The New Copyright Directive: Text and Data Mining (Articles 3 and 4)

Kluwer Copyright Blog
July 24, 2019

Bernt Hugenholtz (Institute for Information Law (IViR))

Please refer to this post as: Bernt Hugenholtz, 'The New Copyright Directive: Text and Data Mining (Articles 3 and 4)', Kluwer Copyright Blog, July 24 2019, http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/

Art. 2(2) of the DSM Directive defines 'text and data mining' as "any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations". Text and data mining (TDM) generally refers to the computer-based analysis of large bodies of data in order to gain knowledge.

With the increasing powers of computer processing and the omnipresence of vast amounts of minable text and data on the Internet, TDM has become a hugely important research tool in science and many other domains. For example, in linguistics TDM can be used to analyze large bodies of text to extract syntactic or grammatical patterns. TDM is also applied in numerous other scientific domains, ranging from astronomy to musicology to the social sciences. Outside the world of science proper, data mining plays a growing role in the arts, as the spectacular Next Rembrandt project that produced an amazing Rembrandt-like portrait based on mining Rembrandt's oeuvre, reveals.

In the industrial and commercial realm TDM has become even more pervasive. Text and data mining is nowadays standard practice in pharmaceutical research, journalism, information retrieval, search, and consumer information – to name just a few areas. TDM is also an essential tool in developing intelligent applications that require vast volumes of raw text and data to 'self-learn' complex tasks such as translation or speech recognition. Much of the current and future development in artificial intelligence, therefore, depends on TDM.

With the InfoSoc Directive's all-inclusive reproduction right extending to every "direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part" (Art. 2 InfoSoc Directive), TDM creates potential conflicts with copyright law when copyright protected content is mined. Similarly, the sui generis database right may also be implicated. Whereas some scholars have advocated a normative interpretation of the reproduction right, which would restrict its scope to exploitative uses "of the work as the work", and would rule out non-exploitative uses such as mining, the EU legislature assumed that an express TDM exception was required for reasons of legal certainty (see EC, Impact Assessment, p. 104-105; see recital 8 DSM).

While the need to ensure broad TDM freedoms in science was always self-evident, the EU's approach towards text and data mining in the commercial realm is more ambivalent. The European Commission's draft DSM Directive merely proposed a mandatory TDM exception for the benefit of non-commercial research organizations and cultural heritage institutions. During the discussions in the Council Working Group, following a proposal by the Dutch delegation, an optional exception was added that conditionally permitted TDM for commercial purposes. In the end, the European Parliament gave mandatory status to that exception as well. Consequently, the DSM Directive now comprises two obligatory DSM provisions: Articles 3 and 4. The two exceptions are not, however, equally robust.

Art. 3 exempts acts of reproduction and extraction committed by "research organisations and cultural heritage institutions". Art. 2(3) defines a 'cultural heritage institution' as "a publicly accessible library or museum, an archive or a film or audio heritage institution". According to Art. 2(1), a 'research organisation' is either a not-for-profit entity or an entity tasked by a Member State with a public service research mission. Public broadcasting organizations and commercial research institutes, for example, are therefore excluded from the scope of Art. 3, but might still find solace in Art. 4.

In addition to permitting mining activities per se, Art. 3(2) allows the secure storage and retention of copies of mined works and other subject matter "for the purposes of scientific research, including for the verification of research results". This is important because empirical scientific research generally requires research data to remain available for corroboration purposes. Nevertheless, Art. 3 permits TDM only in respect of works or other subject matter (e.g. databases) to which beneficiary organizations "have lawful access". According to Recital 14, 'lawful access' covers access to content pursuant to contractual arrangements (e.g. subscriptions or open access licenses), as well as to "content that is freely available online". The requirement of 'lawful access' does not however imply that rightholders may contractually rule out text and data mining in their terms of agreement. Article 7 expressly provides that any contractual provision contrary to Article 5 is unenforceable. Note as well that the option to 'opt out' out of the TDM exemption is provided only in respect of the non-research uses governed by Art. 4.

Rightholders do remain free to "apply measures to ensure the security and integrity of the networks and databases where the works or other subject matter are hosted. Such measures shall not go beyond what is necessary to achieve that objective" (Art. 3(3)). As Recital 16 clarifies, such measures are justifiable only for reasons of systems or database security and integrity (e.g. "in view of a potentially high number of access requests to, and downloads of, their works or other subject matter"), not for purely commercial reasons. Given the limited potential harm (if any) that right holders might suffer from this exemption, Member States need not provide for financial compensation (Recital 17).

The Directive's second TDM exception encompasses a much broader class of users, but is considerably narrower in scope. Art. 4(1) generally allows acts of reproduction and extraction "for the purposes of text and data mining", and reproductions and extractions may be retained for the same purpose. The provision thus permits TDM for all imaginable purposes, regardless of any underlying commercial motive. Art. 4 does, however, allow rightholders to opt out of the exemption. Art. 4(3) applies only on condition that right holders have not expressly reserved their rights "in an appropriate manner, such as machine-readable means in the case of content made publicly available online". According to Recital 18, "it should only be considered appropriate to reserve those rights by the use of machine-readable means, including metadata and terms and conditions of a website or a service. […] In other cases, it can be appropriate to reserve the rights by other means, such as contractual agreements or a unilateral declaration." In other words, Art. 4 right holders may effectively prohibit text and data mining for commercial uses by adding robot.txt type metadata to their content online.

As a result, the Directive effectively creates and legitimizes a derivative market for text and data mining, which right holders may wish to control, license or even entirely prohibit. While most content owners will have no incentive to prohibit or monetize data mining, some right holders will. Scientific publishers, for example, are well aware that their publishing portfolios have informational value beyond the published articles they have aggregated. Indeed, some publishers already offer paid-for text and data mining as value-added services, and will be reluctant to grant TDM licenses to third parties. Other publishers are still in the process of developing licensing strategies to capitalize on this emerging market (see report here).

In conclusion, the TDM provisions of the DSM Directive secure considerably less freedom to text and data mine than they initially appear to do. The opt-out clause of Art. 4, in particular, leaves for-profit miners in the EU at the mercy of the content owners. This puts AI developers, journalists, commercial research labs, and other innovators at a competitive disadvantage in comparison with the United States, where text and data mining is deemed fair use, even if it is done for profit. One may wonder if innovation in Europe would not have been better served without any of the TDM exceptions in the new Directive.

————————————————————————————————

This post is part of a series on the new Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market (CDSM Directive):

The New Copyright Directive: A *tour d'horizon* – Part I by João Pedro Quintais

The New Copyright Directive: A tour d'horizon – Part II (of press publishers, upload filters and the real value gap) by João Pedro Quintais

The New Copyright Directive: Digital and Cross-border Teaching Exception (Article 5) by Bernd Justin Jütte

The New Copyright Directive: Collective licensing as a way to strike a fair balance between creator and user interests in copyright legislation (Article 12) by Johan Axhamn

The New Copyright Directive: Article 14 or when the Public Domain Enters the New Copyright Directive by Alexandra Giannopoulou

The New Copyright Directive: Fair remuneration in exploitation contracts of authors and performers – Part 1, Articles 18 and 19 by Ananay Aguilar