Originally published in: Skadina, Inguna/Eskevich, Maria (Eds.): CLARIN Annual Conference 2018, Proceedings. 8-10 October 2018, Pisa, Italy. – Utrecht: Clarin, 2018. Pp. 39-42.

New exceptions for Text and Data Mining and their possible impact on the CLARIN infrastructure

Pawel Kamocki
IDS Mannheim / Université Paris Descartes

Erik Ketzan
Birkbeck, University of London

Julia Wildgans
IDS Mannheim / Universität Mannheim

Andreas Witt
Universität zu Köln / IDS Mannheim / Universität Heidelberg

Abstract

The proposed paper discusses new exceptions for Text and Data Mining that have recently been adopted in some EU Member States, and probably will soon be adopted also at the EU level. These exceptions are of great significance for language scientists, as they exempt those who compile corpora from the obligation to obtain authorisation from rightholders. However, corpora compiled on the basis of such exceptions cannot be freely shared, which in a long run may have serious consequences for Open Science and the functioning of research infrastructure such as CLARIN ERIC.

1. Overview of the current system of statutory exceptions in European copyright

Copyright grants authors exclusive rights in relation to their works. In principle, every reproduction or communication to the public of copyright-protected material requires authorisation from the rightholder. Obviously, if applied strictly this could have a chilling effect on freedom of expression, art and research; this is particularly true in the digital environment, where every use of a work necessitates a reproduction (in the device's memory), while copying and worldwide sharing is cheap and instantaneous. In order to strike balance between the interests of rightholders and those of the public, legislators introduce statutory exceptions and limitations to exempt certain unauthorised uses from liability (exceptions) or to limit the scope of the rightholders' monopoly (limitations).

In the European Union, national legislators are not entirely free to adopt exceptions and limitations. Rather, the Directive 2001/29/EC of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (hereinafter: InfoSoc Directive) contains (in its art. 5) a limitative list of exceptions and limitations that can be adopted in the national laws of the Member States. Apart from one mandatory limitation (that enables the functioning of the Internet), national legislators are free to choose which exception they want to adopt in their legal systems. National implementations of each of these exceptions can be narrower than allowed by the Directive, but they cannot be broader.

2. New exceptions for Text and Data Mining in certain EU Member States

Text and Data Mining (or text/data analytics) is the process of deriving new information from unstructured data by means of computational analysis. Since the analysed material is necessarily reproduced in the process (even if these reproductions may be just temporary), mining, in order to be lawful, requires authorisation from rightholders. The necessity to adopt statutory exceptions for Text and Data

Mining, at least for research purposes, has been discussed at least since 2011, i.e. the publication of the Hargreaves review (Hargreaves, 2011). In 2013, a group on Text and Data Mining was created within the Stakeholder's Dialogue "Licenses for Europe" (European Commission, 2013). The academic community, unhappy with the adopted approach (focused on licensing rather than on statutory exceptions), largely withdrew from the process (LIBER, 2013). One of the key arguments in favour of a statutory TDM exception is the fact that TDM for research purposes is allowed under the 'fair use' doctrine in the US, or covered by statutory exceptions e.g. Japan and other non-European countries.

In 2014, the UK was the first EU country to adopt a statutory TDM exception. Section 29A of the Copyright, Designs and Patents act allows for making copies of works in order to "carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose". Such copies need to be accompanied by a sufficient acknowledgement (unless this is practically or otherwise impossible) and cannot be transferred or used for any other purpose. The exception is expressly non-overridable by contracts (a contractual clause that purports to restrict the allowed activities is unenforceable), but it only applies to those who have "lawful access" to a work. This latter requirement raises questions on whether this access should be expressly authorised (in a license), or simply not resulting from copyright infringement (in which case e.g. everyone with Internet access could mine openly available websites). There seems to be no clear answer to this question, even though, in our opinion, the second interpretation should prevail.

In 2016, France also introduced a TDM exception (art. L. 122-5, 10° of the French Intellectual Property Code), but its scope remains very unclear. It seems to allow mining of scientific articles for the purposes of non-commercial public research (i.e. research carried out at universities and publicly funded research institutions). Adopted just before presidential and parliamentary elections, the French regulation on TDM is marked by its formal imperfections which an implementing decree was supposed to clarify; unfortunately, a proposal for such a decree was rejected in 2017 (Langlais, 2017) and, to the best of our knowledge, no progress has been made since. Therefore, it seems that the French TDM law is reduced to dead letter.

A much bolder measure was taken by the German legislator in 2017. New §60d of the German Copyright Act (UrhG) which entered into force on 1 March 2018 allows reproductions of copyright-protected content in order to enable automatic analysis of a large number of works for non-commercial scientific research. Furthermore, it also allows necessary modifications of mined content (cf. §23 UrhG). Interestingly, the new law expressly uses the word "corpus". Such a "corpus" can be shared with a "specifically limited circle of persons" (presumably a research team, perhaps also multi-institutional). However, once the research is over, the corpus has to be deleted or transferred to a library or an archive for permanent storage. The new German exception is expressly non-overridable by contractual clauses (cf. §60g), which in practice means that all content openly available on the Internet can be freely mined, even if the terms of service prohibit such uses. On the other hand, the new law requires that flat-rate equitable remuneration be paid to a copyright collecting society for the allowed uses (VG Wort; cf. §60h UrhG).

It shall also be noted that in some countries, such as Poland, the implementation of the research exception seems broad enough to encompass data mining activities (in Poland: only those carried out in public research institutions, cf. art. 27 of the Polish Copyright Act). Other Member States, however, seem to lack a research exception exceeding private copying (e.g. Austria). This fragmentation is particularly troublesome from pan-European projects such as CLARIN. A greater degree of harmonisation, achievable only via an intervention at the EU level, seems urgent.

3. New exception for Text and Data Mining in the Digital Single Market Directive?

In September 2016, the European Commission proposed a draft for a new Directive of on copyright in the Digital Single Market (European Commission, 2016). Art. 3 of the draft proposes a mandatory (i. e. to be implemented in all the Member States) exception for reproductions and extractions "made by research organisations in order to carry out text and data mining (...) for the purposes of scientific research". Only public universities and research institutions can benefit from this exception; however, the exception is no longer limited to non-commercial activities, so public-private partnerships are also within its scope. Like in the UK, the text requires "lawful access" to mined material, which raises the exact same questions as those discussed above.

The proposed exception is, like in the UK and in Germany, non-overridable by contracts. However, it allows rightholders to implement technological protection measures (Digital Rights Management)

"to ensure the security and integrity of the networks and databases". Such measures, however, "shall not go beyond what is necessary to achieve this objective".

Many contrasting views on the proposal have been expressed during the discussions in the European Parliament. The Culture and Education Committee (CULT) advocates to a solution similar to the one adopted in Germany, requiring payment of equitable remuneration and deletion of the compiled corpus upon the completion of the project. Its draft also stipulates that "lawful access" to mined works has to "acquired", which seems to indicate that a license to use the content (for whatever purpose) is necessary, and that content available on the open Internet is not necessarily concerned by the exception (CULT, 2017). According to the Committee on the Internal Market and and Consumer Protection (IMCO), the beneficiaries of the exception shall not be limited to research organisations, and that mining should be allowed also for other purposes than scientific research (IMCO, 2017). The Industry, Research and Energy Committee (ITRE) took a similar position (ITRE, 2017). Arguably the most important of the Committees, the Committee on Legal Affairs (JURI) expressed a more nuanced opinion. On the one hand, JURI advocates that the exception should concern all users and purposes; on the other hand, it also advocates for a narrow interpretation of "lawful access". Research organisations, however, shall be allowed to mine databases of scientific publishers even if they do not meet the "lawful access" requirement. Furthermore, corpora mined for research purposes shall be stored securely in designated facilities and re-used only for the purposes of verification of results of the research (JURI, 2017).

On 25 May 2018, the European Council (under the Bulgarian presidency) published its version of the proposal (European Council, 2018), which contains three important modifications compared to the Commission's original document. Firstly, the beneficiaries of the mandatory TDM exception include (alongside "research organisations") also "cultural heritage institutions" (defined as publicly accessible libraries, museums and archives as well as film or audio heritage institutions). Secondly, the Council's version requires that the corpora used for TDM shall be stored "with an appropriate level of security" and not retained "for longer than necessary" (which may imply the necessity to delete them at the end of the research project). Thirdly, and perhaps most importantly, the Council's proposal adds art. 3a containing an optional exception for TDM, allowing Member States to adopt broad TDM exceptions, potentially covering all categories of beneficiaries and purposes; however, these non-mandatory exceptions can only apply if the users have lawful access to the mined works, and if the use for TDM purposes has not been expressly restricted by rightholders (via Digital Rights Management or simply by an appropriate notice). This changes the paradigm from "TDM only with permission" to "open for TDM by default", but does not really provide the users with means to mine content which its rightholder does not want to be mined.

The final report of the European Parliament's Committee on Legal Affairs (JURI, 2018), adopted on 29 June 2018 was partly inspired by the Council's proposal. JURI postulated that the beneficiaries of the TDM exception shall include research institutions, but also educational establishments and cultural heritage institutions, to the extent that they conduct scientific research the results of which are publicly accessible. Secondly, JURI also added an optional TDM exception, similar to the one proposed by the Council.

JURI's final report was rejected by the European Parliament during a plenary vote on 5 July 2018 (mostly because of other controversial provisions of the Directive). This means that the adoption process has been slowed down, but not completely interrupted. However, it is still impossible to predict the content of the soon-to-be-adopted TDM exception. The Directive will probably have to be implemented twelve months after its entry into force (as per — seemingly undisputed — art. 21 of the Commission's draft).

4. The possible impact of the new exceptions on CLARIN infrastructure

While language researchers will receive substantial benefits and some legal certainty from the new TDM exceptions, even if certain research activities are exempted from the rules of copyright, proper licensing is still necessary to efficiently and widely share the fruits of researchers' work. In this sense, paradoxically, the new exception can have negative consequences on infrastructures such as CLARIN ERIC. In a world where intellectual property rights are prima facie no longer a barrier to access content (because everything can be freely mined, at least for non-commercial research purposes), researchers have fewer incentives to care about proper licensing and sharing of their datasets and results (e.g. within research infrastructures) (Suber, 2012). This may in turn considerably reduce the "knowl-

edge commons" (i.e. immaterial resources that — due to proper licensing — can be freely accessed and re-used by anyone and for any purpose (Hess, Ostrom, 2006) and in a long run hamper the development of Open Science. In such circumstances, even if research activities freed from the requirement to obtain permission from rightholders can flourish, knowledge transfer, citizen science and user innovation (von Hippel, 2017) may paradoxically be more difficult. In order to avoid this, it is important to remember that even if certain research activities are exempted from the rules of copyright, proper licensing is still necessary to efficiently and widely share the fruits of researchers' work. An alternative incentive (other than removing access barriers to primary material) for contributing to knowledge commons shall perhaps be provided by policymakers and research funding agencies. CLARIN ERIC, who declared its dedication to the principles of Open Science, has an important role to play in guaranteeing that language science remains truly open not only for researchers, but for all citizens.

Reference

Hargreaves, I. (2011). "Digital Opportunity. A Review of Intellectual Property and Growth". available at: https://www.gov.uk/government/publications/digital-opportunity-review-of-intellectual-property-and-growth

European Commission (2013). "Licenses For Europe: Structured stakeholder dialogue 2013", available at: https://ec.europa.eu/licences-for-europe-dialogue/

LIBER (Association of European Research Libraries) (2013). "Stakeholders representing the research sector, SMEs and open access publishers withdraw from Licences for Europe", available at: https://libereurope.eu/blog/2013/05/24/stakeholders-representing-the-research-sector-smes-and-open-access-publishers-withdraw-from-licences-for-europe/

Langlais, P.-C. (2017). "L'exception Text & Data Mining sans décret d'application…", Sciences Communes, 10 May 2017, available at: https://scoms.hypotheses.org/category/data-mining

European Commission (2016). "Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market", available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52016PC0593

European Council (2018). Notice from Presidency to Delegations on the Proposal for a Directive of the European Commission and the Council on copyright in the Digital Single Market, 2016/0280 (COD), available at: http://www.consilium.europa.eu/media/35373/st09134-en18.pdf.

CULT (2017). Draft Opinion for the Committee on Legal Affairs on the proposal for a directive of the on copyright in the Digital Single Market: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2F%2FEP%2F%2FNONSGML%2BCOMPARL%2BPE-595.591%2B01%2BDOC%2BPDF%2BV0%2F%2FEN

IMCO (2017). Draft Opinion for the Committee on Legal Affairs on the proposal for a directive on copyright in the Digital Single Market: http://www.europarl.europa.eu/sides/getDoc.do?type=COMPARL&reference=PE-599.682&format=PDF&language=EN&secondRef=01

ITRE (2017). Draft Opinion for the Committee on Legal Affairs on the proposal for a directive on copyright in the Digital Single Market: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2F%2FEP%2F%2FNONSGML%2BCOMPARL%2BPE-592.363%2B01%2BDOC%2BPDF%2BV0%2F%2FEN

JURI (2017). I Draft Report on the proposal for a directive on copyright in the Digital Single Market: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2F%2FEP%2F%2FNONSGML%2BCOMPARL%2BPE-601.094%2B01%2BDOC%2BPDF%2BV0%2F%2FEN

JURI (2018). I Report Plenary sitting on the proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market (COM(2016)0593 – C8-0383/2016 – 2016/0280(COD)): http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2f%2fEP%2f%2fNONSGML%2bREPORT%2bA8-2018-0245%2b0%2bDOC%2bPDF%2bV0%2f%2fEN

Suber, P. (2012). Open Access, MIT Press.

Hess, Ch. and E. Ostrom (2006). Understanding Knowledge as a Commons, MIT Press.

Von Hippel, E. (2017). Free Innovation, MIT Press.