Brief to the INDU Committee on Text and Data Mining

Submitted by the Portage Network

December 4, 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 welcomes the opportunity to contribute to the Standing Committee on Industry, Science and Technology review of the Copyright Act. Portage is a national, library-based, research data management (RDM) network sponsored by the Canadian Association of Research Libraries (CARL). Our primary goal is to make data more accessible, enduring, and useable for Canadian researchers and the public at large.

The relationship between data and copyright

Data and factual information[1] are not classified as original works of authorship protected by copyright in Canada, but certain types of data are made up of copyrighted works. It is also possible that these data may also be compiled to make a new copyrighted work.

An increasingly common technique for conducting research with data is text and data mining (TDM). TDM is the automated process of identifying patterns from data extracted from large quantities of material – some of which may fall under copyright. Researchers take this extracted material, transform it into new machine-readable formats, and mine the data to "discover new knowledge, test hypotheses and identify new relationships."[2] Another application for text and data mining is machine learning, a crucial component in the development of artificial intelligence (AI). Machine learning can entail machines consuming vast amounts of human created works, many of which would be copyrightable. What is crucial to note is that copies made during the text and data mining process cannot be interpreted to compete with the market for the original copyrighted works; they are effectively extracting data to create a new data set or sets. The only practical application for these extracted data is for the associated research or algorithmic refinement and the mined data sets are not made publicly available.

Text and data mining has a wide variety of crucial research applications, many of which, as the Association of European Research Libraries states:

> …will increase the progress of science exponentially. It has the potential to facilitate the discovery of cures for diseases such as cancer and Parkinson's. It has already been used to discover new applications for existing drugs and will act as a foundation for innovation and new industry. For libraries, it means that the researchers we support will be able to fully realise the value of our growing

---

[1] e.g. 'Rainfall or temperature measurements, mortality rates, population numbers, currency values, chemical structures, historical facts and dates, the number of Twitter followers someone has' (from: Simon Fraser University Library, https://www.lib.sfu.ca/help/academic-integrity/copyright/data-copyright)

[2] Liber. Text and Data Mining. The need for change in Europe. https://libereurope.eu/wp-content/uploads/2014/11/Liber-TDM-Factsheet-v2.pdf

collections of scientific content. This will, in turn, ensure a more rigorous approach to research, including more through reviews of the literature.[3]

Other possible research applications of TDM include mining newspapers to find textual indicators of economic uncertainty, political shifts, or social trends, and mining of large-scale library catalogues, other online knowledge repositories, or social media aggregations to understand changes in technologies, publishing, consumer behaviours and so forth.[4]

Researchers are increasingly using TDM and text analytics across several computational research methods, and curating the output of this work has the potential to result in novel discoveries within and across diverse bodies of research. The Portage data curation network supports researchers in the development of reproducible research findings in the application of TDM.

In the US, recent court decisions have recognized the "solid legal basis for non-consumptive research on copyrighted materials." As a consequence, organizations such as the HathiTrust Research Center now provide "access to the text of the complete 16.7-million-item HathiTrust corpus for non-consumptive research, such as data mining and computational analysis, including items protected by copyright." This policy is based on the premise that non-consumptive research use, such as TDM, does not impinge upon or "change the legal status of items protected under copyright."[5]

There are two possible solutions for resolving the issues that copyright causes for machine learning and other big data research applications. The first solution would be to emulate the U.S. fair-use model by making the current list of fair dealing purposes illustrative rather than exhaustive. The second solution would be to create a specific exception for text and data mining or computer informational analysis.[6] While Canadian exceptions like fair dealing may already permit TDM, and its permission can be added into library purchase contracts, a TDM-specific exception would provide

---

[3] Liber. Text and Data Mining. The need for change in 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, Serials Review, 00987913, Sep2014, Vol. 40, Issue 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. 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 https://www.theglobeandmail.com/report-on-business/rob-commentary/why-copyright-law-poses-a-barrier-to-canadian-ai-ambitions/article35019241/

clarity and could give Canadian researchers an advantage, driving new discoveries and innovation.

In addition, many sources that are crucial for TDM are databases with terms of use that are negotiated between libraries and publishers or between users and publishers. In fact, rights holders often use licenses "to override copyright exceptions that were created through transparent legislative processes, at the expense of users and at the cost of the spread of knowledge, discovery and innovation."[7] It must be made clear that any TDM exception cannot be waived or overridden by contract.

Finally, any new exception should not be limited to non-commercial or scientific research, as TDM has a variety of commercial and interdisciplinary applications.

Why do researchers think adding TDM to the Act is important?

A wide-ranging set of testimonials from Canadian researchers is provided at the end of this document. We provide two illustrative examples here:

> *"It's really challenging to carry out text and data mining on copyrighted materials in Canada, so much so that despite the demonstrated potential of being able to do this sort of large-scale analysis, very little work has been done. In many cases, it's simply too onerous – or seems too onerous – to navigate copyright. This is especially pronounced for graduate students and early career researchers: they may have the skills and the energy, but not the time and support to navigate the process. An exemption [for TDM] would open the door to transformational research and pave the way for a new generation of scholars to find new and exciting insights from the ever-increasing digital record."*
>
> Ian Milligan, Associate Professor, Department of History, University of Waterloo, September 2018.

> *"Imagine the following scenario: you go to the library and check out a book but you are NOT allowed to read it. You may look at it, and you may touch it, but reading it is against the law. That is essentially what disallowing text and data mining of books is like. We have all of this information that is legally stored in our libraries, but we are only allowed to understand it in one very old fashioned way. That's absurd. Text and data mining is just another form of reading. It's time to unlock the wealth of information stored in archives and libraries. Researchers are not trying to violate copyright. They are trying to study texts,*

[7] "CFLA Position Statement: Protecting Copyright Exceptions from Contract Override" (CFLA), accessed April 22, 2018 http://cfla-fcab.ca/wp-content/uploads/2018/02/CFLA-FCAB_statement_contract_override.pdf.

which is what they have always done. Only now this research is against the law. We are failing as a society to promote new knowledge."

Andrew Piper, Professor, Department of Languages, Literatures, and Cultures, and director of .txtLAB, McGill University, October 2018.

Recommendation: Amend the Copyright Act to allow text and data mining

Allowing TDM without rights-holder permission would provide clarity and could give Canadian researchers an advantage, driving new discoveries and innovation.

This goal can be accomplished by expanding fair dealing so that it is open to any purpose as it is in the US, or by creating a specific exception for TDM. It must be clear that this new exception is not limited to non-commercial or scientific research and that it cannot be overridden by contract.

About Portage

Portage is an initiative of the Canadian Association of Research Libraries (CARL), launched in 2015 to promote the shared stewardship of research data and to address specific policy, service, and infrastructure gaps in the national RDM ecosystem. Portage has made great headway in this, collaboratively developing and delivering a number of services and platforms designed to support Canadian researchers better manage, store, and share their data. More information about Portage is available here: https://portagenetwork.ca/about/

Further reading

Many organizations that represent researchers and the libraries that support researchers have issued statements or briefs related to TDM.

Association of European Research Libraries https://libereurope.eu/text-data-mining/

European Federation of Academies of Sciences and Humanities https://www.allea.org/wp-content/uploads/2017/11/PWGIPR_Statement_TDM_2017.pdf

Association of Research Libraries: http://www.arl.org/storage/documents/TDM-5JUNE2015.pdf

HathiTrust Research Center: HaithiTrust Research Center Extends Non-Consumptive Research Tools to Copyrighted Materials: Expanding Research through Fair Use: https://www.hathitrust.org/blogs/perspectives-from-hathitrust/hathitrust-research-center-extends-non-consumptive-research-tools

International Federation of Library Associations: https://libereurope.eu/blog/2015/06/09/liber-response-to-stm-statement-on-text-and-data-mining/

Appendix: Further testimonials from Canadian researchers

Geoffrey Rockwell, Professor of Philosophy and Humanities Computing, University of Alberta; Director of the Kule Institute for Advanced Study Arts

> *"It is vital that researchers be permitted to use non-consumptive methods on copyrighted materials if we are to be able to study contemporary culture using big data methods. To not let researchers do so is in effect not letting them read the materials with technological assistance. Can a researcher not use a computer to search a copyrighted e-book for a word? Can they not count the hits? Can they not count words on a page whether manually or with a computer? The important thing is that we are talking about non-consumptive methods from which one cannot reconstitute the text or be said to be "copying" the text. That should be the bar."*

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 analyse 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, Professor, School of English and Theatre Studies, University of Guelph

> *"Permitting the non-consumptive use of copyrighted materials without the permission of the copyright holder is crucial to the future of research in Canada. To take a particular case, the study of Canadian literature and culture using the newest computational techniques is virtually blocked, because the published cultural heritage of our relatively young nation is almost entirely covered by copyright. Current barriers to text and data mining not only set back research and training in these fields. They also deprive Canadians of the insight into our histories and the knowledge of our society that would be offered by the application of these exciting new methods."*