Kluwer Copyright Blog

Limitations to Text and Data Mining and Consumer Empowerment. Making the Case for a Right to “Machine Legibility”

Rossana Ducato, Alain Strowel (Université Catholique de Louvain) · Tuesday, March 19th, 2019

The use of artificial intelligence (AI) tools raises possible issues of bias, discrimination and transparency that need to be investigated by (legal) researchers. But AI tools can also support the implementation of legal principles and rules. This is the case with smart disclosure systems (SDSs). The latter refers “to the timely release of complex information and data in standardized, machine readable formats in ways that enable consumers to make informed decisions”. Smart disclosure systems allow users to have easy and timely access to the relevant pre-contractual information or even receive personalised advice based on their preferences**.** Over recent years, new providers have proposed services to automatically analyse websites’ contractual documents and check their compliance with the applicable consumer and data protection framework (www.usableprivacy.org; https://pribot.org/polisis; https://claudette.eui.eu; http://www.rosels.eu/research/research-project-iop/). One of the main goals of these projects is to increase the awareness of users of the rights, obligations and possible risks in their online transactions, and to mitigate the consequences of the well-known signing-without-reading process. Such tools are primarily directed at consumers, as end-users of the service; however, other potential users are consumer associations or regulatory authorities, which can use them to perform periodical assessments, start investigations or verify complaints more efficiently.

The functioning of SDSs is based on text and data mining (TDM). TDM uses techniques from natural language processing, machine learning, information retrieval, and knowledge management for the automated analysis of digital content (structured and unstructured data), in order to extract information, identify patterns, and discover new trends, insights or correlations. In the proposal for a Directive on copyright in the Digital Single Market, TDM is defined as "any automated analytical technique aiming to analyse text and data in digital form in order to generate information such as patterns, trends and correlations", and is analysed in more detail here.

Despite the development of such promising tools for enhancing the readability and understandability of the conundrum of terms, the current EU legal framework is not generally supportive of TDM. Nor will the TDM-related provisions laid down in the proposal for a Directive on Copyright in the Digital Single Market be particularly effective in the case of SDSs.

The proposal, which is expected to be voted on during the March plenary, includes two TDM exceptions: art. 3 introduces a mandatory exception, not overridable by contract, which would allow research organisations and cultural heritage institutions to text and data mine works to which they have lawful access for the purposes of scientific research; while the second exception (art. 3a), also mandatory, will provide for an exception to make temporary reproductions and extractions of content during a TDM process, if such use "has not been expressly reserved" by the rightholder "including by technical means".

Despite the advantage of providing some legal certainty, those two exceptions suffer from several limitations, as highlighted by copyright scholars (see, for example, here, here, here, here and here). Art. 3 is narrow in scope and does not provide effective protection against technological protection measures (TPMs); while the exception at art. 3a, even if not limited to research purposes, can be put out of play by both contracts and TPMs.

The proposed exceptions are still not enough for embracing the potential of TDM and will exclude many useful applications, including SDSs. The problem is far from being a hypothetical one: the results of empirical research we conducted in 2018 show that the terms and conditions (T&C) of a representative set of platforms operating in the sharing economy most of the time ban the practice of TDM. In particular, 20 out of 21 platforms published the T&C on their website and 14 of them contained specific intellectual property clauses, directly or indirectly related to TDM activities. More specifically:

- four platforms expressly prohibited TDM on the website content;
- three others did not allow the use of any kind of bot, crawler or scraper (i.e., the automated software agents that search through the content of webpages. These are necessary tools for TDM, e.g. when there is no available application programming interface);
- in four cases, the reproduction or copying of website materials – which is usually a preliminary step in the TDM process – was forbidden; and
- in three instances, the formulation was vague or broad enough to exclude TDM.

None of those provisions was clearly limiting the TDM prohibition to data related to the service (e.g. the timetable of flights, the prices, the list of accommodation and contact details of the owners, etc.) so as to avoid free riding from mala fide users. The prohibition or impossibility of mining the text of T&C and privacy policies arguably was not intentional and only a side effect of the general TDM prohibition.

However, such an interpretation can be challenged if we look at the technical instructions embedded in the same websites. Running the robots.txt file, our study demonstrated that the contractual provisions were usually covered by the instructions, and that in one case the protocol excluded the crawling of a directory containing the legal documents, while in two other cases it specifically blocked access to the pages of the T&C and privacy policy.

With regard to SDSs, the prohibition of TDM can dangerously undermine the legitimate activity of consumers, who could use such instruments for the automatic analysis of contractual documents to better understand the terms of the user agreement. If a new "bionic eye" is available to scan a document and extract the relevant pre-contractual information, is it justified to prohibit its use?

Even without an express TDM exception, there are two possible legal interpretations that can ensure the protection and ultimately the empowerment of consumers.

First, if we take into account the rationale of copyright protection we must consider the implicit requirement according to which the reproduction involves a use as a work (see Chapter 7). Such use as a work does not exist in the case of TDM, nor in other cases involving copying for deriving information or checking conduct (e.g., to identify plagiarism). As put by a U.S. Court in *Authors Guild v. Google, Inc.*, 'the purpose of Google's copying of the original copyrighted books is to make available significant information *about those books*, permitting a searcher to identify those that contain a word or term of interest' (emphasis added). This purposive analysis of the act of copying strongly weighs in favour of fair, because highly transformative, use. In the EU, the requirement of use as a work for a copyright infringement to happen can help to reach the same outcome. Indeed, when acts of reproduction are carried out for the purpose of search and TDM, the work is not used as a work, it only serves as a tool or data for deriving other relevant information. The expressive features of the work are not used, and there is no public to enjoy the work, as the work is only an input in a process for searching a corpus and identifying occurrences and possible trends or patterns. Therefore, anyone should be free to perform TDM.

Second, the absolute prohibition of mining the text of T&C and privacy policies is likely to conflict with another legal ground which is recognised in EU consumer and data protection law: the principle of legibility. The latter means that information must be legible and given in plain and intelligible language. As pointed out by Micklitz et al., such a principle implies the possibility to actually read the text of the contract, i.e. through the design of conditions "plainly both from an editing and optical point of view" (no "small-print" for instance). The same principle is enshrined in the General Data Protection Regulation, as interpreted by the European Data Protection Board in its Guidelines on Transparency.

In both the consumer and the data privacy laws, the principle of legibility pursues the same goal: the protection of the weak party against information asymmetries, by requiring that information must be visible, accessible and readable. If we interpret such a requirement in a technologically neutral way, it is possible to argue in favour of a right to "machine legibility". By machine legibility we mean the possibility for an SDS to have access to the precontractual information (T&C) and the information related to the processing (privacy policy) in a format processable by the smart system.

If, thanks to AI, there are now instruments enhancing the human ability to read and understand contractual terms and privacy settings, not only should information be legible to a human eye, but also to the tools that a user can take advantage of. Therefore, either depending on IPRs, contractual or TPM limitations, an absolute prohibition of TDM on pre-contractual and privacy information available online on the website of online platforms will unreasonably restrict a legitimate prerogative of the consumer or the data subject.

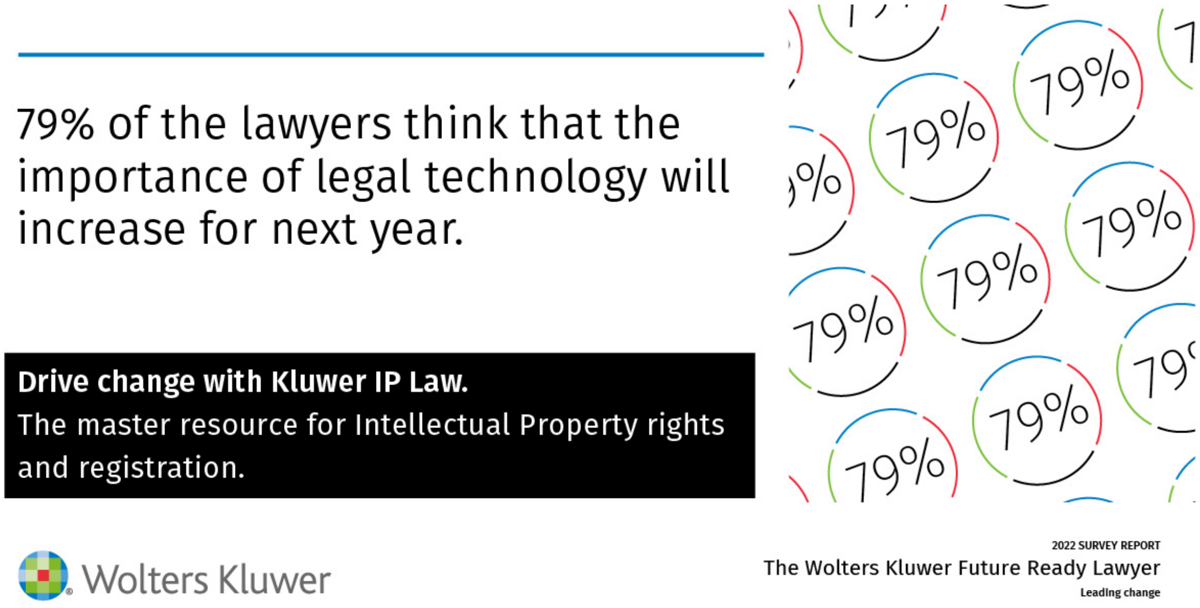
In conclusion, considering the rationale of copyright protection and the principle of legibility deriving from consumer and data protection legislation, there are at least two legal grounds to support the use of TDM for consumer empowerment.

As an alternative, we would equally welcome a well-designed TDM exception able to accommodate the requirement of machine legibility that appears necessary in a society where the automatic treatment of information becomes central. But we are not yet there.

An extended version of this article is available here: https://ssrn.com/abstract=3278901

The updated version of the paper is expected to be published this spring in IIC – International Review of Intellectual Property and Competition Law.

To make sure you do not miss out on regular updates from the Kluwer Copyright Blog, please subscribe here.


79% of the lawyers think that the importance of legal technology will increase for next year.
Drive change with Kluwer IP Law.
The master resource for Intellectual Property rights and registration.
79%
Wolters Kluwer
2022 SURVEY REPORT
The Wolters Kluwer Future Ready Lawyer
Leading change

This entry was posted on Tuesday, March 19th, 2019 at 3:49 pm and is filed under Digital Single Market, European Union, Exceptions and Limitations, Legislative process, Text and Data Mining (TDM)
You can follow any responses to this entry through the Comments (RSS) feed. You can leave a response, or trackback from your own site.