus est en capacité 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 se confond avec le Web, dans toute son étendue, visible et invisible. Les possibilités ouvertes pour la recherche par ce mode de lecture algorithmique ont été bien comprises par certains pays leaders dans ce domaine, comme les États-Unis, la Grande-Bretagne, l'Irlande et le Japon, qui bénéficient depuis plusieurs années déjà d'une législation autorisant la pratique de la fouille de contenus. L'accès régulé à cette technologie constitue d'ores et déjà une ligne de fracture en matière de compétitivité de la recherche, qui explique le retard pris par certains pays - dont la France.

Tableau 1. Quel impact de la législation sur le droit d'auteur pour le TDM ?

	Le « *fair use* » des États-Unis	La proposition de la CE	L'exception française (Loi pour une République Numérique)	L'exception du Royaume-Uni
Quels usages sont couverts ?	Tout usage couvert par le *fair use*	La recherche scientifique	La recherche scientifique	La recherche
Une fin commerciale est-elle explicitement écartée ?	Non, sous réserve que le *fair use* soit respecté	Non, tant que la recherche émane d'un « institut de recherche »	Oui	Oui
Qui est autorisé à fouiller du contenu protégé ?	Quiconque, sous réserve que le *fair use* soit respecté	Les organismes de recherche à but non lucratif / les missions d'intérêt public	Quiconque œuvrant dans le cadre de la recherche publique	Quiconque
Les titulaires du droit d'auteur peuvent-ils limiter l'usage du TDM ?	Non, à moins d'un usage abusif (ex : s'il met en péril la viabilité commerciale du contenu fouillé)	Des mesures de garantie de la sécurité et de l'intégrité des réseaux et des bases de données	Non spécifié	Des mesures techniques de protection qui sont « raisonnables »
Quels types de textes et données peuvent-être fouillés ?	Tous, sous réserve que le *fair use* soit respecté	Les œuvres ou autres objets protégés	Tout type d'œuvres textuelles protégées, ainsi que les données incluses ou associées aux écrits scientifiques; toute BDD protégée contenant du texte et/ou des données incluses ou associées aux écrits scientifiques	Tout type d'œuvres

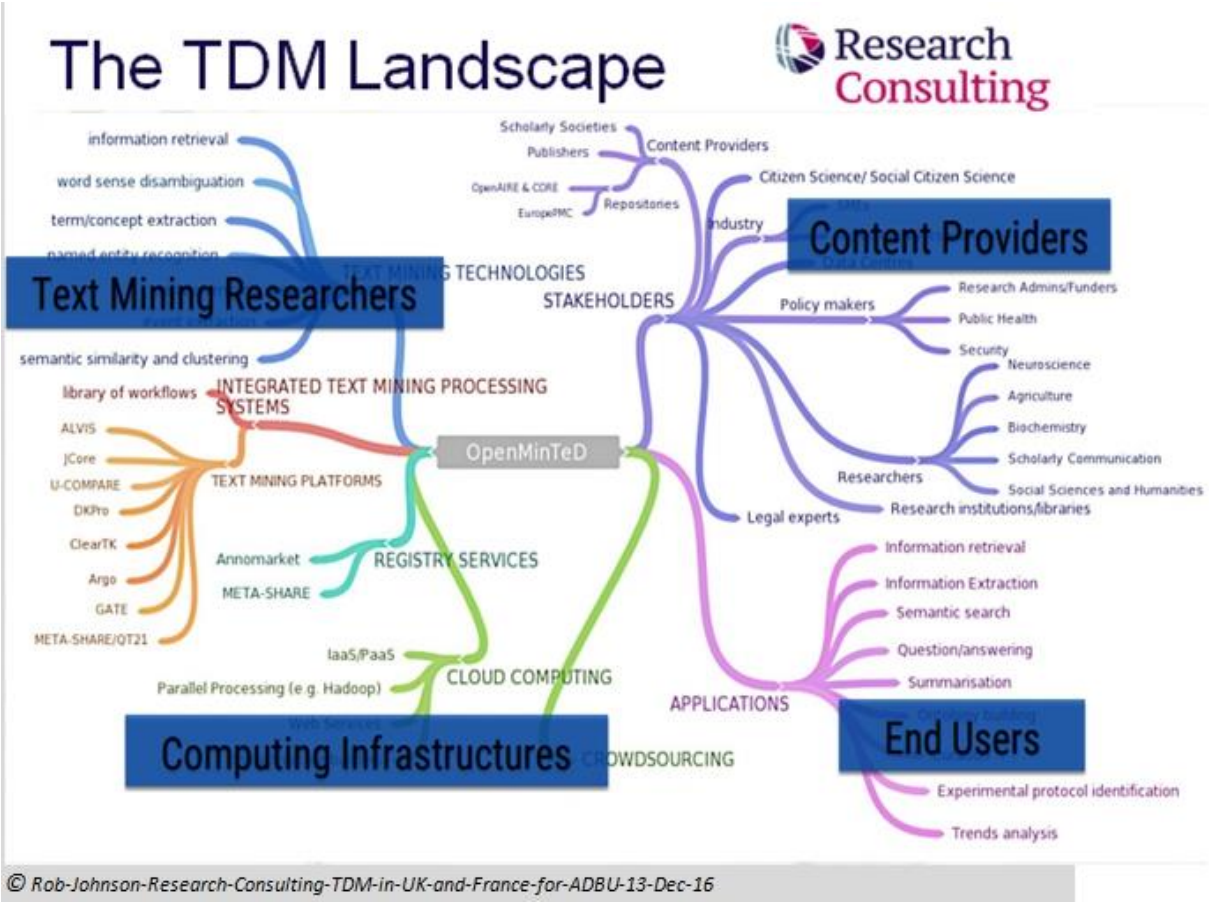
Préparé au nom de L'ADBU par www.research-consulting.com

Mais selon Rob Johnson, pour favoriser et généraliser la pratique du TDM, la directive européenne attendue en 2017 sur l'exception au droit d'auteur ne suffit pas. Certes, une part importante du Web est régie par la réglementation relative au droit d'auteur. Pourtant, il est important de souligner que la fouille de contenus n'a pas pour objectif la dissémination indue de ces contenus sous droits ou leur exploitation commerciale. Si la fouille de contenus soulève un problème juridique, c'est uniquement parce qu'en tant que lecture computationnelle, elle implique techniquement la création d'une copie du corpus à fouiller. Une solution sécurisant les légitimes intérêts des divers ayants droit doit donc être trouvée. De même, il faut garantir l'accès aux données. Trop souvent, les chercheurs interrogés déclarent rencontrer des difficultés pour obtenir l'accès au contenu publié, généralement en raison de dispositifs techniques de verrouillage, de restriction aux API ou bien encore d'hétérogénéité des formats de données entre les éditeurs qui compliquent l'agrégation de contenus.
En plus de clarifier le cadre législatif à l'échelle européenne avec des droits et des attributions non équivoques, il est indispensable de prévoir des clauses dans les contrats avec les éditeurs et les licences adoptées de façon à sécuriser l'accès aux données. À cette étape, le rôle de

sensibilisation des chercheurs et d'expertise des bibliothécaires aux différentes formules juridiques est essentiel, à l'instar de la fonction de chef d'orchestre et d'accompagnement qu'ils occupent depuis ces dernières années dans les dispositifs de gestion et d'ouverture des données de la recherche universitaire.

Développer l'infrastructure

L'amélioration des infrastructures et des outils constitue la 3ème étape clef pour contrer les nombreux projets qui se soldent par un échec ou une interruption faute d'infrastructure solide au départ ou d'initiatives isolées sans réel soutien. Ici, le faible intérêt commercial des éditeurs et fournisseurs de contenus à développer des solutions de TDM dans un contexte législatif d'exception, justifie que l'investissement soit porté par des fonds publics. Par exemple l'investissement de la France dans le projet ISTEX[5] a largement contribué à une avancée significative en matière d'accès au contenu, d'infrastructure et d'expertise.



L'amélioration des technologies utilisées pour compiler, uniformiser, interroger et préserver la matière issue du TDM, doublée d'un encouragement à un usage plus large d'ISTEX pour la fouille de contenus et du développement de services en ligne orientés usagers, accessibles et adaptés aux chercheurs dotés de compétences techniques limitées[6], restent une priorité. L'étude révèle deux initiatives heureuses en la matière. Conjointement avec d'autres partenaires à travers l'Europe, l'INRA, Institut National de la Recherche Agronomique, travaille au développement d'OpenMinTeD (Open Mining INfrastructure for TExt and Data), une e-infrastructure de fouille de textes orientée chercheurs en accès libre. Ce projet vise à promouvoir les outils de fouille de textes et de données, et à les rendre plus accessibles et interopérables à travers des registres appropriés et un niveau d'interopérabilité normalisé. Autre acteur très actif dans la promotion du TDM au Royaume-Uni, ContentMine, une entreprise à but non lucratif basée à Cambridge, qui a pour ambition de « libérer les données scientifiques des revues académiques » pour

permettre à quiconque d'effectuer des recherches à l'aide du TDM. L'entreprise en fait la promotion auprès des chercheurs qui se heurtent à un volume massif de contenus.

Inciter et monter en compétences : le rôle des bibliothèques

Pour achever et consolider l'édifice, l'étude pilotée par l'ADBU ne saurait trop rappeler le rôle d'ambassadeur structurant et actif des bibliothèques auprès des décideurs dans la diffusion des bonnes pratiques du TDM. Leur participation dans le processus constitue ainsi le dernier stade de consolidation. D'abord parce qu'elles sont les mieux placées au sein de l'université pour se faire les « évangélisateurs » des performances de cette technique et de ses potentiels pour les résultats de la recherche auprès des chercheurs. Mais au-delà, la bibliothèque est l'acteur idéal pour leur apporter la formation nécessaire aux manipulations des données, les compétences en matière d'indexation et de conservation ainsi que l'orientation documentaire indispensable à l'expertise juridique. En définitive, il apparaît clairement dans l'étude que le TDM requiert avant tout un bon niveau de culture numérique. C'est pourquoi les experts de la fouille de textes, les départements informatiques et les bibliothèques doivent coopérer pour contribuer à la montée en compétences des chercheurs, sur le modèle des collaborations qui se font jour entre les experts des outils numériques, les bibliothécaires et les scientifiques dans les projets menés en Humanités numériques.

Conclusions

Les modifications apportées récemment à la législation sur le droit d'auteur en matière de fouille de données ne suffisent pas à assurer les conditions propices au développement des usages du TDM dans l'enseignement supérieur. Elles doivent être accompagnées d'améliorations en termes d'accessibilité, d'infrastructures, de compétences et de mesures incitatives. Un pilotage fort et un engagement supplémentaire sont aujourd'hui nécessaires à la création d'un environnement véritablement favorable au TDM. L'adoption du TDM par la communauté des chercheurs ne se fera pas sans des efforts consentis au niveau financier et en termes d'image. Les « leaders » du TDM devront avant tout montrer l'efficacité de la fouille de données et ses retombées stratégiques sur la notoriété des projets afin de justifier la demande de moyens auprès des décideurs et lever les doutes sur la valeur des résultats. Ils devront porter des mesures incitatives tout autant afin de combattre le manque de sensibilisation et de financement au sein des équipes de recherche pour le partage des données et la curation, l'absence d'infrastructure et d'outils faciles d'utilisation et largement accessibles, la diversité des cultures disciplinaires universitaires, que d'accompagner et de soutenir l'investissement nécessaire à l'acquisition d'un seuil de compétences indispensable à la réalisation des premières expérimentations. Les bibliothèques[7] et les consortiums de bibliothèques en particulier ont un rôle stratégique à jouer pour soutenir et favoriser la fouille de textes et de données dans l'université : en offrant des services de soutien dédiés aux chercheurs, en leur proposant des guides de bonnes pratiques, ou bien encore en les accompagnant dans leur expertise juridique.

En résumé, le TDM ne deviendra pas une réalité de pratiques de la recherche universitaire sans :

- le législateur pour clarifier l'application de l'exception en cas de collaboration entre des chercheurs du public et des partenaires commerciaux et contrôler l'interaction de l'exception au droit d'auteur avec les autres régimes juridiques concernés des partenaires commerciaux ;
- les acteurs pilotes de la recherche pour investir dans le développement de services de bibliothèques et examiner les possibilités d'échange de connaissances avec les partenaires commerciaux ;
- les organismes de financement de la recherche et les décideurs politiques pour investir dans l'infrastructure nécessaire et soutenir la contribution du TDM à la construction de la science ouverte ;
- les éditeurs et les fournisseurs d'infrastructures pour faciliter l'accès au contenu protégé et s'accorder sur des normes ouvertes et des formats de données uniformisés.

Notes

[1] Définition partiellement extraite de l'étude comparative d'impact de la fouille de textes sur la compétitivité de la recherche en France et en Grande Bretagne. Pilotée et publiée par l'ADBU avec le soutien du Ministère de l'Education Nationale de l'Enseignement supérieur et de la Recherche (MENESR), l'étude a été confiée au prestataire britannique Research Consulting.

[2] La LOI n° 2016-1321 du 7 octobre 2016 pour une République numérique, dont les décrets d'application sont en cours de préparation, a été initialement proposée par la secrétaire d'État au numérique Axelle Lemaire. Elle a été promulguée le 7 octobre 2016 après une consultation publique (organisée fin 2015 par le Conseil National du Numérique dans le cadre de la concertation nationale sur le numérique), le travail parlementaire et l'adoption du projet de loi par le Sénat le 28 septembre. La loi porte sur 3 enjeux numériques majeurs : la circulation des

données et du savoir, la protection des droits dans la société numérique et l'accès au numérique. La législation sur le TDM relève du chapitre « Economie du savoir » et fait l'objet de l'article 38 qui modifie les articles L. 122-5 et L. 342-3 du code de la propriété intellectuelle.

[3] Research Consulting est un cabinet-conseil du Royaume-Uni spécialisé dans la gestion, la diffusion et la commercialisation de la recherche universitaire. Il conseille les organismes de financement, les universités, les bibliothèques et les éditeurs universitaires sur les changements de politiques et les développements technologiques en matière de recherche et publication de travaux universitaires.

[4] L'étude l'estime à 16 milliards, dont 10 pour l'enseignement supérieur et la recherche d'après les statistiques 2016 d'Eurostat, avec un retour sur investissement qui s'élève à 20%.

[5] Le projet ISTEX a consisté à créer une bibliothèque numérique accessible aux membres de l'enseignement supérieur et aux établissements de recherche en France. Grâce à un investissement dans l'acquisition de contenus qui s'élève à plus 55 millions d'euros depuis 2012, les données des éditeurs sont regroupées dans un même et unique corpus uniformisé, mis à la disposition des chercheurs. Les utilisateurs peuvent également demander le droit de télécharger des sous-corpus d'ISTEX et d'utiliser leurs propres outils pour fouiller le contenu.

[6] On peut citer à ce titre l'initiative du National Centre for Text Mining du Royaume-Uni qui a développé une plateforme web dédiée à la fouille de textes, Argo, afin de développer et d'exploiter des solutions d'analyse de textes. L'interface utilisateur graphique d'Argo est entièrement disponible via un navigateur web, ce qui rend la fouille de textes accessible aux chercheurs sans compétences particulières dans le développement de logiciel.

[7] À l'instar de la Bibliothèque nationale de France qui soutient une politique active en faveur de la fouille de textes et de données.