Légiférer sur le TDM, contexte et propositions

(TDM: *Text and Data Mining*)
Mars 2016

Résumé

La fouille de textes et données à des fins scientifiques est un outil fondamental de la recherche contemporaine. Permettre sa mise en œuvre légale en France est un enjeu de souveraineté scientifique. Il est donc **indispensable** que le projet de loi pour une république numérique inscrive le TDM dans la législation française « *pour les besoins de la recherche publique, à l'exclusion de toute finalité commerciale* » Cette courte note en rappelle les raisons et propose deux approches législatives possibles.

Tous les domaines de l'activité humaine sont bouleversés par la révolution numérique. C'est en particulier le cas du domaine qui a été à l'origine de cette révolution: la recherche scientifique. De nouvelles méthodes et outils voient le jour, permettant d'effectuer des recherches inenvisageables auparavant. La révolution numérique permet par exemple de traiter de manière efficace des masses particulièrement importantes de données, qu'il s'agisse de corpus de textes, de séquençage de génomes, d'images satellitaires ou de signaux issus d'ondes gravitationnelles.

Ainsi, un outil de base des scientifiques est aujourd'hui la *fouille de textes et données*, le « *text and data mining* » (abrégé en TDM dans la suite). Stricto sensu le concept n'est pas nouveau mais il est en plein développement, son automatisation permet aux scientifiques d'exploiter des ensembles gigantesques de données et de textes, hors de portée de traitements manuels.

Dans le cadre maintenant bien avancé d'élaboration de la loi pour une république numérique, le TDM peut être introduit de deux manières, au moins. Ces deux possibilités ne sont pas antagonistes et peuvent se compléter utilement.

La première possibilité, proposée par l'Assemblée Nationale et telle que transmise au Sénat consiste à *introduire une exception dans le Code de la Propriété Intellectuelle*. Cela est exprimé dans l'article 18bis et plusieurs documents, en particulier issus du CNRS, en analysent l'opportunité, nous n'y revenons pas ici.

La seconde possibilité consiste à i*ntroduire une nouvelle disposition au Code de la Recherche* permettant de statuer simplement sur le droit à disposer du TDM comme outil fondamental de la recherche scientifique. L'article 17 serait alors complété, d'un nouvel alinéa, a priori entre les 5 et 6 actuels, statuant:

La recherche publique bénéficie sans restriction du droit à l'extraction d'informations sur toutes les données publiées relevant de ses activités scientifiques (articles de revues scientifiques, actes de conférences, données d'expérimentation ou de simulation, vidéo, informations issues des réseaux sociaux, …).

Dans les attendus d'un tel article, on pourra noter que sans présumer des avancées technologiques ultérieures, la mise en œuvre du TDM pourra être réalisée ou bien via des API (Application Programmable Interface) que devront fournir librement et loyalement les plateformes, ou bien par l'intervention de tiers de confiance qui auront accès librement à toutes les informations scientifiques publiées ou bien encore être mis en œuvre sur des matériaux acquis contractuellement par ailleurs.

Nous détaillons maintenant le contexte et les raisons de la mise en place d'un droit fondamental au TDM pour la recherche scientifique publique.

De quoi s'agit-il ? L'activité scientifique produit un nombre de données considérable, que ce soit des articles de revues[1] ou d'actes de conférences, des documents publiés dans les archives ouvertes[2] ou encore des propositions de projet de recherche, des rapports d'expérience ou d'évaluation, des données d'expérimentation produites par des capteurs ou par des simulations numériques. Le TDM vise à produire des services permettant de découvrir dans ces ensembles de textes ou de données des informations utiles scientifiquement, par exemple :

- tous les articles traitant récemment du traitement d'une maladie, des gènes impliqués et des populations concernées,
- les corrélations entre deux expériences,
- les liens entre des articles scientifiques y compris dans des disciplines différentes,
- les personnes compétentes sur tel ou tel sujet ou résultat,
- …

Le TDM est le télescope numérique de la connaissance et en tant que tel les services qu'il offre deviennent essentiels aux activités scientifiques de haut niveau. Comment accéder aux bonnes informations, faire une synthèse, des corrélations, accéder à l'état de l'art sans ces informations ? **C'est un enjeu de souveraineté scientifique** qui est sous-jacent.

De manière globale et tout particulièrement (mais pas seulement) dans les développements scientifiques, l'enjeu de l'accès aux données est fondamental et nous semble devoir s'appuyer sur deux principes fondamentaux:

- **Une recherche financée par des fonds publics doit être ouverte** et en particulier, l'**accès aux articles et aux données scientifiques publiées, issus de la recherche financée par des fonds publics, doit être ouvert à tous**, académiques, industriels, citoyens;
- Leur exploitation, c'est à dire l**es services basés sur ces articles et données, doivent être ouverts à la concurrence.**

Notons que le TDM peut en fait concerner, dans le cadre d'activités scientifiques, deux types de données :

1. **Le TDM-Science**: qui est **réalisé sur des données issues de l'open data ou d'activités scientifiques** : articles de revues ou d'actes de conférences, données issues de capteurs ou de simulations, vidéos, etc…
2. **Le TDM-Général** : qui de manière plus étendue, **concerne toutes les données disponibles** et donc en plus des données concernées par le TDM-Science, les données de réseaux sociaux, données de l'internet en général, livres, essais, textes de la presse écrite, films, série TV, etc…

Le **projet de loi pour une république numérique tel que communiqué par l'Assemblée Nationale au Sénat** propose d'inscrire le TDM-Général à finalité scientifique sur les données et productions de la recherche dans la législation française « *pour les besoins de la recherche publique, à l'exclusion de toute finalité commerciale* » comme le précise le texte dans son article 18 bis.

Pour bien faire comprendre tout l'intérêt et l'importance de l'introduction de l'accès au TDM dans la loi, quelque soit la manière de le formuler, nous continuons par un exercice de questions-réponses.

- **Le TDM-Science concerne-t-il le dernier roman ou l'essai que je viens de publier ?**

✓ Non, le TDM-Science tel que définit ici concerne la fouille d'articles et de données publiés et conçus dans le cadre de recherches scientifiques. De même que pour un roman ou un essai, une tribune publiée dans un quotidien ou une interview ne seront pas concernés par les dispositions TDM-Science.

- **Y a t il une différence entre le TDM et l'open access ?**

✓ Oui. Mais ces notions sont souvent utilisées dans le même contexte et elles sont reliées. De manière très succincte et pour ce qui concerne le domaine scientifique, l'open access est la capacité à accéder librement à des publications alors que le TDM est la capacité à extraire des informations d'un ensemble,

---

[1] En 2012 plus de 28 000 revues étaient actives et 1,8 million d'articles publiés par an [Mabe 2012]. Actuellement l'éditeur Elsevier traite 1,3 millions d'articles qui lui sont soumis par an et dont quelques 30% sont publiés in fine.

[2] L'archive ouverte HAL comporte actuellement plus d'un million de référence dont quelques 30% sont en texte complet.

souvent très grand, d'articles et de données publiées. La différence est importante. Dans le premier cas, on peut lire et, en fonction de la licence, reproduire, communiquer tout ou partie du texte. Dans le cas du TDM, on utilisera des services[3] pour gérer l'exploration des articles et données sans nécessairement y accéder directement.

- **Est-ce qu'autoriser le TDM va permettre à un chercheur de profiter honteusement du travail d'un autre chercheur en pouvant utiliser abusivement des données accumulées par ce dernier ?**

✓ Le TDM ne permet de fouiller que des données publiées. Les services de TDM ne pourront pas accéder aux données tant que leur découvreur ou inventeur ne les aura pas publiées. Par exemple, rien n'obligera un historien à rendre publiques ses archives personnelles s'il a publié un article basé sur elles.

- **Quelle est la plus-value d'un éditeur commercial dans le monde de l'édition scientifique actuelle ?**

✓ Ce qui fait aujourd'hui la plus value essentielle d'un éditeur commercial c'est la réputation des revues qu'il met en œuvre, traduite en particulier par le fameux, mais controversé, « facteur d'impact ». Mais une revue repose d'abord fondamentalement sur le savoir et le travail des scientifiques: ce sont eux qui mènent les recherches, écrivent les articles et constituent les données. Ce sont eux encore qui constituent les comités éditoriaux des revues et qui font l'analyse des textes et des données qui sont soumis à publication. La plus-value d'un éditeur commercial est donc complètement dépendante des scientifiques.

Les grands éditeurs comme Springer et Elsevier ont parfaitement compris que ``vendre du pdf'' n'est pas l'avenir de l'édition, mais que les services autour des textes et des données constituent l'enjeu majeur. Leur stratégie s'est donc très clairement orientée vers 1) la maîtrise des textes et des données et 2) les services sur ces corpus. Elsevier, par exemple, a mis en place une politique d'open access vigoureuse avec des services permettant de valoriser les publications. L'offre de services Scopus est de ce point de vue remarquable et se base sur les données possédées par Elsevier, mais aussi sur toutes les données accessibles librement. Une autre valeur ajoutée sur les données textuelles réside dans leurs inter-relations. Là aussi l'exemple d'Elsevier est éclairant et son acquisition du réseau social Mendeley il y a deux ans lui permet par ce service d'accéder aux discussions et échanges entre scientifiques.

- **Existe-t-il déjà des revues dont tout le comité scientifique a démissionné pour créer à côté une revue « clone » en open-access?**

✓ Les exemples sont assez nombreux. L'un des premiers est le « Journal of Logic Programming » dont l'ensemble du comité éditorial a démissionné en 1999 pour passer d'Elsevier à Cambridge University Press, à l'époque une revue ``quasi open access''. Le comité éditorial du journal ``Lingua: An International Review of General Linguistics '', aussi publié chez Elsevier a démissionné en bloc en 2015 pour fonder la revue Glossa en association avec Ubiquity Press.

- **Existe-t-il des exemples de revues un peu « fameuses » en Open Access ?**

✓ Un exemple remarquable est la revue « Logical Methods in ComputerScience », dans le top 10 des revues à haut facteur d'impact en informatique fondamentale et dont le comité éditorial a validé début 2016 le passage dans epiSciences, une série de revues en open access natif, basé sur les archives ouvertes telles que HAL ou ArXiv avec le fort soutien d'Inria et du CNRS.

- **Quelle est la législation en vigueur dans les principaux pays étrangers ?**

✓ Le Japon et le Royaume-Uni ont mis en place une exception au droit d'auteur pour exploration des données scientifiques en 2009 et 2013 respectivement. En Israël, aux Etats-Unis et au Canada, les opérations de TDM à des fins scientifiques publiques sont permises par une notion appropriée d'*utilisation juste* adaptée à leur droit d'auteur i.e. le « fair use ».

« What my testimony should emphasise I think is the MASSIVE competitive advantage that the UK and Japan have over the rest of Europe when it comes to content mining. I can do a huge number of things now, quickly and easily and legally -- which my colleagues in Germany, France and elsewhere can't dream of doing because copyright law is so different in these jurisdictions. » Ross Mounce (Université de Cambridge, 2015).

- **Que penser des propositions consistant à permettre l'accès des scientifiques au TDM via des contrats passés par exemple avec les éditeurs ?**

✓ Pour prolonger la métaphore du télescope numérique, une telle situation reviendrait à vendre aux astronomes des télescopes permettant d'explorer uniquement une partie fixée de l'univers. Les différents télescopes étant par ailleurs difficilement compatibles entre eux. Plusieurs établissements de

[3] Typiquement des API (Application Programming Interfaces) ou l'implication de tiers de confiance

recherche public français travaillent actuellement avec plus de 80 éditeurs. Explorer la voute céleste en un point donné nécessiterait donc quelques 80 télescopes différents, clairement un contre sens.

- **Quelles conséquences auraient pour les chercheurs français le retrait de l'amendement autorisant le TDM à des fins de recherche ?**

✓ Comme présenté dans l'introduction, le TDM constitue un nouvel outil fondamental permettant aux scientifiques de faire de la recherche au meilleur niveau international. De même que sans microscope ou sans télescope un biologiste ou un astronome ne pourra pas faire de recherche au niveau international, de même, sans télescope numérique, i.e. sans TDM, les scientifiques français ne pourront pas développer leur recherche au niveau international. Sans TDM en France, les scientifiques iront pratiquer leur métier ailleurs, en particulier aux USA, au Royaume-Uni ou au Japon.

- **Quelles seraient les conséquences pratiques d'une loi française en contradiction avec la loi d'un autre pays concernant le TDM ? Est-ce que c'est la loi du pays du chercheur, du pays de l'éditeur, … qui s'appliquerait ?**

✓ Comme développé dans plusieurs des références données à la fin de cette note, c'est bien la loi du pays dans lequel la recherche est pratiquée qui s'applique. C'est ce qui est mis en œuvre par les pays qui ont mis en place une exception TDM scientifique comme le Royaume-Uni.

- **Comment s'assurer que des entreprises commerciales ne pourront pas profiter de la tolérance de la loi pour utiliser le TDM à des fins mercantiles ?**

✓ Le texte de la proposition de loi précise sans ambiguïté que le TDM est autorisé « *pour les besoins de la recherche publique, à l'exclusion de toute finalité commerciale* ». Une université qui utiliserait son droit au TDM à des fins mercantiles serait très clairement hors la loi.

- **Si en utilisant entre autres leur droit au TDM, les scientifiques font des découvertes qu'ils brevettent ou dont leur établissement tire profit, cela contredit-il l'engagement de non finalité commerciale ?**

✓ Non car dans ce cas il y a une valeur ajoutée, dérivée des informations obtenues.

- **Une entreprise pourra-t-elle développer des activités commerciales sur le traitement des connaissances ?**

✓ Oui et ceci est souhaitable. Les entreprises comme les éditeurs développent des services, dont du TDM, sur les textes et les données auxquelles elles ont légalement accès: typiquement les textes et les données qu'elles possèdent ou qui sont en open access à des fins commerciales. L'ouverture des données permet l'ouverture à la concurrence des services rendus aux scientifiques. Un établissement académique pourra acheter un service lui donnant accès à Scopus ou au Web of Science (Wos) afin de mettre en place sa stratégie ou comprendre les impacts de sa production ou encore comprendre avec quelle université étrangère il serait pertinent de collaborer sur telle ou telle thématique.

- **Un établissement académique devra-t-il nécessairement faire appel à une entreprise commerciale pour accéder à des services d'accès à la connaissance ?**

✓ Non bien sur. L'établissement pourra développer ses propres outils ou utiliser ceux mis en place par d'autres universités ou établissements français ou étranger. Il pourra aussi s'appuyer sur des outils mis en place par le service public au-dessus des archives ouvertes telles que HAL[4] ou ArXiv. Cela pourra aussi faire l'objet de coopérations internationales entre archives institutionnelles nationales.

- **Comment gérer le risque de voir une entreprise comme Google proposer des outils de TDM aux chercheurs ? Ainsi Google aurait là encore accès aux données de manière indirecte.**

✓ Ce risque est aujourd'hui avéré: la fouille de données est souvent réalisée avec des outils comme Google ou de manière plus sophistiquée aujourd'hui avec Watson (d'IBM). Dans les deux cas, l'entreprise ne peut accéder qu'à ses propres données ou à celle qui sont en Open Access, par ailleurs la reproductibilité et la loyauté ne sont pas assurées. C'est pour se donner d'autres possibilités que ces initiatives actuellement quasi uniques qu'il est important de permettre la mise en place d'un TDM maîtrisé au niveau de la recherche publique française et si possible européenne. C'est un enjeu fondamental de souveraineté.

---

[4] Un tel service se mets actuellement en place dans le cadre d'Inria et du CCSD: anHALytics.