THE CONVERSATION
Academic rigour, journalistic flair

Data mining: why the EU's proposed copyright measures get it wrong

May 24, 2018 8.35am BST

Shutterstock

Data mining: why the EU's proposed copyright measures get it wrong

May 24, 2018 8.35am BST

Authors

Martin Kretschmer
Professor of Intellectual Property Law, University of Glasgow

Thomas Margoni
Senior Lecturer in Intellectual Property and Internet Law, University of Glasgow

Data that is mined with the help of machine learning techniques has been a rapid area of technological advancement – with good and bad consequences for everyone. And EU copyright law is currently caught in the crossfire.

Cambridge Analytica and Facebook's recent data scandal, which involved the profiling of users from their online behaviour facilitated by social networks, brought important issues to the surface about web privacy, only after it was reported that millions of people had their data harvested and improperly shared with a political consultancy.

But the same data mining technique also offers great societal benefit in fields such as traffic prediction, natural language processing and the identification of potential cures for diseases.

Many people think that regulating the use of data is a matter of data protection or privacy laws. However, where the raw material subjected to analysis is not "personal data" but material protected under copyright law, such as texts or certain structured databases, another set of legal norms come into play. This has far reaching and little understood consequences.

Fit for the digital age?

The proposal for a Digital Single Market Copyright Directive, which is currently at a critical stage in the European Parliament and in the Council of the EU, contains a number of provisions intended to modernise the bloc's copyright law, making it – in EU jargon – "fit for the digital age".

One of the provisions relates to a mandatory text and data mining (TDM) exception that would benefit research organisations acting for research purposes, for example a university researcher scanning and analysing scientific articles.

At first sight, permitting such uses appears to be a progressive idea, because copyright law exists to protect new expressions, and by doing so, to promote creativity. However, in our view, the debate around text and data mining and the need for a dedicated copyright exception has been focusing on the wrong premise.

Why should text and data mining of copyright protected materials be a copyright matter at all? There would be no need for a copyright exception if the labelling of TDM as a copyright infringement was removed.

TDM refers to the use of ideas, principles, facts and correlations contained in literary works, other types of texts or in datasets. Put simply, TDM techniques do not use copyright works as works, they access the information stored in them.

The informational value in texts is not, and should not, be protected by copyright law.

> *Copyright protects original expressions. Its protection does not extend to ideas, procedures, methods of operation or mathematical concepts as such.*

This principle can be found not only in all major copyright traditions, but also in the treaties that establish the international copyright framework.

Pinball wizard

Under the proposed directive, here's how copyright law may treat text and data mining. While the idea of a young wizard studying magic with his fellow students is not protected by copyright law, the original expression rendered by JK Rowling with her series of Harry Potter novels is protected.

If a temporary reproduction of a Harry Potter book is made in order to produce unauthorised copies, this is clearly a copyright infringement. But if the temporary copy is made within the scope of a machine learning application to study natural language processing, only the informational value of Rowling's text is extracted. These text extracts are annotated with labels, such as named entities and sentiment tags, which are then processed, for example, to create or improve automatic translation tools.

Wizard works: Harry Potter books. Shutterstock

But the novel is not replicated as a novel, so such an extraction method should not constitute a copyright infringement.

It's far from clear whether the current EU legal framework recognises this, however. The Information Society Directive gives a broad definition to the right of reproduction and provides a narrowly defined "closed" list of exceptions. It means that, at present, most TDM activities in the EU are deemed copyright infringements, or their legal status is so uncertain that it is safer to assume a violation has occurred.

By contrast, the US generally considers text and data mining to be "fair use".

Informational value

It is important to test the wording of the exception proposed in Article 3 of the EU's Copyright Directive against the goal of making the informational value of copyright works available to all. In our view, it remains unsatisfactory in at least three respects:

1. It limits the availability of the exception to research institutions for research purposes, thereby excluding innovative firms or journalists;
2. It does not address the backdoor of preventing text and data mining for technical reasons, which allows publishers to force licensing opportunities;
3. It only covers the right of reproduction but not the rights of distribution or communication to the public.

Societies are better off with more knowledge not less. They will be more culturally attractive, economically competitive and technologically advanced, all of which leads to more innovation, more

jobs and well-informed citizens. Copyright law is an instrument to achieve this public policy goal.

We suggest that – in order to rebalance the relationship between investment and innovation – anyone who has lawful access to copyright materials, including public interest research organisations and businesses, should be able to conduct text and data mining.

Read more:* *Germany's legal crackdown on social media: four misconceptions dispelled

---

Sustained lobby pressure from publishers is moving fast against this proposal, which is supported by a planned amendment of the draft report in the European Parliament.

EU copyright law really needs an open exception that is responsive to technological development, similar to those already available in many other jurisdictions that permit TDM activities. But it seems that legislators in the EU are unable to offer even modest certainty.

Intellectual property **Copyright** **Privacy law** **Machine learning** **Data mining** **copyright law**