Opinion on the legality of JNU Data Depot

Dr Arul George Scaria

Context

I, Arul George Scaria, upon the request of Mr Carl Malamud and Prof (Dr) Andrew Lynn, have reviewed the copyright related aspects of the JNU Data Depot project. This opinion is provided without taking any monetary or non-monetary remunerations. This opinion is drafted with the sole intention of promoting research and innovations in India, by providing more clarity on the legality of a project that aims to use text and data mining (TDM) as a knowledge discovery tool.

I am furnishing this in my personal capacity, as a researcher who has been working in the area of intellectual property rights for a substantial period of time. I would like to clarify that the views provided herein does not reflect the position of my employer, National Law University, Delhi. I may also be publishing all or substantial parts of this legal opinion in different platforms, including blogs and journal articles.

Relevant experiences for providing this opinion

I'm an Assistant Professor of Law at National Law University, Delhi, since 2014. I'm also a co-director of the Centre for Innovation, Intellectual Property and Competition (CIIPC). I'm also also an Affiliate Faculty of the CopyrightX program, which is a course offered each year from January to May under the auspices of Harvard Law School, HarvardX distance-learning initiative, and Berkman Center for Internet and Society, Harvard University. I did my doctoral research (2008 October – 2011 December) in the area of copyright law at the International Max Planck Research School for Competition and Innovation (IMPRS-CI), which was an interdisciplinary doctoral program jointly offered by the Max Planck Institute for Innovation and Competition (Munich) and the Ludwig Maximilians Universität (Munich). For my doctoral research, I had received the IMPRS-CI Scholarship from the Max Planck Institute and I was awarded doctoral degree with *summa cum laude* ('with the highest distinction'). My post-doctoral research (2012 January to 2014 February) at the Catholic University of Louvain (UCL), Belgium, focused on the issue of open access to large scale research data and I was part of different EU funded projects. I have also worked as a CSIR-NIF Fellow in the IP management division of the National Innovation

Foundation of India (2007 February to 2008 September), an autonomous organisation under the Department of Science and Technology, Government of India.

I have two sole-authored books in the area of intellectual property rights. My first book, *Ambush Marketing: Game within a Game*, was published by the Oxford University Press in 2008. My second book, *Piracy in the Indian Film Industry: Copyright and Cultural Consonance*, was published by the Cambridge University Press in 2014. I have also contributed to different international journals by drawing from my research. Detailed list of my publications can be found on my webpage (http://ciipc.org/people/co-directors/arul-george-scaria/). I have also made presentations on diverse IP related issues in different international forums like the World Intellectual Property Organisation (WIPO) and the European Commission (EC), apart from speaking at different international conferences. I was also a member of the advisory committee constituted by the National Council of Educational Research and Training (NCERT), India, in 2014 for addressing the copyright license issues for the open e-textbooks project of NCERT. I'm currently a member of the IPR Expert Group constituted by Department of Science and Technology, Government of Rajasthan.

Primary focus of this legal opinion

The primary legal question which this opinion addresses is whether the researchers are violating copyright law when they engage in TDM at the facility referred to as 'JNU data depot'? This question is relevant as at least some of the articles in the JNU data depot ("data depot") may still be under copyright protection and it is probable that they might have been included in the database without permission from the copyright owners. India is yet to see any specific litigations with regard to TDM and this note aims to share my perspectives on the legality of the facility in question.

What is TDM?

As defined by the UK IPO, text and data mining refers to "use of automated analytical techniques to analyse text and data for patterns, trends and other useful information."[1] TDM has enormous potential in knowledge discovery and some of the fields that have already witnessed the enormous

[1] https://www.gov.uk/guidance/exceptions-to-copyright#text-and-data-mining-for-non-commercial-research

potential of TDM are biomedical sciences, linguistics, and machine learning.[2] It's important to recogonise that the application and potential of TDM is not limited to any specific fields and the potential fields of application include law. As some scholars have pointed out, TDM can be conceptualised as happening in four stages – access, extraction, mining, and use.[3]

Some relevant aspects of the data depot project

Following four factual aspects are relevant, while considering the legality of the data depot.

1. As is evident from different documents relating to data depot, certain specific access related restrictions have been imposed on the users of the data depot. For example, though the facility will have access to 73 million journal articles, no one will be allowed to read or download those works from that facility. The only thing that the facility permits is use of computer software to crawl over the text and data without allowing those users to actually "read" the works.
2. The users have to physically visit the facility to use it.
3. The researchers are currently allowed to use the facility only for non-commercial research purposes.
4. The terms and conditions for the use of the depot are very similar to the conditions used by the HathiTrust.

Copyright law and the concerned TDM activities

In most of the TDM projects, at least some of the contents used for mining might be the ones which are still under copyright protection. This could include journal articles, photographs, or even sound/ video clippings. While some of the works might have already entered the public domain and some of them might be open access works, it is very much probable that some are works wherein the copyright owners have reserved all their rights and are still under copyright protection. In such a scenario, researchers are often worried as to whether they will be infringing the copyrights in those works when they engage in TDM. The explicit demand for TDM licenses from the side of some publishers have increased the anxieties among researchers and institutions.

[2] For an interesting discussion on benefits of TDM, see https://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining.
[3] Matthew Sag, 'The New Legal Landscape for Text Mining and Machine Learning', 47, available online at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3331606.

Recogonising the potential of TDM in research and innovations, different countries are creating exceptions for TDM through different approaches.

For example, the UK included a specific exception (Sec. 29A) under the Copyright, Designs and Patents Act 1988, allowing TDM for non-commercial purposes. But the UK exception provision can be availed only when the user has lawful access to the concerned copyrighted material (for example, having subscription to the concerned journals). The recent EU directive on copyright in the single market (Directive 2019/ 790) has also included two specific TDM related provisions. Art. 3 of the Directive mandates the member states to provide exceptions for reproductions and extractions made by research organisations and cultural heritage institutions for the purposes of "scientific research". As one may notice from the manner in which the directive has defined the term 'research organisation', this exception provision will benefit public funded, not-for-profit entities like universities (including university libraries), research institutes, and other entities whose primary goal is to conduct scientific research or to carry out educational activities involving scientific research. Similar to the UK exception, this exception provision can be used only with regard to works to which they have lawful access. Art. 4 of the Directive also mandates member states to provide exceptions for TDM and unlike the exception provided under Art. 3, there are no purpose related restrictions in Art. 4. It can be used by for-profit organisations and independent (unaffiliated) researchers. While Art.4 might give the impression of a broad exception in the first reading, it is in effect very narrow in scope, as it has provided an opt-out system for right holders. According to the provision, if a right holder has explicitly reserved her rights, the exception will not be applicable. To assess the practical effectiveness of both Art. 3 and Art. 4 of the DSM directive, one may have to wait for the implementation of those provisions in domestic laws of EU member states. It is also hoped that those domestic laws and the Court of Justice of the European Union (CJEU) might give more clarity on the scope of different terms used in the exception provisions, including 'scientific research'.

The country considered as the most favourable jurisdiction for TDM activities might be USA. Though it doesn't have any specific exceptions for TDM, the fair use exception under the US copyright system is broad enough to cover most TDM activities, including those for commercial purposes. The fair use exception under US copyright law is not limited to any specific purposes and a judge will have to decide on the fair use argument, based on certain important factors evolved through case-laws. The most prominent factors in this regard are - (1) purpose and character of use; (2) nature of the copyrighted work; (3) amount and substantiality of the portion taken; and (4) impact on the potential market. These are inter-related factors and a review of the recent fair use cases suggest that one of the sub-factors of the first factor, i.e., whether the use was a

transformative use, can have substantial impact on analysis in most factors and final outcome in a litigation. The more transformative a use is, the higher the likelihood of a court reaching a finding of fair use.

As Prof. Matthew Sag has pointed out in a recent article published in the Journal of the Copyright Society of the USA, in general, the use of a copyrighted work can be classified as expressive use or non-expressive use, depending on the purpose of the use.[4] When we use a copyrighted work for appreciating the expressive aspects of that work, it is referred to as expressive use.[5] The examples cited by him in this regard are downloading a movie to watch it or making photocopies of texts for reading them. On the other hand, non-expressive uses generally refer to acts of "reproduction that is not intended to enable human enjoyment, appreciation, or comprehension of the copied expression as expression".[6] TDM is a good example for non-expressive use, as the purpose of TDM is not to read those individual articles, but to discern through automated analysis of the combined data certain information such as patterns, trends, or correlations. As Prof. Sag points out with the help of landmark cases such as Google Books case and the HathiTrust case, copying expressive works for non-expressive uses is generally considered by courts as fair use.[7]

As indicated above, one of the important aspects of the US approach in this area is that even TDM for commercial purposes might be well within the fair use exception. While the question of whether the use in question was for a commercial purpose might have some relevance for analysis under the fourth factor (impact on the potential market of the plaintiff), the courts have held that there cannot be any presumption against a finding of fair use just because the use in question was commercial in character. For example, in the Google Books case, Google was engaging in a commercial activity and it didn't prevent the court from reaching a conclusion of fair use. In a nut-shell, the current fair use system might be providing sufficient protection for researchers in the US from any copyright infringement liability, when they engage in TDM.

Legal position in India

In my view, TDM activities like the ones provided by the data depot, should be considered as legal in India on three different, but inter-related, grounds which are explained below.

[4] Ibid., 9.
[5] Ibid.
[6] Ibid.
[7] Ibid., 32.

Firstly, it is a basic principle of copyright law that there cannot be any copyright in facts or ideas. Copyright law only protects the expression of ideas. When a researcher is using a copyrighted article (or for that matter any other copyrighted work) for the purpose of TDM, they are using the work only as data or large sets of data, which are clearly outside copyright protection. As Prof. Sag has pointed out, the use is clearly non-expressive use of an expressive work and the copyright holder do not have any legal or moral grounds to prevent such uses.

Secondly, if facts and ideas are clearly outside copyright protection, it is morally and legally impermissible to allow publishers to use End User License Agreements (EULA) or any other contractual tools to restrict researchers from using articles for TDM. Courts should take a strong position against such contracts which are trying to create property rights over non-protectable subject matter.

Thirdly, it is important to remember that copyright law is not just about the rights of creators of works, but also of users of those works. As rightly explained by Justice Endlaw in *The Chancellor, Masters & Scholars of the University of Oxford and others v. Rameshwari Photocopy Services and another* (Delhi High Court, Single Bench decision, September 2016), "...the rights of persons mentioned in Section 52 are to be interpreted following the same rules as the rights of a copyright owner and are not to be read narrowly or strictly or so as not to reduce the ambit of Section 51, as is the rule of interpretation of statutes in relation to provisos or exceptions...".[8] In other words, one has to treat the rights given to the users at par with those given to creators. This position is very much in recognition of the fact that knowledge creation is a cumulative process and access to existing knowledge is important for creation of future knowledge. Hence one has to also look at the exceptions provided under Indian copyright law, before jumping into any conclusion as to whether the TDM activity in question is illegal under Indian copyright law.

If one looks at the exceptions in India, it can be seen that the country follows a hybrid system of exceptions wherein a relatively broad fair dealing exception is complemented with a long list of specific, enumerated exceptions.[9] As many readers are aware of, the often cited distinction between the fair use system and the fair dealing system is that the fair dealing exception is limited to the specific purposes mentioned in the provision, whereas the fair use system is not limited to any specific purposes and it can be applied for a broader range of activities. However, if we look at the evolving jurisprudence in this area (particularly cases from countries like Canada which has a fair

[8] The Chancellor, Masters & Scholars of the University of Oxford and others v. Rameshwari Photocopy Services, https://indiankanoon.org/doc/135895592/, para 41.
[9] See Sec. 52 of Copyright Act 1957.

dealing provision similar to India), it can be seen that the fair dealing provision can also be a dynamic tool that can respond adequately to the developments in technology and changing user requirements.

There are three essential steps in a fair dealing analysis. First question a court may ask is whether there was a dealing. As long as the plaintiff can show that the defendant has made use of her work, this requirement would be met.[10] In the context of the data depot, it is certainly evident that researchers would be "using" some copyrighted materials, even though as explained earlier, it is good to remind ourselves that the use is just a non-expressive use of an expressive work. In the second step, the court may look into the question of whether the use was for a purpose specifically mentioned under the provision and an objective analysis is required in this regard.[11] The Indian fair dealing provision specifically includes private or personal uses, including research, as a purpose for which the fair dealing provision is applicable. The access restrictions explicitly put in by the data depot indicates that the users of the facility will be using it only for research purposes, and only in their personal capacity. As the activity in question is for a purpose specifically mentioned in the provision, the second requirement is also met. The third and the final step involved in a fair dealing analysis is to ask whether the dealing was "fair". While the copyright statute does not define the term "fair", different judgments have provided some guidelines for analysing "fairness". The most famous among them is the observations of Lord Denning in *Hubbard v. Vosper*[12]:

"It is impossible to define what is "fair dealing". It must be a question of degree. You must first consider the number and extent of the quotations and extracts. Are they altogether too many and too long to be fair? Then you must consider the use made of them. If they are used as a basis for comment, criticism or review, that may be fair dealing. If they are used to convey the same information as the author, for a rival purpose, that may be unfair. Next, you must consider the proportions. To take long extracts and attach short comments may be unfair. But, short extracts and long comments may be fair. Other considerations may come to mind also. But, after all is said and done, it must be a matter of impression."[13]

As is evident from the observations of Lord Denning, fairness is a question of degree and impression. Another landmark decision that has provided a better analytical framework for fairness analysis is the decision of the Canadian Supreme Court in *CCH Canadian Ltd. v. Law Society of Upper*

[10] Lionel Bently and Brad Sherman, *Intellectual Property Law* (4th Edn, Oxford University Press, 2014), 224.
[11] Ibid.
[12] Hubbard v Vosper – [1972]2 QB 84
[13] *Ibid.*, 94

Canada.[14] The Canadian Supreme Court highlighted six factors that might be taken into consideration in the fairness analysis and it might be a good idea to analyse the "fairness" of the data depot in light of those six factors.

(1) Purpose of the dealing - According to the court, "purpose of the dealing will be fair if it is for one of the allowable purposes under the Copyright Act". It is beyond doubt that the TDM facility at JNU is intended only for research purposes and it is a purpose clearly permitted under Indian copyright law.

(2) Character of the dealing - Under this factor, the Court suggests that we must examine how the works were dealt with. According to the Court, "[i]f multiple copies of works are being widely distributed, this will tend to be unfair. If, however, a single copy of a work is used for a specific legitimate purpose, then it may be easier to conclude that it was a fair dealing. If the copy of the work is destroyed after it is used for its specific intended purpose, this may also favour a finding of fairness. It may be relevant to consider the custom or practice in a particular trade or industry to determine whether or not the character of the dealing is fair." When we analyse this factor in the context of the data depot, it can be seen that the copying of the articles has been done with a very specific purpose – enabling research and knowledge discovery through TDM. TDM cannot be done in the absence of access to copyrighted works and extraction of data from those works. Hence it is evident that this factor will also be in favour of the data depot.

(3) Amount of the dealing – According to CCH Canadian decision, the amount of the dealing and importance of the work allegedly infringed should be considered in assessing fairness. But the Court itself has pointed out that the relevance of the amount taken would vary depending on the purpose. The Court has pointed out that for some purposes such as research or private study, it may be essential to copy an entire academic article. In the context of TDM, it is important to remind ourselves that copying of the full text is very relevant for effective TDM and a court should not rule this factor against the data depot, just because the complete text in a copyrighted work has been copied for the purpose of TDM.

(4) Alternatives to the dealing - Under this heading, the CCH decision has highlighted the importance of exploring alternatives to dealing with the infringed work. In case of TDM, it is a fact that the level of access to contents will determine the quality of the outputs. While some publishers might try to point out their licensing schemes as an alternate, it is not a viable option to take licenses from thousands (if not millions) of copyright owners merely for the purpose of TDM.

[14] CCH Canadian Ltd. v. Law Society of Upper Canada, 2004 SCC 13.

We should also remind ourselves that those licenses are in effect trying to extract money on non-protected subject matter (data) under copyright law. Hence the analysis under this factor also stands in favour of data depot.

(5) Nature of the work - According to CCH Canadian decision, the nature of the work in question should also be considered by courts while assessing whether a dealing is fair. Some of the questions that the court may take into consideration is whether the plaintiff's work was an unpublished one and whether the work was confidential. While publication of unpublished works is considered as fair by the court, publication of a confidential work may drive the court more in favour of a finding that the dealing was unfair. In the context of the contents currently available in the data depot, it is evident that the TDM is relating to published works (and not confidential works) and hence this factor will also be in favour of the data depot.

(6) Effect of the dealing on the work – CCH Canadian decision has also suggested looking at the effect of the dealing on the work, as a factor in the fairness analysis. In this regard, the Court will be primarily looking at the question of whether the reproduced work is likely to compete with the market of the original work. As is clear from the description of TDM technology, results of TDM (new discoveries) are not competing with the market of copyright holders (journal articles). Hence there are no market displacements happening through TDM. Hence this factor will also be in favour of the data depot.

CCH Canadian decision had also very specifically pointed out that these six factors are not exhaustive and the relevance of the factors would depend on the factual context. Hence we may take into consideration one more question during a fairness analysis – is it fair to allow copyright holders to prevent an activity that is clearly outside the scope of rights provided to them under copyright law? As discussed earlier, TDM involves only non-expressive uses of an expressive work and we should not extend copyright protection to the ideas and facts behind an expressive work. All these factors should lead to a ruling in favour of the data depot, in the event of a litigation from the side of copyright holders.

Conclusion

To summarise, from a legal and policy perspective, I am of the view that the JNU data depot in its present form is not violating any provisions of Indian copyright law. I consider the collaborative TDM facility at JNU as a bold and innovative approach to unlock knowledge discovery. The

current copyright law in India provides sufficient safeguards for TDM and copyright related threats should not be allowed to act as a hindrance in the progress of science.

August 20, 2019

Dr Arul George Scaria