Mémoire au Comité INDU sur l'exploration de textes et de données

Présenté par le Réseau Portage

Le 4 décembre 2018

portage
SERVICES PARTAGÉS POUR LES DONNÉES DE RECHERCHE
SHARED STEWARDSHIP OF RESEARCH DATA

CARL ABRC
CANADIAN ASSOCIATION OF RESEARCH LIBRARIES
ASSOCIATION DES BIBLIOTHÈQUES DE RECHERCHE DU CANADA

www.carl-abrc.ca

Introduction

Portage se réjouit de l'occasion qui lui est donnée d'apporter une contribution à l'examen de la *Loi sur le droit d'auteur* par le Comité permanent de l'industrie, des sciences et de la technologie. Portage est un réseau national de gestion des données de recherche (GDR) rassemblant le milieu des bibliothèques, parrainé par l'Association des bibliothèques de recherche du Canada (ABRC). Notre objectif principal est d'augmenter l'accessibilité, la longévité et l'utilisabilité des données pour les chercheuses et chercheurs canadiens et le grand public.

Le lien entre données et droit d'auteur

Les données et les renseignements factuels[1] ne sont pas classés comme des œuvres de l'esprit originales protégées par le droit d'auteur au Canada, mais certains types de données sont composés d'œuvres protégées. Ces données peuvent aussi être compilées pour former une nouvelle œuvre protégée par le droit d'auteur.

L'exploration de textes et de données (ETD) est une technique de recherche de plus en plus courante. Il s'agit d'un processus automatisé de recherche de tendances dans des données extraites de grandes quantités de documents, dont certains sont susceptibles d'être protégés par le droit d'auteur. Les chercheurs prennent le matériau extrait, le transforment pour obtenir un nouveau format lisible par machine et explorent les données pour « découvrir de nouvelles connaissances, tester des hypothèses et trouver de nouvelles relations »[2]. L'exploration de textes et de données est également appliquée à l'apprentissage automatique, qui est une composante cruciale du développement de l'intelligence artificielle (IA). L'apprentissage automatique peut impliquer la consommation par des machines de grandes quantités d'œuvres créées par des êtres humains, dont un grand nombre est susceptible d'être protégé par le droit d'auteur. Il faut noter que les copies réalisées pendant le processus d'exploration de textes et de données ne peuvent pas être interprétées comme constituant une concurrence sur le marché aux œuvres originales protégées par le droit d'auteur, car les données extraites produisent un ou plusieurs ensembles nouveaux de données. La seule application pratique des données extraites se trouve dans la recherche connexe ou le perfectionnement d'algorithmes, et les ensembles de données extraits ne sont pas rendus publics.

L'exploration de textes et de données conduit à diverses applications cruciales pour la recherche, dont beaucoup, comme l'affirme la Ligue des bibliothèques européennes de recherche :

> « ... accroîtront de façon exponentielle les progrès scientifiques. Elles pourraient faciliter la découverte de traitements pour des maladies comme le cancer et la maladie de Parkinson. L'exploration de textes et de données a déjà servi à découvrir de nouvelles applications de médicaments connus et constituera un fondement de l'innovation et des nouvelles industries. Pour les bibliothèques, cela signifie que les chercheurs que nous

[1] Par exemple : « les mesures de précipitations et de température, les taux de mortalité, le nombre d'habitants, les valeurs monétaires, les structures chimiques, les faits historiques et les dates, le nombre d'abonnés à un compte Twitter » (tiré de : Bibliothèque de la Simon Fraser University, https://www.lib.sfu.ca/help/academic-integrity/copyright/data-copyright).

[2] Liber. Text and Data Mining. The need for change in Europe. [L'exploration de textes et de données. Pour un changement en Europe]. https://libereurope.eu/wp- content/uploads/2014/11/Liber-TDM-Factsheet-v2.pdf.

soutenons seront en mesure de tirer pleinement parti de la valeur de nos collections scientifiques en essor constant. Cela entraînera des méthodes de recherche plus rigoureuses, notamment des revues de littérature plus exhaustives[3]. »

L'ETD peut s'appliquer à d'autres domaines de recherche comme : l'exploration de la presse écrite, des indicateurs textuels d'incertitude économique, des changements politiques ou des tendances sociales, mais aussi l'exploration de catalogues de bibliothèque à grande échelle, d'autres dépôts de connaissances en ligne ou d'agrégations de médias sociaux pour comprendre les changements qui s'opèrent dans les technologies, l'édition, les comportements des consommateurs, etc.[4]

Le monde de la recherche a de plus en plus recours à l'ETD et à l'analyse de textes dans plusieurs méthodes computationnelles de recherche, et l'organisation des résultats de ces travaux pourrait donner lieu à de nouvelles découvertes au sein d'organismes de recherche ou par la coopération entre entités de recherche. Le réseau d'organisation des données de Portage aide les chercheuses et les chercheurs à produire des résultats de recherche reproductibles dans l'application de l'ETD.

Aux États-Unis, des décisions judiciaires récentes ont reconnu le « fondement juridique solide de la recherche respectueuse sur des documents protégés par le droit d'auteur ». En conséquence, des organisations comme le HathiTrust Research Center donnent maintenant « accès à l'ensemble des textes du corpus d'HathiTrust, soit 16,7 millions d'articles, aux fins de recherche respectueuse, comme l'exploration de données et l'analyse computationnelle, y compris à des articles protégés par le droit d'auteur ». Cette politique repose sur la prémisse selon laquelle l'utilisation respectueuse de recherches, comme l'ETD, n'empiète pas sur le statut juridique des articles protégés par le droit d'auteur ni le « modifie »[5].

Deux solutions permettraient de résoudre les problèmes posés par le droit d'auteur en matière d'apprentissage automatique et d'autres applications de recherche sur des mégadonnées. La première serait d'imiter le modèle américain d'utilisation équitable en faisant en sorte que la liste actuelle des fins d'utilisation équitable soit indicative plutôt qu'exhaustive. La deuxième consisterait à établir une exception particulière pour l'exploration de textes et de données ou l'analyse informatique[6]. Bien que l'ETD puisse déjà être autorisée en vertu d'exceptions canadiennes comme l'utilisation équitable, et qu'elle puisse être autorisée par des dispositions dans les contrats d'achat des bibliothèques, une exception propre à l'ETD apporterait de la clarté et donnerait au milieu canadien de la recherche un avantage qui stimulera les découvertes et l'innovation.

---

[3] Liber. Text and Data Mining. The need for change in Europe. [L'exploration de textes et de données. Pour un changement en Europe]. https://libereurope.eu/wp- content/uploads/2014/11/Liber-TDM-Factsheet-v2.pdf.

[4] Dyas-Correia, Sharon, Alexopoulos, Michelle. Text and Data Mining: Searching for Buried Treasures [L'exploration de textes et de données. À la recherche de trésors enfouis], Serials Review, 00987913, sept. 2014, vol. 40, Numéro 3, https://www.tandfonline.com/doi/abs/10.1080/00987913.2014.950041.

[5] HathiTrust Research Center Extends Non-Consumptive Research tools to Copyrighted Materials: Expanding Research through Fair Use [L'HathiTrust Research Center étend les outils de recherche respectueuse aux documents protégés par le droit d'auteur. Développer la recherche par l'utilisation équitable]. https://www.hathitrust.org/blogs/perspectives-from-hathitrust/hathitrust-research-center-extends-non-consumptive-research-tools.

[6] Geist, Michael. Why copyright law poses a barrier to Canadian AI ambitions [Pourquoi le droit d'auteur constitue-t-il une entrave aux ambitions canadiennes dans le domaine de l'IA?], https://www.theglobeandmail.com/report-on-business/rob-commentary/why-copyright-law-poses-a- barrier-to-canadian-ai-ambitions/article35019241/.

De plus, de nombreuses sources essentielles dans l'ETD sont des bases de données dont les modalités d'utilisation sont négociées entre bibliothèques et éditeurs ou entre utilisateurs et éditeurs. Dans bien des cas, les détenteurs des droits utilisent les licences « pour outrepasser les exceptions du droit d'auteur qui ont été créées par des processus législatifs transparents, aux dépens des utilisateurs et au détriment de la diffusion des connaissances, de la découverte et de l'innovation »[7]. Il faut explicitement énoncer que toute exception relative à l'ETD ne peut être annulée ou outrepassée par contrat.

Enfin, la nouvelle exception ne doit pas se limiter à la recherche scientifique ou à des fins non commerciales, car l'ETD a diverses applications commerciales et interdisciplinaires.

Pourquoi le milieu de la recherche pense-t-il qu'il est important d'ajouter des clauses relatives à l'ETD dans la Loi?

Un nombre important de témoignages de chercheuses et chercheurs canadiens est fourni à la fin du présent document. En voici deux exemples :

> *« Il est vraiment difficile d'explorer des textes et des données sur des documents protégés par le droit d'auteur au Canada, à tel point que, bien qu'il soit prouvé que ce genre d'analyse est possible à grande échelle, très peu de travaux de la sorte ont été réalisés. Dans de nombreux cas, il est tout simplement trop onéreux, en apparence en tout cas, de s'y retrouveren matière de droits d'auteur. Cela est particulièrement vrai pour les étudiantes et étudiants des cycles supérieurs et les chercheuses et chercheurs en début de carrière, qui peuvent avoir les compétences et l'énergie nécessaires, mais ni le temps ni le soutien qu'il faut pour se frayer un chemin dans le processus. Une exemption [relative à l'ETD] ouvrirait la voie à la recherche transformationnelle et à une nouvelle génération d'universitaires qui tirerait des idées inédites et passionnantes du nombre sans cesse croissant de documents numériques. »*
>
> Ian Milligan, Professeur agrégé, Département d'histoire, University of Waterloo, septembre 2018.

> *« Imaginez ce scénario : vous allez à la bibliothèque et empruntez un livre, mais vous n'avez PAS le droit de le lire. Vous pouvez le regarder, le toucher, mais le fait de le lire est illégal. C'est fondamentalement à cette situation que ressemble l'interdiction de l'exploration de textes et de données. Nous avons à notre disposition toute cette information légalement entreposée dans nos bibliothèques, mais nous ne sommes autorisés à la comprendre qu'au moyen de méthodes dépassées. C'est absurde. L'exploration de textes et de données est tout simplement une autre forme de lecture. Il est grand temps de libérer les trésors entreposés dans les archives et les bibliothèques.*

[7] « Énoncé de position de la FCAB : Protection des exceptions sur le droit d'auteur de la préséance de contrats » (FCAB), consulté le 22 avril 2018 http://cfla-fcab.ca/wp-content/uploads/2018/02/FCAB-CFLA_enonce_preseance_contrats.pdf .

Les chercheurs n'essaient pas d'enfreindre le droit d'auteur. Ils cherchent seulement à étudier des textes, comme ils l'ont toujours fait, à ceci près que, maintenant, cette recherche est illégale. Cette entrave à la promotion de nouvelles connaissances est un échec pour notre société. »

Andrew Piper, Professeur, Département des langues, littératures, et cultures, et directeur de .txtLAB, à l'Université McGill, octobre 2018.

Recommandation : modifier la *Loi sur le droit d'auteur* afin d'autoriser l'exploration de textes et de données

L'autorisation de l'ETD sans la permission du titulaire de droits apporterait de la clarté et est susceptible de donner aux chercheuses et chercheurs canadiens un avantage qui stimulera les découvertes et l'innovation.

Pour y parvenir, on peut élargir la notion d'utilisation équitable afin qu'elle s'applique à toutes les fins comme aux États-Unis ou établir une exception propre à l'ETD. Il doit être explicitement énoncé que cette nouvelle exception ne se limite pas à la recherche à des fins scientifiques ou non commerciales et qu'elle ne peut être annulée par contrat.

À propos de Portage

Lancé en 2015 par l'Association des bibliothèques de recherche du Canada (ABRC), Portage est une initiative visant à promouvoir l'intendance partagée des données de recherche et à combler les lacunes de l'écosystème national de gestion des données de recherche en matière de politiques, de services et d'infrastructure. Portage a accompli de grands progrès dans sa mission : de façon collaborative, il a élaboré et offert plusieurs services et plateformes conçus pour aider les chercheuses et chercheurs canadiens à mieux gérer, stocker et partager leurs données. De plus amples renseignements sur le réseau Portage se trouvent à l'adresse suivante :
https://portagenetwork.ca/fr/a-propos-de-portage/

Lectures complémentaires

De nombreuses organisations représentant le monde de la recherche et des bibliothèques appuyant les chercheurs ont publié des énoncés ou des mémoires sur l'ETD.

Ligue des bibliothèques européennes de recherche https://libereurope.eu/text-data-mining/

Fédération européenne des académies nationales des sciences et des humanités https://www.allea.org/wp-content/uploads/2017/11/PWGIPR_Statement_TDM_2017.pdf

Association of Research Libraries [Association des bibliothèques de recherche] : http://www.arl.org/storage/documents/TDM-5JUNE2015.pdf

HathiTrust Research Center : HathiTrust Research Center Extends Non-Consumptive Research Tools to Copyrighted Materials: Expanding Research through Fair Use [L'HathiTrust Research Center étend les outils de recherche respectueuse aux documents protégés par le droit d'auteur. Développer la recherche par l'utilisation équitable]. https://www.hathitrust.org/blogs/perspectives-from-hathitrust/hathitrust-research- center-extends-non-consumptive-research-tools

Fédération internationale des associations et institutions de bibliothèques (IFLA) : https://libereurope.eu/blog/2015/06/09/liber-response-to-stm-statement-on-text- and-data-mining/

Annexe : Autres témoignages de chercheuses et chercheurs canadiens

Geoffrey Rockwell, Professeur de philosophie et d'humanités numériques, University of Alberta; Directeur du Kule Institute for Advanced Study Arts :

> *« Il est essentiel que les chercheurs soient autorisés à utiliser des méthodes respectueuses sur des documents protégés par le droit d'auteur si nous voulons être en mesure d'étudier la culture contemporaine en exploitant des mégadonnées. Les en empêcher revient à leur interdire de lire des documents avec l'aide de la technologie. Les chercheurs ne seraient pas autorisés à chercher un mot dans un livre électronique protégé par le droit d'auteur? Ils ne seraient pas autorisés à compter les occurrences? Ils ne seraient pas autorisés à compter les mots d'une page, que ce soit manuellement ou aidés par un ordinateur? Il est important de rappeler que nous parlons de méthodes respectueuses à partir desquelles on ne peut pas reconstituer le texte ni être considéré comme étant en train de le « copier ». C'est cela qui devrait être interdit. »*

Patrick Drouin, Professeur titulaire, Département de linguistique et de traduction, Université de Montréal; directeur de l'Observatoire de linguistique Sens-Texte

> *« Nous avons besoin, sur une base régulière, de constituer des corpus textuels importants (je dirais même énormes) pour les analyses statistiques afin de procéder à des descriptions lexicographiques et terminographiques. La seule solution viable pour nous consiste à aspirer des quantités de données importantes du Web (sans savoir l'objectif de les diffuser ensuite). Pour le moment, nous ne cherchons pas à obtenir le consentement tout simplement parce que cette procédure nous prendrait des mois et qu'elle demanderait une énergie et une expertise que nous n'avons pas. »*

Chantal Gagnon, Professeure agrégée, Département de linguistique et de traduction, Université de Montréal

> *« Je fais du forage et de l'exploitation de texte depuis un peu plus d'une dizaine d'années. Je travaille notamment sur les textes de la presse écrite canadienne, en français et en anglais. La question des droits d'auteur est un véritable casse-tête pour les chercheurs comme moi. Dans l'un de mes projets de recherche, j'ai passé un temps fou à échanger avec l'un des titulaires des droits d'auteur, et à manœuvrer*

pour trouver une façon d'analyser les données tout en respectant les multiples restrictions instaurées par le titulaire. Résultat : j'ai perdu un temps précieux et encore aujourd'hui, je ne peux pas partager mon corpus ouvertement, comme le recommandent pourtant les organismes subventionnaires. Le partage des données fait pourtant partie des bonnes pratiques de la recherche, et favorise l'innovation. »

Sylvie Vandaele, Professeure titulaire, Département de linguistique et de traduction, Université de Montréal

« Accéder rapidement à des corpus est crucial pour pouvoir assurer l'efficacité de la recherche. Certaines questions de recherche imposent d'avoir recours à des textes récents sous droits d'auteur. Or, les démarches sont chronophages, sans toutefois être toujours couronnées de succès. »

Susan Brown, Professeure, School of English and Theatre Studies [École de langue anglaise et de théâtre], University of Guelph

« Pour l'avenir de la recherche au Canada, il est crucial d'autoriser l'utilisation respectueuse de documents protégés par le droit d'auteur sans la permission du détenteur du droit d'auteur. Pour évoquer un cas particulier, l'étude de la littérature et de la culture canadiennes à l'aide des techniques computationnelles les plus récentes est pratiquement bloquée, car le patrimoine culturel publié de notre pays relativement jeune est presque totalement protégé par le droit d'auteur. Non seulement les obstacles actuels à l'exploration de textes et de données retardent la recherche et la formation dans ces domaines, mais ils privent aussi les Canadiens de la connaissance de nos histoires et de notre société que leur donnerait l'application de ces nouvelles méthodes passionnantes. »