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

La fouille de textes et de données dans l'enseignement supérieur et la recherche publique

Une analyse d'études de cas menées au Royaume-Uni et en France

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

Research Consulting

L'importance du TDM

Qu'est ce que le TDM ?

« Le TDM (Text & Data Mining, ou fouille de textes et de données) désigne toute technique d'analyse automatisée visant à analyser des textes et des données sous forme numérique afin d'en dégager des informations telles que des constantes, des tendances et des corrélations. »

On observe 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 s'avère compliqué pour les chercheurs de passer en revue et d'exploiter cet ensemble de connaissances qui ne cesse de s'enrichir.

La fouille de textes et de données (TDM) repose sur des logiciels de pointe pour l'analyse de ces connaissances et d'autres sources d'informations numériques, qu'il serait impossible de passer au crible manuellement.

Le TDM favorise la diffusion et la performance de la recherche scientifique

Des études préalables ont indiqué que le TDM permettait de :

- Multiplier **par quatre** la couverture de la connaissance dans le domaine de la biologie des systèmes
- Identifier les études pertinentes en matière de politique publique avec seulement **25 % du temps de travail habituellement nécessaire**
- Améliorer la productivité en termes de curation de la documentation biomédicale de **50 %,** permettant des économies de **70 000€ par an** pour une seule base de données
- Identifier les meilleures études dans le secteur de la santé parmi les **1,3 million de publications** par an - une opération qui prendrait **2 à 3 années** et coûterait **100 000€** si elle était effectuée manuellement
- Accélérer la découverte de nouveaux médicaments, et réduire de **12 ans la période** nécessaire à la mise sur le marché de ces médicaments

Les chercheurs européens risquent de prendre du retard sur le reste du monde

Au cours de ces dix dernières années, l'Europe a été dépassée par l'Asie et lui a laissé sa place de premier centre mondial de recherche universitaire sur le TDM, au vu du nombre d'articles publiés. Aux États-Unis, les chercheurs peuvent profiter de la doctrine du *fair use* pour procéder à l'exploration de textes et de données à partir de la littérature scientifique accessible, et éviter les processus complexes de signature de licence.

L'introduction d'exceptions au droit d'auteur en faveur du TDM au Royaume-Uni et en France permet à nos chercheurs d'être sur un pied d'égalité avec l'Asie et l'Amérique du Nord. Néanmoins, les modifications apportées à la législation sur le droit d'auteur 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.

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

Favoriser l'usage du TDM

Fournir un cadre juridique clair

La Chine est le leader mondial de dépôt de brevets pour des techniques de TDM et les États-Unis produisent le plus grand nombre de publications basées sur de la fouille de textes. Dans le même temps, la crainte d'enfreindre accidentellement la législation sur le droit d'auteur empêche les chercheurs européens de profiter pleinement des avantages du TDM.

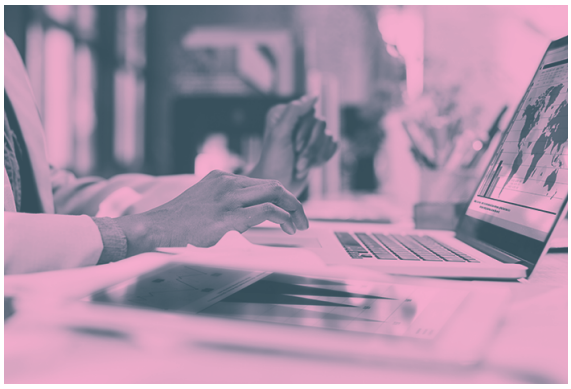
> « La définition qui oppose l'usage commercial et non commercial suscite des incertitudes au sein du secteur universitaire. »
>
> Petr Knoth, Open University (UK)

Les exceptions au droit d'auteur permettent d'améliorer les choses, mais les chercheurs sont encore confrontés à des incertitudes juridiques :

- La charge incombe aux chercheurs et à leur institution de prouver qu'ils répondent aux critères de l'exception, tandis que le conseil et l'accompagnement d'experts dans ces démarches font défaut.
- Les chercheurs au Royaume-Uni et en France craignent que leurs travaux de recherche ne remplissent pas les conditions d'usage non-commercial, et ne puissent donc pas bénéficier de l'exception.
- L'exception française s'applique à tout type d'œuvres textuelles protégées, ainsi qu'aux données incluses ou associées aux écrits scientifiques, sous forme d'œuvre individuelle ou de bases de données.
- Seuls les organismes désignés par décret sont habilités à conserver et communiquer les copies techniques des bases de données, produites aux fins de TDM.

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



Faciliter l'accès

Les chercheurs déclarent rencontrer des difficultés pour obtenir l'accès à du contenu publié, et doivent souvent en rabattre sur leurs ambitions de recherche en conséquence. Le contenu peut être récupéré via le téléchargement de masse ou l'exploration du web, mais dans les deux cas les chercheurs risquent de rencontrer des obstacles :

- **Les Mesures Techniques de Protection.** Les éditeurs peuvent imposer des limites sur la vitesse ou le volume des téléchargements afin de s'assurer que la fouille n'altère ou ne réduise pas la qualité de leur service.
- **Les pièges à robots d'indexation.** Les éditeurs introduisent des pages web qui ressemblent à un article universitaire dans un programme de téléchargement automatique, et bloquent l'accès au reste du contenu de l'éditeur lorsque le logiciel de fouille est détecté.
- **L'accès restreint aux API.** Les Interfaces de Programmation Applicative (API) peuvent être utilisées par les chercheurs pour récupérer du contenu sous une forme exploitable par une machine. Toutefois, tous les éditeurs ne donnent pas accès à une API, d'autres demandent aux utilisateurs de signer un accord supplémentaire avant d'autoriser l'accès, et les différents formats de données entre les éditeurs compliquent l'agrégation de contenus.

Accompagner la montée en compétences

Le TDM requiert un bon niveau de culture numérique. Les experts de la fouille de textes, les départements informatiques et les bibliothèques peuvent tous jouer un rôle dans la montée en compétences des chercheurs. L'amélioration des niveaux de collaboration entre les experts de la fouille de textes et les scientifiques de la discipline s'avèrera nécessaire à mesure que le TDM évoluera.

Développer l'infrastructure

L'existence d'une exception au droit d'auteur peut décourager les éditeurs et fournisseurs de contenu à développer des solutions de TDM, l'investissement doit donc provenir de fonds publics. L'investissement de la France dans le projet ISTEX (www.istex.fr) a permis une avancée significative en matière de contenu, d'infrastructure et d'expertise, très prometteuse pour l'avenir. La poursuite des efforts est à présent nécessaire pour :

- **Améliorer les technologies utilisées pour compiler, uniformiser, interroger et préserver la matière issue de la fouille de textes et de données**
- **Encourager un usage plus large d'ISTEX pour la fouille de textes et de données, et améliorer la disponibilité de contenus récents**
- **Développer des services en ligne orientés usagers, accessibles et adaptés aux chercheurs dotés de compétences techniques limitées**

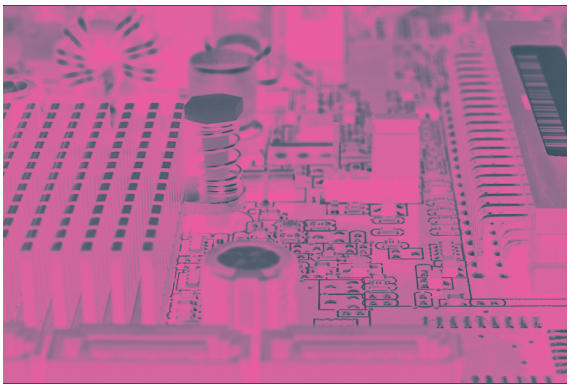
> « À moins que nous ne réduisions ces écarts au niveau des compétences, des logiciels et de l'appui juridique, l'Europe va devenir une nation du tiers monde dans le domaine de la fouille de textes et de données. »
>
> Peter Murray-Rust,
> Université de Cambridge

Mettre en place des mesures incitatives

Le TDM fait globalement face aux mêmes défis culturels que le mouvement plus global vers la science ouverte, avec notamment :

- **Le manque de sensibilisation sur le sujet**
- **La force et la diversité des cultures disciplinaires**
- **Les efforts nécessaires à l'acquisition des compétences suffisantes à la réalisation des premières expérimentations**
- **L'absence d'infrastructure et d'outils faciles d'utilisation et largement accessibles**
- **Les doutes quant à la valeur des résultats**
- **Les financements limités pour le partage des données et la curation**

De plus grands efforts devront être réalisés au niveau financier et en termes d'image afin de faciliter l'adoption du TDM par la communauté des chercheurs.



« Le monde réalise aujourd'hui que nous disposons de beaucoup de données, mais on continue à poser des questions de recherche qui datent du siècle dernier. »

Francois Rioult, Université de Caen (France)

Favoriser l'accès au contenu

Le projet ISTEX consiste 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.
www.istex.fr

Développer des outils web

Le National Centre for Text Mining du Royaume-Uni 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.
http://argo.nactem.ac.uk/

Créer une e-infrastructure en accès libre

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. 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.
www.openminted.eu/about/partners/inra/

Communiquer sur l'importance du TDM

ContentMine est une entreprise à but non lucratif basée à Cambridge, au Royaume-Uni, qui a pour ambition de « libérer les données scientifiques des revues académiques » et permettre à quiconque d'effectuer des recherches à l'aide du TDM. L'entreprise fait la promotion du TDM auprès des chercheurs qui se heurtent à un volume massif de contenus. Ses conférences et ateliers ont déjà réuni plus de 2000 personnes. www.contentmine.org

Le rôle des bibliothèques

Les recommandations à la communauté des bibliothèques

Les bibliothèques et les consortiums de bibliothèques peuvent soutenir et favoriser la fouille de textes et de données grâce à :

1. **La mise en place de mécanismes de suivi des expériences menées par les chercheurs**
2. **L'élaboration d'études de cas et la définition de guides de bonne pratiques**
3. **L'appui soutenu de la bibliothèque nationale en faveur de la fouille de textes et de données**
4. **L'intégration de clauses relatives à la fouille de textes et de données dans les accords de licence**
5. **Le développement de services de soutien dédiés à la fouille de textes et de données pour les chercheurs**

« Nous devons parvenir au stade où la bibliothèque sera en mesure de dire : '*si vous avez la moindre question à propos de la fouille de textes et de données, nous sommes là pour y répondre*' »

Danny Kingsley, Université de Cambridge (UK)

Le soutien des bibliothèques en faveur de la fouille de textes et de données

Un service complet de soutien des bibliothèques pour la fouille de textes et de données fonctionnerait sur la base d'un partenariat avec les chercheurs afin d'offrir :

1. **La promotion des bénéfices de la fouille de textes et de données à tous les niveaux de l'organisation**
2. **Des conseils juridiques sur l'utilisation de l'exception TDM, sur les restrictions de licences qui peuvent être ignorées et sur la manière d'attribuer les sources, notamment en cas d'utilisation de données ouvertes**
3. **L'accès à une expertise juridique**
4. **Le développement de compétences sur l'indexation et la curation de métadonnées, et l'accès à des formations techniques de codage ou de l'utilisation de systèmes de calcul haute performance (HPC)**
5. **Des conseils sur les outils et sources de données disponibles dans les collections des bibliothèques et plus largement en ligne**

Faire de la fouille de textes et de données une réalité

Les législateurs

- Énoncer les dispositions légales de la fouille de textes et de données et apporter des réponses juridiquement sécurisées et clairement exprimées
- Clarifier l'application de l'exception en cas de collaboration entre des chercheurs du public et des partenaires commerciaux
- Contrôler l'interaction de l'exception au droit d'auteur avec les autres régimes juridiques concernés

Les pilotes de la recherche et les responsables institutionnels

- Faire connaître les bénéfices de la fouille de textes et de données à la communauté des chercheurs
- Investir dans le développement de services de bibliothèques pour encourager la fouille de textes et de données
- Examiner les possibilités d'échange de connaissances en matière de fouille de textes et de données avec les partenaires commerciaux

Les organismes de financement de la recherche et les décideurs politiques

- Investir dans l'infrastructure nécessaire au soutien et au développement de la fouille de textes et de données
- Identifier ou créer un forum national afin de répondre aux défis en termes d'accès
- Considérer les besoins des chercheurs qui pratiquent la fouille de textes et de données comme faisant partie intégrante du mouvement vers la science ouverte

Les éditeurs et les fournisseurs d'infrastructures

- Poursuivre le développement de services dans les nuages en faveur de la fouille de textes et de données
- Prendre des dispositions pour faciliter l'accès à tout type de contenu protégé, à des fins de fouille
- Adopter des normes ouvertes et des formats de données uniformisés

Cadre de l'étude

Ce document de synthèse est fondé sur un rapport, préparé par Research Consulting au nom de l'ADBU, qui traite des enjeux d'ordre pratique, organisationnel et juridique relatifs à l'usage de la fouille de textes et de données (ou TDM, pour Text and Data Mining) pour la recherche universitaire.

Cette étude fournit une analyse comparative du cadre réglementaire en France, au Royaume-Uni et dans l'Union européenne, et examine les obstacles et les éléments favorables à l'usage de la fouille de textes et de données, tels qu'ils apparaissent à la lumière d'un certain nombre d'études de cas. Ces études de cas s'appuient sur les différentes expériences des chercheurs et des spécialistes de la fouille de textes et de données qui opèrent en France, au Royaume-Uni, dans d'autres pays de l'Union européenne et aux États-Unis.

Le rapport complet est disponible sur www.adbu.fr/etude-tdm

Contacts

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

L'ADBU est l'association des directeurs et des personnels de direction des bibliothèques universitaires françaises et de la documentation. Elle œuvre, en collaboration avec les décideurs politiques, à la reconnaissance des bibliothèques universitaires et des établissements publics de recherche, au maintien des normes de qualité de l'information technique et scientifique, et apporte son appui à la communauté des universitaires et professionnels des bibliothèques.

Contact : Julien Roche
vp@adbu.fr
www.adbu.fr

Research Consulting

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.

Contact : Rob Johnson
rob.johnson@research-consulting.com
www.research-consulting.com