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

Pawel Kamocki
ELDA / IDS Mannheim
pawel.kamocki@gmail.com

Erik Ketzan
Birkbeck, University of London
eketza01@mail.bbk.ac.uk

Julia Wildgans
IDS Mannheim / Universität Mannheim
j.wildgans@googlemail.com

Andreas Witt
IDS Mannheim / Universität Mannheim / Universität Heidelberg
witt@ids-mannheim.de

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 infrastructures 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[1]. In principle, every reproduction[2] or communication to the public[3] of copyright-protected material requires authorisation from the rightholder[4]. 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

[1] A work can be defined as an original creation in the literary (including computer programmes), artistic or scientific domain. The threshold of originality (*'author's own intellectual creation'*) is relatively easy to meet and one can say that, especially in the case of works of language, originality is *de facto* presumed

[2] The exclusive right of reproduction is construed broadly and includes 'direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part' (art. 2 of the InfoSoc Directive)

[3] The exclusive right of communication to the public refers to *'any communication to the public (...), by wire or wireless means, including the making available to the public (...) in such a way that members of the public may access [the material] from a place and at a time individually chosen by them'*, (i.e. uploading on the Internet — art. 3 of the InfoSoc Directive)

[4] This authorisation is typically granted in an agreement called 'a licence' (Latin *licentio* — permission).

Pawel Kamocki, Erik Ketzan, Julia Wildgans and Andreas Witt 2019. New exceptions for Text and Data Mining and their possible impact on the CLARIN infrastructure. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 66–71.

related rights in the information society (hereinafter: InfoSoc Directive) contains (in its art. 5) a limitative[5] 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)[6], 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. Art. 5.3 (a) allows Member States to adopt exceptions for *use for the sole purpose of (...) scientific research, as long as the source, including the author's name, is indicated (...) and to the extent justified by the non-commercial purpose to be achieved.*

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, especially for research purposes, has been discussed at least since 2011, i.e. the publication of the Hargreaves review[7]. In 2013, a group on Text and Data Mining was created within the Stakeholder's Dialogue *Licenses for Europe*[8]. The academic community, unhappy with the adopted approach (focused on licensing rather than on statutory exceptions), largely withdrew from the process[9]. 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. Meanwhile, some EU Member States decided to adopt TDM exceptions within the current legal framework (i.e. art. 5.3(a) of the Infosoc Directive, cf. *supra*).

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)[10], 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[11], 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

---

[5] Cf. recital 32 of the InfoSoc Directive: *'This Directive provides for an exhaustive enumeration of exceptions and limitations (...)'*

[6] Art. 5.1 of the InfoSoc Directive (so-called *'temporary acts of reproduction'*)

[7] 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 (v. pp. 41-52, esp. p. 48)

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

[9] 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/

[10] Section 29A, sub-section 5, Copyright, Designs and Patents Act 1988

[11] Art. L. 122-5, 10° of the French Intellectual Property Code

decree was supposed to clarify; unfortunately, a proposal for such a decree was rejected in 2017[12] 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[13] 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[14]. Interestingly, the new law expressly uses the word *corpus* to designate a collection of normalised, structured and categorised data created as part of the TDM process. Such a *corpus* can be shared with a *specifically limited circle of persons* (presumably a research team, also multi-institutional). However, once the research is over, the *corpus* has to be deleted or transferred to a specialised library or an archive for permanent storage[15]. The new German exception is expressly non-overridable by contractual clauses[16], 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[17]. Moreover, the adopted solution may turn out to be temporary, as it has an 'expiration date': on 1 March 2023, the new rules will cease to apply. However, before that date the German legislator may decide to maintain them in force, or — more likely — adapt them to ensure compatibility with the upcoming EU Directive (cf. *infra*)

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[18]). 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 on copyright in the Digital Single Market[19]. 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) advocated a solution similar to the one adopted in Germany, requiring payment of equitable remuneration and deletion of the compiled corpus upon the

---

[12] 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

[13] Introduced by the *Urheberrechts-Wissensgesellschafts-Gesetz* (*UrhWissG*) of 7 September 2017

[14] §23 UrhG (also modified by UrhWissG)

[15] §60d(3) UrhG

[16] §60g UrhG

[17] §60h UrhG; the amount of the remuneration should be specified in an agreement concluded between the German states (Länder) and the relevant collecting society (for text data: VG Wort); to the best of our knowledge, no such agreement has been concluded as of yet, and it is quite impossible to predict its content

[18] Cf. art. 27 of the Polish Copyright Act

[19] 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

completion of the project. Its draft also stipulates that *lawful access* to mined works has to be *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[20]. 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 mining should be allowed also for other purposes than scientific research[21]. The Industry, Research and Energy Committee (ITRE) took a similar position[22]. 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[23].

On 25 May 2018, the European Council (under the Bulgarian presidency) published its version of the proposal[24]. As far as TDM exceptions are concerned, this version 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, cf. *supra* about the solution adopted in Germany). 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 would change the paradigm from *"TDM only with permission"* to *"open for TDM by default"*, but would not really provide 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, adopted on 29 June 2018[25] was partly inspired by the Council's proposal. JURI advocates 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.

---

[20] 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, available at: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2F%2FEP%2F%2FNONSGML%2BCOMPARL%2BPE-595.591%2B01%2BDOC%2BPDF%2BV0%2F%2FEN

[21] 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

[22] 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

[23] 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

[24] 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

[25] 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

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), but approved its slightly modified version in the second vote on 12 September 2018. The final stage of the adoption process: three-party negotiations (trilogue) could officially start; however, it did not run very smoothly. The compromise text presented by Romanian presidency as 11 countries (including e.g. Germany, the Netherlands, Italy, Finland, Poland, Portugal and Slovenia) rejected the proposal, and the final vote (initially scheduled for 21 January 2019) was postponed. This was mostly due to the controversies concerning other articles of the proposed Directive (especially 11 and 13), and not the TDM exceptions.

Somewhat unexpectedly, the trilogue reached compromise on 13 February 2019[26]. The text was then debated at JURI and presented for a plenary vote by the European Parliament. On 26 March 2019, the Parliament adopted the Directive (with 348 MEPs votes for, 274 votes against and 36 abstentions)[27]. At the moment (as of 1 April 2019), it still has to be approved by the Council before it can enter into force, but this is usually a formality.

The TDM exceptions in the adopted text are similar to those proposed by the Council and approved by the Parliament in 2018. The mandatory exception (in article 3) benefits only (public) research organisations and cultural heritage institutions, and it is limited to research purposes (including commercial research). What has changed, however, is that the copies (which still have to be stored *with appropriate level of security*) may be retained for research purposes (so, unlike in the previous versions and in the German exception, they do not have to be deleted upon the completion of the project). Like in the original proposal, rightholders may use Digital Rights Management *"to ensure the security and integrity of the networks and databases"*, but without going beyond what is necessary to achieve this objective. The exception is not overridable by contracts.

The newly added (and renumbered) article 4 contains an optional exception with potentially unlimited beneficiaries and scope of purposes, the only limitation being that this exception can only apply to the content for which the rightholders have not expressly reserved the right to mine (so, potentially everything can become '*mineable by default*'). This leaves a lot of leeway to Member States in allowing TDM for other purposes than research, and to other actors than public research organisations and cultural heritage institutions. However, these optional exceptions, unlike the mandatory one, will probably be overridable by contracts.

The Directive will have to be implemented within two years of its entry into force (article 24). However, the transposition process may not run very smooth (because of the aforementioned controversial provisions unrelated to TDM) and may be significantly delayed.

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

Language researchers will receive substantial benefits and some legal certainty from the new TDM exceptions. However, 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 and conduct (*in-house*) research, researchers have fewer incentives to care about proper licensing and sharing their datasets and results (e.g. within research infrastructures)[28]. This may in turn considerably reduce the *knowledge commons* (i.e. immaterial resources that — due to proper licensing — can be freely accessed and re-used by anyone and for any purpose[29]) 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[30] may paradoxically become more difficult, as they require sharing of data between various groups of stakeholders. In order to

---

[26] http://www.europarl.europa.eu/news/en/press-room/20190212IPR26152/agreement-reached-on-digital-copyright-rules

[27] http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+TA+P8-TA-2019-0232+0+DOC+PDF+V0//EN

[28] On incentives for Open Access in the academic community, see esp. Suber, P. (2012). Open Access, MIT Press

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

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

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.

References

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.