UCLouvain
Institut pour la recherche interdisciplinaire en sciences juridiques

CRIDES
Centre de recherche interdisciplinaire Droit, Entreprise et Société

CRIDES Working Paper Series no. 1/2021

Ensuring Text and Data Mining: Remaining Issues With the EU Copyright Exceptions and Possible Ways Out

by Rossana Ducato and Alain Strowel

FEBRUARY 2021

Cite as: R. Ducato, A. Strowel, *Ensuring Text and Data Mining: Remaining Issues With the EU Copyright Exceptions and Possible Ways Out*, CRIDES Working Paper Series no. 1/2021; forthcoming in European Intellectual Property Review.

The **CRIDES**, *Centre de recherche interdisciplinaire Droit Entreprise et Société*, aims at investigating, on the one hand, the role of law in the enterprise and, on the other hand, the function of the enterprise within society. The centre, based at the Faculty of Law – UCLouvain, is formed by four research groups: the research group in economic law, the research group in intellectual property law, the research group in social law (Atelier SociAL), and the research group in tax law.

www.uclouvain.be/fr/instituts-recherche/juri/crides

ENSURING TEXT AND DATA MINING: REMAINING ISSUES WITH THE EU COPYRIGHT EXCEPTIONS AND POSSIBLE WAYS OUT

Rossana Ducato[*] and Alain Strowel[†]

ABSTRACT

Text and Data Mining (TDM) is a vital tool in the Big Data economy. 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, discover new trends, insights or correlations.

The importance of TDM has been understood by the European legislator, which has introduced two specifically tailored exceptions in the Copyright in the Digital Single Market Directive. After a critical analysis of the new provisions, the paper argues that they still present several flaws that risk to stifle AI developments in Europe. Thus, the contribution outlines an interpretative framework, based on the analysis of the infringement test, to rethink the rights of reproduction and extraction in line with the economic rationale of copyright and the database right. Furthermore, the paper makes suggestions to improve the TDM exceptions at national level. In conclusion, it points out the remaining challenges of private ordering and trade secrets for research and AI innovation.

KEYWORDS

Copyright – sui generis right - exceptions and limitations – text and data mining – research – technological protection measures – contract

[*] **Rossana Ducato** (rossana.ducato@abdn.ac.uk) is Lecturer in IT Law and Regulation at the University of Aberdeen, School of Law. She is also fellow of CRIDES and lecturer of the Jean Monnet course "European IT Law by Design" at UCLouvain. https://www.abdn.ac.uk/law/people/profiles/rossana.ducato

† **Alain Strowel** (alain.strowel@uclouvain.be) is Professor at UCLouvain, Université Saint-Louis – Bruxelles, KULeuven, Munich IP Law Centre, and senior partner at Pierstone. https://www.usaintlouis.be/sl/100297.html

ACKNOWLEDGEMENTS

This article updates and expands the work presented in A. Strowel and R. Ducato, "Artificial intelligence and text and data mining: a copyright carol" in E. Rosati (ed.), Handbook of EU Copyright Law, Routledge, forthcoming 2021.

A sincere thanks to Roberto Caso and Ula Furgal for the constructive discussion on an early draft of this article.

The authors have jointly conceived the paper and share the views expressed therein. Nonetheless, while Section 4 is attributable to Alain Strowel, Section 3 is specifically attributable to Rossana Ducato. Both authors equally contributed to the drafting of the remaining sections.

1. Introduction

Text and Data Mining (TDM) is an essential tool for the data economy. TDM uses techniques from natural language processing, machine learning, information retrieval, and knowledge management for the automated analysis of texts and digital content (structured and unstructured data), in order to extract information, identify patterns, discover new trends, insights or correlations.

Considering its capacity to retrieve and analyse huge amounts of data (Big Data), TDM represents one of the main tools for research and for artificial intelligence (AI) applications. It can work as a technological enabler in several areas, from predictive algorithms in the health sector[1], to sentiment analysis[2], fact checking[3], smart disclosure systems,[4] biometric recognition systems,[5] etc.

[1] Like the BlueDot project, which predicted the spread of Covid-19 virus well before the World Health Organization (WHO). This example is reported in Sean Flynn, Christophe Geiger, João Pedro Quintais, 'Implementing User Rights for Research in the Field of Artificial Intelligence: A Call for Action at International Level' (Kluwer Copyright Blog, 21 April 2020) <http://copyrightblog.kluweriplaw.com/2020/04/21/implementing-user-rights-for-research-in-the-field-of-artificial-intelligence-a-call-for-action-at-international-level/> accessed 31 July 2020. See, also, Neesha Jothi, Nur'Aini Abdul Rashid, Wahidah Husain, 'Data Mining in Healthcare – A Review', Procedia Computer Science, Volume 72, 2015, p. 306-31; Wullianallur Raghupathi, 'Data mining in health care' in Stephan P. Kudyba (ed), *Healthcare informatics: improving efficiency and productivity* (CRC Press 2016), 353; Carsten Eickhoff, Kim Yubin and Ryen W. White, 'Overview of the Health Search and Data Mining Workshop' (2020) Proceedings of the 13th International Conference on Web Search and Data Mining 901.

[2] Jalel Akaichi, Zeineb Dhouioui and Maria José López-Huertas Pérez, *Text mining facebook status updates for sentiment classification* (IEEE 2013).

[3] Stefano Guarino, Noemi Trino, Alessandro Chessa e Gianni Riotta, 'Beyond Fact-Checking: Network Analysis Tools for Monitoring Disinformation in Social Media' in H. Cherifi et al. (eds) *Complex Networks and Their Applications VIII. COMPLEX NETWORKS 2019. Studies in Computational Intelligence* (Springer 2019), 436.

[4] For an overview of the first examples of smart disclosure systems in the consumer and data protection domain, see Rossana Ducato and Alain M Strowel, 'Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to Machine Legibility' (2019) 50 IIC - International Review of Intellectual Property and Competition Law 649.

[5] Wen-Kwang Tsao and others, 'A data mining approach to face detection' (2010) 43 Pattern Recognition 1039; Shiv Naresh Shivhare and Sri Khetwat Saritha, 'Emotion detection from text documents' (2014) 4 International Journal of Data Mining & Knowledge Management Process 51.

Despite the advantages of TDM for research and for new tools and services, legal uncertainties under EU and national laws combined with the use of technical blocks and contractual restrictions are likely to hinder the practice of TDM in Europe. [6]

The legal issues raised by TDM techniques involve a diversity of legal domains and fundamental rights, from privacy and data protection[7] to freedom of expression[8] and intellectual property.[9] Each of them requires a different assessment and, in some cases, will justify restrictions to TDM for the protection of other interests and values.

This paper focuses on the intellectual property (IP) issues raised by TDM, and in particular the scope of the TDM exceptions to copyright, database and computer programs, and press publishers protections.

In principle, TDM may clash with the IP framework if a work or a database qualify for protection under Directive 2001/29/EC ("InfoSoc Directive")[10], Directive 2009/24/EC ("Software Directive")[11], or Directive 96/9/EC ("Database Directive")[12]. Depending on the technique used, TDM may involve: 1) the reproduction of a copyrighted content; 2) the extraction of a substantial part of the database; 3) the reproduction and

[6] See the survey on content blocking conducted by LIBER (the Association of European Research Libraries) and the short report of March 2020 entitled "Europe's TDM Exception for Research: Will It Be Undermined by Technical Blocking From Publishers?" (https://libereurope.eu/blog/2020/03/10/tdm-technical-protection-measures/).

[7] Sandra Wachter and Brent Mittelstadt, 'A right to reasonable inferences: re-thinking data protection law in the age of big data and AI' (2019) Columbia Business Law Review, https://osf.io/preprints/lawarxiv/mu2kf/. See, in particular, p. 72 ff.

[8] Sorrell v. IMS Health Inc., 564 U.S. 552 (2011). Bonnie Kaplan, 'Selling health data: de-identification, privacy, and speech' (2015) 24 Cambridge Quarterly of Healthcare Ethics 256.

[9] The regulation of TDM is tackled as a legal exception to copyright and can thus raise the fundamental right of intellectual property protection under Article 17(2) EU Charter of fundamental rights. See Alain Strowel, 'Intellectual Property Strengthened by the Court of Justice Interpretation of Article 17(2) of the EU Charter of Fundamental Rights' in Pollicino O, Riccio GM, Bassini M (eds), *Copyright and Fundamental Rights in the Digital Age* (Edward Elgar 2020).

[10] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, OJ L 167, 22.6.2001, p. 10–19.

[11] Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs, OJ L 111, 5.5.2009, p. 16–22.

[12] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, OJ L 77, 27.3.1996, p. 20–28.

adaptation of a computer program.[13] At the same time, the system of exceptions and limitations to those IP rights did not provide sufficient space for enabling some TDM activities.[14]

In order to solve legal uncertainties and compete with systems offering a favourable legal framework (e.g., Japan or the US), the European legislator introduced two *ad hoc* TDM exceptions in the 2019/790 Copyright in the Digital Single Market Directive (hereinafter "CDSMD")[15].

The aim of the paper is to critically analyse those provisions, assessing whether the scope of the TDM exceptions introduced in the CDSMD are sufficient to promote an adequate development of research, especially in the field of data-driven technologies (like AI applications).

The article is structured as follows. Section 2 introduces the two new TDM exceptions: Article 3 CDSMD, which establishes an exception for research organisations and cultural heritage institutions when the TDM is performed for research purposes; and, Article 4 CDSMD, which introduces a broader TDM exception which does not limit its beneficiaries and is not conditioned by a research purpose. In the latter case, the rightholder can nevertheless restrict TDM by contract or via a machine-readable means.

In Section 3, the critical points of the new provisions are discussed. Despite the advances of the CDSMD regime, the latter risks not to fulfil their initial promises. The criticisms to the TDM exceptions are developed in five points: 1) narrow scope of the exceptions; 2) systematic inconsistencies in the scope of application; 3) unsatisfactory

[13] A communication to the public or a reuse do not always occur in the case of TDM: the latter usually elaborates the information and publishes the results of the analysis in the form of aggregate data, statistics, reports, etc. Therefore, unless the output of the TDM shows the whole or the excerpts of the protected work or the database, there will be no communication to the public or reuse. J.P. Triaille, J. de Meeûs d'Argenteuil and A. de Francquen, *Study on the legal framework of text and data mining (TDM)*, March 2014; M. Caspers and L. Guibault, *Baseline report of policies and barriers of TDM in Europe*, 2016.

[14] Ducato and Strowel (n 4) .

[15] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, OJ L 130, 17.5.2019, p. 92–125. The Directive has introduced a definition of TDM in the following terms: "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" (Article 2.2, CDSMD). The definition is sufficiently broad to embrace the current TDM application panorama. For a technical definition of TDM, see Marti A. Hearst, 'Text Data Mining' in Ruslan Mitkov (ed), *The Oxford Handbook of Computational Linguistics* (Oxford University Press 2003). Specifically on text mining, Ronen Feldman and James Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data* (Cambridge university press 2007). For an extensive analysis of the definition of TDM, see Triaille, de Meeûs d'Argenteuil and de Francquen (n 13).

regime settled for technological limitations; 4) lawful access as a pre-condition for TDM; 5) frustration of the legislative intent to provide more legal certainty.

In Section 4 we question whether the freedom to conduct TDM would have been better framed with a normative interpretation of the scope of the reproduction right. If copyright is aimed at protecting the expressive features of a work (the work as an expression addressed to a person or public), TDM cannot be considered an infringing activity, since it does not use the "work as an expressive work". Such a logic, however, cannot be applied to the rationale of the right to extraction covered by the sui generis database right which is a related right for an entrepreneur and belongs to industrial property. The substantial investment in the 'making' of the database is protected against any extraction of a substantial part (no matter the purpose of its use). Nevertheless, a purposive interpretation of the sui generis right (which is infringed only by an act leading to reconstituting a substantial part of the content) might leave some room for lowering the barrier for TDM on databases.

In conclusion, the paper sets forth recommendations and suggestions for the national transposition of the TDM exceptions, and outlines the challenges beyond copyright and the database right that remain to be addressed to unleash the full potential of TDM and AI research in Europe.

2. TDM exceptions and limitations

During a three-year legislative process, several provisions of the draft CDSMD were hotly debated, including the initial Article 3 on TDM. At the end, the Directive introduces two new exceptions for TDM (enshrined at Articles 3 and 4) that Member States will have to transpose in their national legislation. In the following paragraphs, the main aspects of those provisions are detailed.

2.1 The TDM exception for research

Article 3 contains the TDM exception for the purpose of scientific research.[16] Such an exception was included in the 2016 Commission's proposal for the Directive on copyright in the DSM.[17]

[16] On the notion of scientific research in different legal contexts, see EDPS, 'Preliminary Opinion on data protection and scientific research' (2020), https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf.

[17] Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market, COM/2016/0593 final - 2016/0280 (COD).

Four components of this TDM exception can be distinguished: 1) the rights affected; 2) the beneficiaries; 3) the scope; 4) the pre-existing condition.

As to the first element (rights affected), Article 3 provides for an exception to:

1) the right of reproduction of whole or part of databases protected by copyright (Article 5(a) Database Directive);

2) the right of extraction of whole or a substantial part of databases covered by the sui generis right (Article 7(1) Database Directive);

3) the right of reproduction in whole or part of works, fixations of performances, phonograms, fixations of broadcasts, the original and copies of films (Article 2 InfoSoc Directive);

4) the right of reproduction of on-demand press publications[18] (the new press publisher rights established by Article 15(1) CDMSD).[19]

With regard to its beneficiaries, the exception is granted to research organisations and cultural heritage institutions only. Research organisations are defined as "a university, including its libraries, a research institute or any other entity, the primary goal of which is

[18] For the purpose of the CDSMD, press publication means "a collection composed mainly of literary works of a journalistic nature, but which can also include other works or other subject matter, and which: (a) constitutes an individual item within a periodical or regularly updated publication under a single title, such as a newspaper or a general or special interest magazine; (b) has the purpose of providing the general public with information related to news or other topics; and (c) is published in any media under the initiative, editorial responsibility and control of a service provider" (Article 2(4) CDSMD). Recital 56 adds that the new right exists for "journalistic publications, published in any media, including on paper, in the context of an economic activity that constitutes a provision of services under Union law". For instance, the notion includes "daily newspapers, weekly or monthly magazines of general or special interest, including subscription-based magazines, and news websites" (Recital 56). Press publications include the article, as literary work, but also other subject matter accompanying the article, such as photo and videos. The definition, however, does not extend to scientific journals and blogs "that provide information as part of an activity that is not carried out under the initiative, editorial responsibility and control of a service provider, such as a news publisher" (Recital 56 and Article 2(4) *in fine* CDSMD).

[19] Article 15(1) CDSMD confers to press publishers not only the exclusive right recognised at Article 2 of the InfoSoc Directive, but also the right at Article 3(2) InfoSoc Directive. However, Article 3 CDSMD refers to the acts of reproduction and extraction only, and the TDM exception does not extend to the right of making available press publications to the public. The drafting of Article 3 referring to Article 15(1) confirms that TDM is limited to the analysis of text and data in order to generate something different from the original corpus subject to mining.

to conduct scientific research or to carry out educational activities involving also the conduct of scientific research: (a) on a not-for-profit basis or by reinvesting all the profits in its scientific research; or (b) pursuant to a public interest mission recognised by a Member State; in such a way that the access to the results generated by such scientific research cannot be enjoyed on a preferential basis by an undertaking that exercises a decisive influence upon such organisation" (Article 2(1) CDSMD). While, the notion of cultural heritage institution refers to "a publicly accessible library or museum, an archive or a film or audio heritage institution" (Article 2(3) CDSMD).

Those who can benefit from the TDM exception are essentially institutions that provide a cultural or public service in the interest of society on a non-for-profit basis.[20] The reality of research is however more complex. It is not unusual for a public university to be involved in a consortium with industry and SMEs (it is actually encouraged by many research programs supported by the European Commission). The issue is partially addressed in the CDSMD Recitals where it is stated that research organisations and cultural heritage institutions "should be able to rely on their private partners for carrying out text and data mining, including by using their technological tools" (Recital 11). Therefore, in the context of public-private partnerships, the CDSMD leaves some room for the private actors to benefit from the exception at Article 3 CDSMD, if required by the needs of the project. At the same time, this implies that such condition will not extend beyond the scope of the collaborative project or after its conclusion.

The third element of the TDM exception concerns its scope: TDM activities shall be directed to research purposes only. By "scientific research" it is meant research both in natural and human sciences (Recital 12).[21]

Fourthly Article 3 requires from the TDM beneficiary to "have lawful access" (Article 3(1)) to the work or other protected subject matter to be mined, what has been labelled here the "pre-existing" condition. In other words, the exception only works under the condition that research organisations and cultural heritage institutions have already lawful access to the resource. By lawful access the CDSMD means "access to content based on an open access policy or through contractual arrangements between rightholders and research organisations or cultural heritage institutions, such as subscriptions, or through other lawful means [...] Lawful access should also cover access to content that is freely available online" (Recital 14 CDSMD). Therefore, if the content is protected and the user does not enjoy a right to access and use it, TDM cannot be performed for research purposes. This is a severe limitation to the exception that will be addressed below.

[20] Hospitals carrying out research may be included in such a definition, as expressly mentioned in Recital 12 CDSMD.

[21] On the notion of scientific research within the CDSMD, see moreover in para. 3.1.

If all the conditions listed in Article 3(1) are met, the research organisation or the cultural heritage institution is allowed to perform the TDM activity and to retain the copies of the works and subject matter made according to the exception.[22] This is a relevant addition that recognises the importance of maintaining the copies of the protected work and of the data to fulfil the scientific rationale, for instance by allowing the peer review and the verification of the results. The TDM exception, though, comes with the obligation to store the copies "with an appropriate level of security".

The "security measures" can be established "to ensure the security and integrity of the networks and databases where the works or other subject matter are hosted" by the rightholders (Article 3(3) CDSMD). However, such security or integrity measures on the files stored should in no way limit the possibility of applying the TDM tools. The Directive leaves the exact definition of the measures required from the lawful miner under Article 3(2) and of those applicable by the rightholder (Article 3(3)) to the determination of the parties, but under the final approval by the Member States. The latter are encouraged to define best practices concerning the application of the above-mentioned measures involving the relevant stakeholders (i.e., rightholders, research organisations and cultural heritage institutions).[23] In practice, it is likely that those measures will make it more complex or cumbersome to conduct TDM on the corpuses, and it will be very difficult to disentangle the measures objectively justified by the security or integrity of the corpuses and those that go beyond (and hamper TDM).

Finally, one important aspect that has been clearly established in the Directive is the prohibition of contractual provisions overriding the TDM exception for research purposes.[24]

2.2. TDM exceptions and limitations for everyone

Article 4 CDSMD contains another provision favourable to TDM activities that was specifically added during the legislative process leading to the adoption of the Directive.[25]

[22] See Article 3(2) CDSMD.

[23] See, Article 3(4) CDMSD.

[24] See, Article 7(1) CDSMD.

[25] In particular, see the version of the text dated 25 May 2018 (Council of the EU, Interinstitutional File: 2016/0280(COD), doc. 9134/18, available here: https://www.consilium.europa.eu/media/35373/st09134-en18.pdf).

Analogously to Article 3, the provision here commented establishes that the miner must have lawful access to the resource as a *sine qua non* condition.[26] Apart from this common ground, all the other elements of the exception differ.

Notably, Article 4 is broader in terms of the plethora of beneficiaries. There are in fact no limitations or qualifications to comply with: everyone is in principle entitled to the exceptions or limitations that will be implemented by the Member States under Article 4. The scope of the exception is wider as well: not only research purposes are covered, but any TDM activity, whether non-profit or for profit, as long as it falls under the definition of TDM in Article 2(2) CDSMD.

Another difference concerns the object covered by the exception here at stake. In addition to the list of rights that can be limited by TDM for research purposes, Article 4 includes: the right to reproduction (Article 4(1)(a) Software Directive) and the right to adaptation of computer programs (Article 4(1)(b) Software Directive). Indeed, TDM activity may also concern software code and a limitation to the exclusivity of the rightholder is welcome in this area as well, if the legislative intent is to lower the barriers for TDM.

3. The TDM exceptions: the good, the bad and… the weird.

After this preliminary overview, it is possible to engage in the critical analysis of the regime introduced by the CDSMD. In this section the positive (or at least the satisfactory) aspects of the new regime ("the good") will be discussed, but also the potential shortcomings and drawbacks ("the bad"), and the points that remains unclear, controversial or inexplicable ("the weird"). The main elements are reported in Table 1 below.

Starting from the "good": a first, albeit trivial, clarification made by the Directive is that TDM does not necessarily interfere with copyright and neighbouring rights (Recital 9 CDSMD). TDM may involve the processing of mere facts or data, which are not copyrightable as such. Any limitation to the processing would conflict with the fundamental dichotomy between unprotected ideas (facts, data or information) and protected expression, which lies at the core of copyright. The database right is subject to a similar limitation as the protection only applies to the "whole or a substantial part" of the contents of a database (Article 7 Database Directive), not to the data (or other non-protected elements) comprised in the database. But in practice, many corpuses contain a

[26] The language used in Article 4 slightly varies as it refers to "lawfully accessible" works or other subject matter, while Article 3 refers to the works or other subject matter "to which [the beneficiaries, i.e. the research organisations and cultural heritage institutions] have lawful access", but this does not affect the pre-condition.

mix of protected and unprotected elements, for instance the scientific publications and the data used for the demonstrations explained in the scientific journal.

Furthermore, some TDM activities on protected materials might not involve reproduction at all;[27] or, when they do, they might be satisfactorily exempted as temporary reproductions (Article 5(1) InfoSoc Directive)[28] or as reproductions for scientific research (Article 5(3)(a) InfoSoc Directive). However, the interface between TDM and the national exceptions for research is complex, and one can conclude that the additional TDM for research exception was needed. Indeed, the pre-existing exceptions to copyright and sui generis right, limited *per se* or per the effect of the inconsistent national implementations, were unable to sufficiently cover the operations performed by TDM and ensure clarity about the rights of the users.[29] Not surprisingly, even before the CDSMD, some Member States decided to adopt specific provisions to support TDM.[30] It is the case of UK (2014)[31], France (2016)[32], Estonia (2017)[33], and Germany (2017)[34]. Therefore, the goal of improving the legal certainty about uses, including cross-border, of protected materials in case of TDM activity is to be welcomed.

As said, the IP framework is only one possible obstacle in the way of TDM and its applications: notably contracts and technological protection measures may restrict its uses

[27] According to some scholars, there is no reproduction if the tool simply spots one or two words through the text without making a copy of the work (e.g., spotting and counting the occurrences of the word "malaria"). Triaille, de Meeûs d'Argenteuil and de Francquen (n 13), 31. Similarly, I.A. Stamatoudi, 'Text and data mining' in I.A. Stamatoudi (ed), *New Developments in EU and International Copyright Law* (Wolters Kluwer 2016), p. 1261; Maria Lillà Montagnani and Giorgio Aime, 'Il text and data mining e il diritto d'autore' (2017) AIDA 376, pp. 379 ff.

[28] Provided that the copy is transient/incidental, forms an integral and essential part of a technological process and whose sole purpose is to enable: (a) a transmission in a network between third parties by an intermediary, or (b) a lawful use. Article 5(1) InfoSoc Directive.

[29] An excursus of the limitation of the current legislative framework of TDM is offered in Ducato and Strowel (n 4). For an overview of the different state of implementation of copyright exceptions and limitations in Europe, see the interactive map realised by Kennisland and supported by a grant from the Open Society Foundation, https://copyrightexceptions.eu.

[30] See, Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko, *The exception for text and data mining (TDM) in the proposed Directive on Copyright in the Digital Single Market: legal aspects: in-depth analysis* (European Parliament 2018), pp. 17-18.

[31] Copyright, Designs and Patents Act 1988, § 29A (UK).

[32] Article 38 of the Law No. 2016-1231 of for a Digital Republic added paragraph 10 to Article L122-5 and paragraph 5 to Article L342-3 of the Intellectual Property Code (Code de la propriété intellectuelle).

[33] Estonian Copyright Act, Article 19(3).

[34] Section 60d, Law on Copyright and Related Rights (Urheberrechtsgesetz).

(even when the text or the data are not protected by any exclusive right.)[35] A positive point added by the Directive is that at least the TDM exception for research is binding for the parties and cannot be overridden by contract.[36] This contrasts with the non-binding character of the broader exception under Article 4 CDSMD.

Another important step, that materialised during the legislative procedure, has been the recognition of the importance of TDM not just for research purposes, but also due to its various applications in other fields (see Recital 18). Therefore, to the initial and only TDM research exception proposed by the Commission in 2016, the final text of the Directive added a second TDM exception available to anyone for any purpose (Article 4 CDSMD).

Furthermore, both exceptions are now mandatory, meaning that Member States shall introduce them into their national law.

However, the loosening of copyright stops short of permitting TDM. The exceptions introduced by the CDSMD appear in fact highly limited and, more broadly, the complex and often unclear drafting of some provisions does not facilitate their future interpretation and application.

The criticisms to the TDM exceptions can be organized around five main points: 1) narrow scope of the exceptions; 2) systematic inconsistencies in the objective scope of application; 3) technological limitations; 4) lawful access as a pre-condition for TDM; 5) frustration of the legislative intent to provide more legal certainty.

3.1. The narrow scope of the exceptions

The formulation of the TDM research exception (Article 3) raises a first preliminary question concerning its precise scope of application. As mentioned in para. 2.1., scientific research is not expressly defined in the CDSMD, but Recital 12 states that it refers to works both in natural and human sciences. This formulation, that relies on the popular dichotomy of natural v. human sciences, is actually quite vague, and, paradoxically, can lead to restrictive interpretations concerning the scope of application of Article 3. Compared to the broad formulation of the equivalent research exception under the

[35] With reference to the database protection, see CJEU, Case C-30/14, *Ryanair Ltd v PR Aviation BV* [2015], ECLI:EU:C:2015:10.

[36] Article 7(1) CDSMD. This contrasts with the non-binding character of the broader exception under art.4 CDSMD. See more in the next paragraph.

GDPR[37], Article 3 could be interpreted narrowly. For example, would scientific research in the field of medicine be included under the scope of Article 3 CDSMD? What about statistical research in the field of economics? Would it fall under the label of human science? If we think, for example, to the classification of the field of sciences proposed in the OECD Frascati Manual, we find natural science and humanities, but also – as separate entities - social sciences, engineer and technology, medical and health sciences, agricultural sciences.[38] The scientific sectors distinguished by the European Research Council (ERC), another popular classification among scholars, do not refer to the categories of natural and human sciences. Instead, the main branches are: 1) physical sciences and engineer; 2) life sciences; 3) human and social sciences.[39] These classifications do not aim to create a taxonomy of the scientific branches, but are distinguished to serve different purposes, e.g. to organise research calls. Nevertheless, they show clearly that there might be different nuances in (the interpretation of) the classification of science. Applied research, for example for public health or relating to technology, might fall outside the scope of the Article 3 exception (let's imagine that computer science is not classified as natural science *stricto sensu*. This might have implications for AI development).

Recital 12 could have been better formulated by referring to scientific research *tout court*, understood as any activity, performed according to the pertinent methodological standards, aimed at generating new knowledge and advancing the state of the art in a given field.[40] There would be no compelling reason for restricting the scope of application of Article 3 depending on the scientific area where the research is conducted. Thus, we suggest that the national implementations of the Directive adopt a broad notion of scientific research in the context of the TDM exception.

This conclusion is even more necessary considering that Article 3 is already very narrowly designed, being subject to a double requirement. The exception works only for two categories of beneficiaries (research organisations and cultural heritage institutions) and for a specific objective (research purposes). This means that, for example, independent

[37] In the General Data Protection Regulation, recital 156 defines research "in a broad manner including for example technological development and demonstration, fundamental research, applied research and privately funded research. In addition, it should take into account the Union's objective under Article 179(1) TFEU of achieving a European Research Area. Scientific research purposes should also include studies conducted in the public interest in the area of public health". For an analysis of the notion of scientific research within the data protection context, see Rossana Ducato, 'Data protection, scientific research, and the role of information' (2020) 37 *Computer Law and Security Review* 105412.

[38] See, OECD, Working Party of National Experts on Science and Technology Indicators, Revised field of science and technology (FOS) classification in the Frascati Manual, 2007, https://www.oecd.org/science/inno/38235147.pdf

[39] See, for instance, the ERC panels' classification: https://erc.europa.eu/sites/default/files/document/file/erc%20peer%20review%20evaluation%20panels.pdf

[40] On the notion of scientific research (although in the data protection context), see EDPB, *Guidelines on Consent under Regulation 2016/679 (wp259rev.01)*, 2018, p. 27.

researchers carrying out projects in the public interest can be automatically cut out from the benefit of Article 3. Similarly, Article 3 will not exonerate any TDM activity that is not related to scientific research. For instance, a university is not entitled to rely on this exception for the pursuit of other institutional goals, such as educational, administrative or governance purposes (but Article 4 helps to bypass this limitation). Usually, educational uses when performed by not-for-profit organizations are exonerated along the research uses; this is not the case with the Article 3 exception. In practice, it is not clear whether this limitation as to the purpose will seriously block research institutions to use the outcome of TDM for illustrating teaching or for other non-research purposes.

Furthermore, the formulation of Article 3 fails to recognise the reality of scientific research nowadays. First of all, the public benefit that can be pursued by scientific research is something that goes beyond the walls of traditional research institutions.[41] For instance, in light of the growing attention for behaviourally informed regulation, TDM could be used by policy makers to test draft policies and new legislative interventions. Journalists could benefit from such an exception for researching sources and check the authenticity of a news. Commercial private actors, which play a decisive role in research and development in the field of AI, would be similarly excluded from the exception, whether they are Big Techs or start-ups. Furthermore, considering the level of job insecurity in academia, researchers between two short-term employment contracts would in principle be excluded from Article 3 when they mostly need to use those tools for remaining actively involved in research and visible (or when they have to prepare a grant application to obtain a new funding). Therefore, the narrow definition of research organisations risks to penalise a category of researchers when they are in a delicate professional situation.

Second, the status of the copies of the protected material realised during the TDM process is not entirely clear: they can be retained, but it is not spelled out whether such materials can be used for the purpose of scientific research (including further verification) by someone *external* to the organisation. Fortunately, the Recital 15 *in fine* refers to the possibility of relying on the exception for research provided for in Article 5(3)(a) InfoSoc Directive, but reproduction and extraction for scientific peer review (and joint research) are not necessarily exempted under the national implementations of the rather unspecified Article 5(3)(a). Recital 10 correctly recognizes that the existing exceptions are "not fully adapted to the use of technologies in scientific research". To eliminate any uncertainty and ensure the needs of research, Article 3 could have been drafted along the lines of Section 60d of the *Urheberrechtsgesetz*. The German rule explicitly establishes that the corpus created

[41] Many scholars have argued that the exception for research should be broadened. For instance, M. Caspers and Lucie Guibault, *A right to 'read' for machines: Assessing a black-box analysis exception for data mining* (2016); Thomas Margoni and Giulia Dore, 'Why We Need a Text and Data Mining Exception (But it is Not Enough)' Zenodo; Thomas Margoni and Martin Kretschmer, 'The Text and Data Mining exception in the Proposal for a Directive on Copyright in the Digital Single Market: Why it is not what EU copyright law needs' CREATe Blog; Reto Hilty and Heiko Richter, 'Position Statement of the Max Planck Institute for Innovation and Competition on the Proposed Modernisation of European Copyright Rules Part B Exceptions and Limitations (Art. 3–Text and Data Mining)' (2017); Geiger, Frosio and Bulayenko (n 32); Eleonora Rosati, 'An EU Text and Data Mining Exception for the Few: Would it Make Sense?' (2018) 6 *Journal of Intellectual Property Law & Practice* 429.

through TDM must be accessible to individual third parties to check the quality of scientific research.[42]

On the other hand, Article 4 is only apparently broader than the TDM research exception: the rightholder has the power to opt-out, by expressly reserving the right to make reproduction or extraction for TDM in an appropriate manner.[43] For example, when the content is made publicly available online, this can be done via the terms and conditions of the website or via machine-readable means.[44] In other words, the exception and limitation enacted on the basis of Article 4 can be overridden by any expression of will, whether unilateral or by contract.[45] As we have shown elsewhere,[46] many online platforms are already prohibiting TDM in their online terms of use, and are thus engaged in what is called private ordering. They are contractually prohibiting users from performing any act of reproduction on the website content or/and using an exclusion protocol to impede crawling or indexing (the robot.txt protocol). In light of this, the impact of Article 4 will be rather limited in enabling TDM over content available on the internet.

Furthermore, the storage of the copies done for TDM is confined to the activity of mining as such. Therefore, once the process is completed, all reproductions and extractions should in principle be deleted. Because the storage of such content is likely to be transient, Article 4 does not oblige the miner to any particular security measure (on the contrary, Article 3(2) CDSMD imposes an appropriate level of security for the stored data corpus).

3.2. Systematic inconsistencies in the scope of application

When comparing the two TDM exceptions and analysing them in relation to other provisions of the CDSMD, some inconsistencies appear.

---

[42] See Christophe Geiger, Giancarlo F. Frosio and Oleksandr Bulayenko (n 32) 23.

[43] Article 4(2) and Recital 18, CDSMD.

[44] The use of the robot.txt will suffice to prohibit TDM on a content available online. P. Bernt Hugenholtz, *The New Copyright Directive: Text and Data Mining (Articles 3 and 4)* (Kluwer Copyright Blog 2019), http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/ . The robot.txt is an exclusion protocol that content providers can insert into the root directory to prevent crawling or indexing activities on certain pages of their website. See, http://www.robotstxt.org/robotstxt.html.

[45] Since Article 7(1) CDSMD does not refer to Article 4.

[46] Ducato and Strowel (n 4)'. On the negative impact of such a provision on the development of AI creativity, see moreover Eleonora Rosati, 'Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and its Role in the Development of AI Creativity' (2019) 2 *Asia Pacific Law Review* 198.

First of all, both Articles 3 and 4 include an exception or limitation to the press publisher right of reproduction. However, such an extension in the context of TDM for research purposes (Article 3) does not appear straightforward. The protection of Article 15(1) CDSMD is granted to the publishers for the online use of their press publications by *information society service providers*, such as online news aggregators and media monitoring services.[47] Such provision does not grant the publishers a right against any third party. Research organisations or cultural heritage institutions are not likely to be qualified as information society service providers. The reference to Article 15 might make sense if the purpose is to cover services used by universities and cultural heritage institutions for research purposes. It would be the case, for example, of services provided by third parties for crawling newspaper articles.

Second, the limitation of Article 4 applies to the economic rights granted by the Software Directive.[48] Here it is possible to observe two sets of internal inconsistency.

First of all, one might wonder why the Article 3 does not create a limitation to the copyright on software. Are research organisations and cultural heritage institutions not entitled to perform TDM on software for research purposes? Why is it possible, in the absence of any reservation by the owner of the copyright on software, for a commercial entity to conduct TDM on code, while researchers in the computer departments of research institutions would not enjoy this privilege?[49]

A possible answer is that Article 5(3) of the Software Directive already provides for the so-called "black box analysis" exception.[50] The lawful user can, without the authorisation of the rightholder, observe, study or test the functioning of the program "in order to determine the ideas and principles which underlie any element of the program if he does so while performing any of the acts of loading, displaying, running, transmitting or storing the program which he is entitled to do". Therefore, one might argue that anyone can already benefit from a research exception on computer programs. Thus there was no need for Article 3 CDSMD to refer to the Software Directive 2009/24/EC. On the

[47] See, Joao Pedro Quintais, *The New Copyright Directive: A tour d'horizon – Part II (of press publishers, upload filters and the real value gap)* (2019), http://copyrightblog.kluweriplaw.com/2019/06/17/the-new-copyright-directive-a-tour-dhorizon-part-ii-of-press-publishers-upload-filters-and-the-real-value-gap/. For the definition of information society service providers, see Article 1(b), Directive (EU) 2015/1535.

[48] Articles 4(1)(a) and (b) Directive 2009/24/EC.

[49] As pointed out by Roberto Caso, 'Il conflitto tra diritto d'autore e ricerca scientifica nella disciplina del text and data mining della direttiva sul mercato unico digitale' (2020) Trento LawTech Research Paper nr. 38, available at https://ssrn.com/abstract=3533401.

[50] On Article 5(3) of the Software Directive, see, P Goldstein and PB Hugenholtz, *International Copyright. Principles, Law, and Practice (Third Edit). New York: Oxford University Press* (2013), p. 385. The decompilation exception (Article 6) adds another freedom to make reproductions and adaptations of software code when it is indispensable to obtain the interoperability information, which belongs to the "ideas" underlying a program.

contrary, since Article 5(3) of the Software Directive does not cover TDM activities beyond a research purpose, Article 4 CDSMD has clarified that reproduction and adaptation of software for TDM outside the research context is permitted, but nevertheless subject to the rightholder's reservation, possibly through the use of machine-readable means.

However, this answer would not be fully satisfactory. Indeed, under Article 5(3) of the Software Directive, the acts of "loading, displaying, running, transmitting or storing" can be exempted under the black box analysis test, which however does not extend to the translation, adaptation, arrangement and any other alteration of a computer program and the reproduction of the results. Therefore, TDM for research purposes might eventually be performed on the ground of Article 5(3) of the Software Directive but it is not clear whether it can cover any translation or adaptation.

As an alternative, research organisations and cultural heritage institutions could rely on Article 4 CDSMD. However, as already underlined, Article 4 is an imperfect solution for research purposes: not only TDM can be blocked by contracts and machine-readable means, but researchers would not be allowed to maintain the copies generated via TDM for the validation of their results (in theory, such copies should be already deleted at the stage of submission of a paper for a publication in a scientific journal).

A second systematic inconsistency relates to the relationship between Article 4 CDSMD and the "black box exception" in the Software Directive. The formulation of Article 4 might create a Schrödinger's paradox: the activity of observing, studying, and testing the functioning of a program by the lawful acquirer of the software could be limited by contract if we apply Article 4 CDSMD, and not be limited by contract if we consider the Software Directive.[51]

3.3. Technological protections

All the exceptions come with limitations: to begin with, the three-step-test will apply.[52] This means that the TDM exceptions are deemed to be narrowly detailed and interpreted ("shall be applied in certain specific cases"), cannot be in conflict with the "normal exploitation of the work" and must "not unreasonable prejudice the legitimate interest of the rightholder". As known, the European implementation of the three-step-test in Article 5(5) InfoSoc Directive is usually interpreted by the European Court of Justice restrictively.[53] It is yet to be seen whether this test could limit the application of the TDM

[51] Article 8 Software Directive establishes that Article 5(3) cannot be overridden by contract.

[52] Article 5(5) of Directive 2001/29/EC is recalled at Article 7(2) CDSMD. See, João Pedro Quintais, 'The New Copyright in the Digital Single Market Directive: A Critical Look ' (2020) 1 European Intellectual Property Review 28.

[53] See par. 58, *Infopaq International A/S v Danske Dagblades Forening* (C-5/08), Judgment of the Court (Fourth Chamber) of 16 July 2009, ECLI:EU:C:2009:465; par. 24-26, *ACI Adam BV and Others v*

exception in practice. At first sight, most traditional copyright owners, for example those of scientific or press publications, do not have a business model based on the revenues from TDM activities, and it is not likely they would be able to establish that the limited TDM exceptions conflict with the "normal exploitation" of their journals, books or newspapers. However, some publishers already offer paid-for TDM as value-added services. With the result that the test could work as an additional limit to the TDM exceptions.[54]

A second order of limitations comes indirectly from the unsatisfactory regulation of the relationship between the TDM exceptions and the protection of TPMs. Article 7(2) CDSMD establishes that the first, third and fifth subparagraphs of Article 6(4) of Directive 2001/29/EC shall apply to Articles 3 to 6 of the CDSMD. This means that TPMs are considered an important tool to protect the legitimate interests of the rightholders. However, they should ensure that users remain able to enjoy their copyright exceptions they benefit from. Rightholders are primarily entrusted with the task of adopting "voluntary measures" (not defined in the InfoSoc Directive) to balance the protection of their entitlements and the regime of exceptions and limitations. In case of *inertia* by the rightholders, Member States may kick in and establish appropriate measures.[55] However, the latter have not been consistently adopted, neither implemented, by the Member States.[56]

Therefore, if the mechanism provided for at Article 6(4) CDSMD have proven to be largely ineffective over the last 20 years, it is reasonable to question why it should work

Stichting de Thuiskopie and Stichting Onderhandelingen Thuiskopie vergoeding (C-435/12), Judgment of the Court (Fourth Chamber) of 10 April 2014, ECLI:EU:C:2014:254; par. 63, *Stichting Brein v Jack Frederik Wullems* (C-527/15), Judgment of the Court (Second Chamber) of 26 April 2017, ECLI:EU:C:2017:300.

[54] Despite several authors have argued in favour of a more flexible interpretation of the three-step-test that could prevent unbalanced restrictions to exceptions and limitations to copyright. For instance, Christophe Geiger and others, 'Declaration on a balanced interpretation of the "three-step test" in copyright law' (2008) 39 IIC 707; Jonathan Griffiths, 'The'Three-Step Test'in European Copyright Law-Problems and Solutions' (2009) Queen Mary School of Law Legal Studies Research Paper; Reto M Hilty, 'Declaration on the Three-Step Test: Where Do We Go from Here' (2010) 1 J Intell Prop Info Tech & Elec Com L 83; Martin Senftleben, 'The international three-step test: a model provision for EC fair use legislation' (2010) 1 J Intell Prop Info Tech & Elec Com L 67; Christophe Geiger, Daniel Gervais and Martin Senftleben, 'The three-step test revisited: How to use the test's flexibility in national copyright law' (2013) 29 Am U Int'l L Rev 581; João Pedro Quintais, 'Rethinking normal exploitation: enabling online limitations in EU copyright law' (2017) 6 AMI-tijdschrift v oor auteurs-, media-en informatierecht.

[55] Severine Dusollier, 'Exceptions and technological measures in the European Copyright Directive of 2001' (2003) 34 International Review of Industrial Property and Copyright Law 62.

[56] Margoni and Kretschmer (n 41) .

for TDM now.[57] In our opinion, the CDSMD has missed an important occasion for harmonising the framework and establishing in the black letter of the law a mandatory prohibition for TPMs to override the exceptions.[58]

Within this context, there is however something positive. The controversial fourth paragraph of Article 6(4) of the InfoSoc Directive, that provides for an exclusion for on-demand services to comply with the obligation to preserve the exceptions, is not expressly recalled by the CDSMD.[59] This means that the TDM exceptions for content made available online for interactive on-demand use shall benefit from the safeguard mechanisms under the first subparagraph of Article 6(4) InfoSoc.

A final point to be stressed with reference to technological protection measures is the ambiguous provision at Article 3(3) CDSMD, which states that "[r]ightholders shall be allowed 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". Recital 16 enumerates few possible examples, such as IP address validation or user authentication, which fall within the concept of access control measures. However, the Directive mentions also integrity protection mechanisms, which comprise cryptography, watermarking, etc. Those protection measures are already known, however, they will be used in another context where they will even less related to the copyright-protected works and acts.[60] Indeed, while the TPMs referred to in the InfoSoc Directive are directed to technologically protect the entitlement of the rightholders and can be limited as long as it is necessary to let the user benefit from the exceptions, the measures at Article 3(3) protect the security and integrity of the technology system (comprising servers and databases). The limitations to the deployment and use of such technical measures are not determined by the scope of the copyright exceptions, but result from a proportionality test between the technical measures selected and the objective of ensuring the security and integrity of the system. Recital 16 just recalls that such measures should "not undermine the effective application of the exception". The security measures seem to have therefore a different and complex nature, more remote, although not completely disconnected from

[57] Begoña Gonzalez Otero, 'Las excepciones de minería de textos y datos más allá de los derechos de autor: La ordenación privada contraataca' in Concepción Saiz García and Raquel Evangelio Llorca (eds), *Propiedad intelectual y mercado único digital europeo* (Tirant lo Blanc 2019).

[58] As observed by Quintais (n 52).

[59] Ted Shapiro and Sunniva Hansson, 'The DSM copyright directive: EU copyright will indeed never be the same ' (2019) 41 European Intellectual Property Review 404; Quintais (n 52) .

[60] Caso (n 49).

the copyright objective and zone of exclusivity. The abovementioned balance should be ensured with reference to other fundamental rights, such as data protection and privacy.[61]

3.4. Lawful access as a pre-condition for TDM

The TDM exceptions are not a tool designed to enable the access to content. On the contrary, the lawful access must pre-exist the TDM activities. Hence, the CDSMD leaves open the problem for those research organisations and cultural heritage institutions that do not have the means to afford paid subscriptions and for those citizens or entities that do not have "lawful access" to the content.[62]

The notion of lawful access is mentioned at Recital 14 CDSMD, where it is defined as the "access to content based on an open access policy or through contractual arrangements between rightholders and research organisations or cultural heritage institutions, such as subscriptions, or through other lawful means [...] Lawful access should also cover access to content that is freely available online". In our opinion, such provision must be read together with Recital 10, that warns against the disadvantages that research organisations might experience if the terms of the licenses over the content to which they have lawful access could exclude TDM, and Recital 17 that recognises that Member States should not provide for compensation for rightholders.[63]

Article 3 (read in conjunction with Recitals 10 and 17) might support the view that rightholders would not be able to license separately normalised datasets for the purpose of TDM and that compensation should not be provided since "any potential harm created to

[61] On the risks that TPMs in general would have on privacy and data protection, see Severine Dusollier, 'Electrifying the Fence: The Legal Protection of Technological Measures for Protecting Copyright' (1999) European Intellectual Property Review 3: 285.

[62] Hilty R and Richter H (n 41). The two commentators propose that research organisations should be allowed to cary out TDM on normalised datasets, even without having lawful access to the underlying content. See also, Geiger C, Frosio G and Bulayenko O (n 32).

[63] The drafting of Recital 17 is another example of poorly written provision in the CDSMD. From its drafting it is not entirely clear whether it refers to both exceptions or to Article 3 only. On the one hand, a systematic interpretation would lead us to conclude that the legislator exclusively referred to TDM for research purposes here: the incipit of Recital 17 refers essentially to the TDM exception carried out by research organisations and exceptions and limitations for scopes different from research are named only at Recital 18. At the same time, Recital 17 speaks of "text and data mining *exceptions* introduced by this Directive" in plural [emphasis added]. Hence, it does not seem to refer exclusively to Article 3. A logical and functional interpretation of Recital 17 should suggest that if the reason behind the waiver of compensation is the minimal harm created to rightholders, such a rationale should be extended also to the exceptions and limitations under Article 4 on a case-by-case basis (for example, if the uses under TDM are not done for commercial purposes or, even if for profit, are not performed by direct competitors).

rightholders through this exception would be minimal"[64]. In other words, where there is lawful access, TDM for research must always be permitted without additional burdens.

However, the Recitals of a directive are not binding and Recital 17, in particular, appears optional. Therefore, unless the resource is otherwise publicly available online, one can predict that this goal could be circumvented by the publishers' business models, and with unwanted effects. Publishers could simply raise the subscription fees indistinctly for all the research organisations.[65]

Recital 18 CDMSD reiterates lawful access as pre-requisite for the exercise of the exception under Article 4. Such condition clearly includes content made available to the public online as long as the rightholders have not reserved in an appropriate manner the rights to make reproductions and extractions for TDM.[66] If a website impedes the crawling of all of some of its pages via exclusion protocols or terms and conditions, anyone wishing to perform TDM cannot in principle do so. Unless, a specific authorisation is negotiated with the rightholder. The verification of the existence of any reservation of the rights on resources accessible online can be particularly tricky for the miner. In the absence of machine-readable licences, the user working on the online corpus should check one by one the terms and conditions of each website before performing TDM. This kind of scenario appears to be impractical considering that TDM is carried out with automated tools and on a variety of sources (potentially subject to different national laws).[67]

A final point needs to be stressed. The relationship between TDM and access to content is going to be a crucial knot to untangle for the research in AI domains. TDM is an enabling instrument for machine learning and artificial intelligence. Large parts of the AI innovation, at least if based on machine-learning, rely on the volume and quality of the data available to train the algorithms. If the input data are scarce, incomplete, not-well curated and not representative the resulting output will be poor and unreliable. Training dataset are fundamental to ensure the efficacy of the algorithm in relation to its proposed scope. Using TDM can also contribute to mitigate algorithmic bias and potential discrimination.[68] Several studies have demonstrated the failure of some predictive

[64] Recital 17 CDSMD.

[65] Geiger C, Frosio G and Bulayenko O (n 32), 22.

[66] See Recital 18 CDSMD.

[67] Although with reference to the French TDM provision, see Daniel Gervais, 'Exploring the Interfaces between Big Data and Intellectual Property Law' (2019) 10 *JIPITEC* <https://www.jipitec.eu/issues/jipitec-10-1-2019/4875> accessed 31 July 2020.

[68] EXEC. OFFICE OF THE PRESIDENT, PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE (2016), https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf [https://perma.cc/YN5X-KKAX].

algorithms that ended up discriminating minorities, because the training dataset was based on Caucasian white population.[69] A trend that is even more worrisome in the biomedical field, where the absence or the lack of access to representative datasets can exacerbate health disparities.[70]

A TDM exception that allows to train the algorithm with copyrighted works, permitting the reverse engineering of the software and supporting algorithmic accountability process, could help reduce the bias effect.[71] However, the concept of lawful access, enforced by contract, is likely to hinder such an ambitious (and needed) purpose in the AI field.

3.5. Partial frustration of the initial legislative intent

The main intent behind the introduction of the TDM exceptions has been the creation of a harmonised framework, allowing all the actors, including researchers, to have clear guidance on what they are allowed to do or not, also in a cross-border context. However, the objective of the proposal might not be completely fulfilled in practice. For instance, Member States remain free to maintain the already enacted TDM exceptions or to adopt broader provisions compatible with those established in Directive 96/9/EC and 2001/29/EC (Article 25 CDSMD) as long as they do not limit the scope of the mandatory exceptions or limitations provided for in the CDSMD (Recital 5).

The following table summarizes the comments and criticisms made above.

	The good	**The bad**	**The weird**
TDM scope of application	The Directive specifies that: - TDM does not always involve a conflict with the IP framework (data and facts are not protected by copyright)		Data are indirectly protected by the database rights (copyright and sui generis), contracts and/or TPMs.

[69] For an overview, see Eli M. Cahan and others, 'Putting the data before the algorithm in big data addressing personalized healthcare' (2019) 2 npj Digital Medicine 78.

[70] Heidi Ledford, 'Millions of black people affected by racial bias in health-care algorithms' (2019) 574 Nature 608.

[71] Amanda Levendowski, 'How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem' (2018) 93 Wash L Rev 579.

	- TDM does not always involve reproduction. When it involves reproduction, the latter may be covered by Article 5(1) InfoSoc Directive.		
Nature of the exceptions	The two TDM exceptions are mandatory		
Beneficiaries	The exceptions and limitations at Article 4 are available to anyone	Art 3 exception is available only to research organisations and cultural heritage institutions for research purposes	Uncertainty remains as to the scope of art. 3, because Recital 12 refers to research in natural and human sciences. Applied research might for instance not benefit for the exception.
Object		Article 3 does not include an exception to the rights of reproduction and translation in the Software Directive	Article 3 provides an exception to the press publisher right (Article 15 CDSMD), but it is not clear how research organisations' or cultural heritage institutions' reuse of news could fall under the press publishers right against *information society services providers*.
Contractual limitations to TDM	Article 3 exception for research cannot be overridden by contract	Article 4 exceptions and limitations can be overridden by contract and machine-readable means	
TPM protection	The provision of the first subparagraph of Article 6(4) InfoSoc applies where works and other subject matter are made available to the public through on-demand services	No clear mandatory measure against TPMs preventing users to enjoy the exceptions. Users cannot circumvent TPMs	
Security and integrity measures			Unclear nature of the measures to be introduced by the rightholders to ensure the security and integrity of the networks and databases.

			The provision is established only with reference to TDM for research (Article 3)
Pre-existing condition		Lawful access to the resource is a precondition for TDM both under Article 3 and 4	Lawful access for the TDM exception at Article 4 includes content made available to the public online, insofar as the rightholders have not reserved in an appropriate manner the rights to make reproductions and extractions for TDM. If the reservation of rights is not intelligible to the user (e.g. through machine-readable means) it is going to be difficult for the user to understand whether they can mine an online resource.
Compensation to the rightholder	Member States should not provide for compensation for rightholders as regards uses under the TDM exceptions (Recital 17)	The provision is contained in a Recital	
Harmonisation effort	The CDSMD ensures some harmonisation through the introduction of mandatory exceptions		Member States remain free to maintain the already enacted TDM exceptions or to adopt broader provisions compatible with those established in Directive 96/9/EC and 2001/29/EC as long as they do not limit the scope of the mandatory exceptions or limitations provided for in the CDSMD.

3.6. Suggestions for national transpositions

Despite the legislative intent to remove the divergence between copyright exceptions, there is a silver lining: Articles 3 and 4 serve as a minimum benchmark with whom Member States have to comply with. However, because there is still, as we will see in the next section, a great uncertainty about the scope of the reproduction and extraction rights, and whether

they cover the TDM activity at all, we claim that national laws can introduce provisions that clarify the TDM issues.

At the time of writing the CDSMD has yet to be transposed into national legislation (the deadline is 7 June 2021). It is thus desirable that, during this phase, national legislators improve the two TDM exceptions *cum grano salis*. This is actually possible: Article 25 CDSMS provides that Member States might adopt "broader provisions, compatible with the exceptions and limitations provided for in Directives 96/9/EC and 2001/29/EC, for uses or fields covered by the exceptions or limitations provided for in this Directive"[72] "as long as they do not limit the scope of the mandatory exceptions or limitations"[73] introduced with the CDSMD.

Therefore, when transposing the CDSMD, Member States could create "regulatory sandboxes" and enact more favourable rules for pursuing public or constitutional interests, such as the freedom of scientific research. Just to mention a few possible instances, TDM can play a decisive role for: independent researchers who are not affiliated to universities or cultural heritage institutions; journalists researching sources and checking the authenticity of a news item; policy makers testing draft policies and new legislative interventions. Furthermore, if one goal of the TDM exceptions is to create the conditions for not putting national businesses at a competitive disadvantage compared with other international players, Members States might then consider granting a broader exception to commercial private actors, which play a decisive role in research and development in the field of AI. Furthermore, during the national transposition phase, Member States should remove any doubt about the possibility to share the TDM-generated *corpora* with other researchers, at least for verification purposes[74] and specifically allow TDM on software for research purposes.

4. Assessing the infringement of the reproduction right and of the extraction right

Considering the TDM exceptions in the CDSMD, one might legitimately wonder whether the legislative intervention was *too little, too late*. Regrettably, the CDSMD is a missed opportunity to regulate the TDM phenomenon in a clear and effective way. From the previous analysis, it appears that some conditions for TDM activity are too narrow (in terms of its beneficiaries, Article 3 is quite limited) or too vague (the pre-condition of "lawfully accessible" content gives too much room for limiting TDM initiatives by contract or by

[72] Article 25 CDSMD.

[73] Recital 5 CDSMD.

[74] Along the lines of Section 60d of the *Urheberrechtsgesetz*.

TPMs). This might have consequences. For instance, it appears from a survey conducted after the adoption of the CDSMD that publishers are using TPMs to block the access to various sources and databases which are useful for the research conducted at research organisations having paid for accessing the online databases.[75]

Rather than proposing some tweaks to the CDSMD exceptions (in particular to Article 3), this section supports a purpose-oriented reading of the conditions for infringing the reproduction and extraction rights that exonerates the acts of copying or extracting made during the TDM process.

4.1. A purposive test for assessing the infringement of the reproduction right

By searching for some legal certainty, guarded by statutory exceptions, the Directive seems to confirm that TDM, even with the clarifications in Recital 9, is a copyright-relevant (and database right-relevant) activity.[76] As argued elsewhere,[77] a critical clarification about the conditions for an infringement of the right of reproduction would have represented a viable and less costly alternative to follow. This doctrinal proposal, coupled with a normative intervention on specific provisions of the copyright/sui generis right framework (notably a serious rethinking of the regulation of TPMs) could fill the gaps of the brand-new TDM exceptions.

That copyright might extend to the TDM process contradicts its own rationale: the promotion and dissemination of information and knowledge. When it appeared more than three centuries ago, copyright was not only intended to remunerate authors (and publishers) for creating (and disseminating) works, but was also aimed at favouring public learning. This is not only true for the copyright systems rooted within an incentive-based view; the same *raison d'être* is at the origin and core of the continental *droit d'auteur* systems as well.[78] Copyright, in fact, on both sides of the Atlantic, does not cover facts (including correlations), information or other elements considered as non-protectable ideas.

---

[75] See https://libereurope.eu/blog/2020/03/10/tdm-technical-protection-measures/. According to the survey on TDM barriers conducted by libraries associations, "[a]ctions taken by publishers included 1) suspension of campus-wide access to paid for electronic subscriptions 2) threats to cut off access to content unless TDM was stopped 3) technically limiting downloads to one document only 4) a request for additional payments and 5) the introduction of CAPTCHA technology to frustrate TDM".

[76] However some exceptions just clarify that an operation or conduct does not constitute an infringement, without implying that the operation or conduct falls within the scope of the right in the first place. This applies for instance for the exceptions justified by the freedom of expression.

[77] These thoughts were firstly elaborated in Alain Strowel, 'Reconstructing the Reproduction and Communication to the Public Rights: How to Align Copyright with Its Fundamentals' in P. Bernt Hugenholtz (ed), *Copyright Reconstructed Rethinking Copyright's Economic Rights in a Time of Highly Dynamic Technological and Economic Change* (Wolters Kluwer 2018).

[78] See the review of the history and principles of copyright in both legal traditions: Alain Strowel, *Droit d'auteur et copyright, Divergences et convergences* (Bruylant et Paris, L.G.D.J. 1993). More recently, Alain

At the same time, it cannot be denied that digital technologies and the ubiquity of copies have seriously expanded the "footprint" of copyright without help from the legislator[79]: it has just happened through the extension of the existing synthetic and technology-neutral notion of reproduction.[80] However, the development of the digital world has raised the expectations for economic return for rightholders in a manner that, in our opinion, is not commensurate to the economic fundamentals that apply to digital uses, especially on the internet. The explosion of digital copies cannot translate into the same increase of value that would be generated by additional non-digital copies: many digital copies have no value in themselves and are solely the tools or basic ingredient for value-adding services (for instance, copies made for detecting meaningful occurrences of words). The technological-neutral notion of reproduction is ill-suited to the digital realm because of the massive technology-driven copies that lack any commercial value and the inflated hopes generated by the expectations of rights owners who tend to believe that any reproduction, even without economic reason, should be covered by copyright.

Therefore, a way to put this drift back on the (copy)right track could pass through a qualitative test, i.e., distinguishing between the copies which are infringing from those which do not infringe.

As known, a communication must be addressed 'to a public' in order to fall under copyright.[81] The definition of reproduction, in contrast, does not refer to any public and to the possibility for it to access the work.[82] However, that does not mean that no public or audience is required for an infringement of the reproduction right. We have argued

---

Strowel, 'Droit d'auteur et copyright. Convergences des droits, régulation différente des contrats', *Mélanges en l'honneur du Professeur André Lucas* (LexisNexis 2014).

[79] In many jurisdictions, the notion of reproduction came to cover many ephemeral and transient copies just by application of the existing law. Sometimes, as in Article 2 of the InfoSoc Directive, the legislator has made it clear that the reproduction right encompasses 'temporary' reproductions "by any means and in any form, in whole or in part", thus expanding copyright's reach in the technological environment.

[80] The principle of technological neutrality relies on the text of Article 2 InfoSoc Directive, which refers to copies 'by any means and in any form'. The technology-neutral definition of reproduction has allowed its application outside the printing press context where it originated. In parallel to the expansion of the notion of work, which applies to new products of technology and creativity (films, computer games, etc.), the notion of copy was able to adapt to new technologies so that it now applies to new ways of making copies (for example, to various digital copies, but also non-digital copies such as canvas transfer). In that sense, the legal notion is neutral or independent from the technology involved. There are, however, other ways to define the principle of technological neutrality. For a short and recent discussion of its meanings, see Kendrick Lo, 'What is Technological Neutrality (in Copyright) Anyway? Revisiting CBC v. SODRAC' (CanLII Connects, 1 August 2017) <http://canliiconnects.org/en/commentaries/46245> accessed 31 July 2020.

[81] Article 3 InfoSoc Directive.

[82] Article 2 InfoSoc Directive.

elsewhere[83] that EU law has not yet defined the test for an *infringement* of the reproduction right (although it contains a statutory definition and a case law delineation of the reproduction right).[84] This gap permits to design and support an infringement test requiring that for the reproduction right to be infringed the work should be used *as a work* and perceived *as a work* by a public.

Such use as a work does not exist in the case of TDM nor in other cases involving copying for checking conduct (e.g., to identify plagiarism) or for deriving information (e.g., on patterns or trends). As put by the 2nd Circuit in the *Authors Guild v. Google, Inc.* U.S. case 'the purpose of [defendant]'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).[85]

In addition, this purposive analysis of the act of copying leading to a non-infringing conclusion mirrors the analysis under the fair use doctrine in U.S. law (Article 107 U.S. Copyright Act) which leads to the conclusion that highly transformative uses are excluded from the scope of exclusivity.[86] That TDM use can be exempted where the copyright system envisages an open clause is an additional argument for leaving the making of intermediate copies, such as those made during the TDM process, out of copyright's reach. In the absence of a fair use clause, the requirement of "use as a work" in the infringement test can help reaching the same outcome in the EU. Indeed, when acts of reproduction are carried out for the purpose of search and TDM, the work, although it might be reproduced in part, is not used as a work: the work only serves as a tool or data for deriving other relevant

[83] See Strowel (n 79).

[84] On the contrary, the EU legislation specifically defines the conditions for a design or trade mark infringement, and the criterion for infringement refers to the confusing effect or impression on a consumer or a user. For example, Article 10 of the Design Regulation (Council Regulation (EC) No. 6/2002 of 12 December 2001 on Community designs, OJ No. L 3/1, 5 January 2002) provides that 'The scope of the protection conferred by a Community design shall include any design which does not produce on the informed user a different overall impression'. There is no similar provision on infringement in the copyright directives. However, *Infopaq* and other decisions of the CJEU could be read as providing some indications as to when there is a copyright infringement. This case law requires to assess the substantial similarities from the point of view of a public. See also Julien Cabay, 'L'objet de la protection du droit d'auteur: Contribution à l'étude de la liberté de création' (DPhil thesis, Université Libre de Bruxelles 2016) 299. However, this case law leaves enough room for adding (and proposing) a requirement to use the work as a work.

[85] *Authors Guild v. Google, Inc.* No. 13-4829-cv (2d Cir. Oct. 16, 2015). On April 18, 2016, the U.S. Supreme Court denied the petition for a writ of certiorari, leaving the Second Circuit ruling in Google's favour intact. To make available parts of the corpus of books, Google has scanned the digital copies and established a publicly available search function, the Ngram tool.

[86] However, Lemley and Casey warned that courts might not treat machine copying as fair use, since machine learning systems "rarely transform the databases they train on; they are using the entire database, and for a commercial purpose at that". Mark A. Lemley and Bryan Casey, 'Fair Learning' (forthcoming 2021) Texas Law Review, available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3528447> accessed 31 July 2020.

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.

Even if the new Article 4 CDSMD more clearly allows for some commercial activities to be exempted, the analysis based on "the use of the work as a work" requirement for copyright infringement remains useful to assess the copyright-relevance of the acts performed during the TDM process. In addition, the infringement threshold remains necessary to address other types of copying for the purpose of providing information which are not expressly exempted. These include copies made for checking mistakes and plagiarism, copies to use or repair a protected work with a utilitarian function, non-transient copies made on proxy servers, smart disclosure systems and many other intermediate and non-expressive uses that could be considered.

When datasets which are original in the selection or arrangement of elements or which contain original works are mined so as to find correlations, patterns and links between information points, the dataset or its original elements are not used as works for a public. This is why searching for correlations or information in the many available online sources should not constitute a copyright infringement, even if partial copies are made. For instance, the BlueDot project, which warned about the Coronavirus outbreak well before the WHO did, analysed "a variety of information sources, including chomping through 100,000 news reports in 65 languages a day"[87] and, after comparing the information with flight records, was able to identify patterns between the propagation of the virus and the travelling routes of people. Similarly, no infringement is committed when TDM tools are used to mine, copy, and compare original press articles so as to identify the spreading of disinformation on the social networks.

A TDM activity might also be conducted within datasets that qualify as databases under the Database Directive (Article 1(2)). The discussion around the permissibility of TDM in this case is framed within the database rights and the limitations and exceptions to the various rights contained in the Database Directive, as will be discussed in the next section.

4.2. A purposive interpretation of the extraction right

It is not only the reproduction right within copyright the infringement of which can be interpreted as limited to the output-copies. One can propose a similar reading of the

[87] Mark Prosser, 'How AI Helped Predict the Coronavirus Outbreak Before It Happened' (SingularityHub, 5 February 2020) <https://singularityhub.com/2020/02/05/how-ai-helped-predict-the-coronavirus-outbreak-before-it-happened/> accessed 31 July 2020. This example is quoted in Sean Flynn, Christophe Geiger, João Pedro Quintais, Thomas Margoni, Matthew Sag, Lucie Guibault, and Michael Carroll, 'Implementing User Rights for Research in the Field of Artificial Intelligence: A Call for International Action' (2020) 7 *EIPR* 393.

extraction right under the Database Directive, arguing that the TDM acts are not to be intended as infringing the sui generis (database) right.

The conditions for infringing the extraction right are not expressly mentioned in the Database Directive (in the same way that the assessment test for the infringement of the reproduction right is ignored by the EU legislator). To neglect the criteria for assessing an infringement of the extraction right would constitute a similar mistake.

The infringement test can be reconstructed at the light of the purpose of the right as interpreted in CJEU case law. For the CJEU, "the objective" is to protect the maker of the database "against the unauthorised *appropriation of the results of that investment* by acts which involve, in particular, *the reconstitution* by a user or a competitor of that database or a substantial part of it at a fraction of the cost needed to design it".[88] If TDM could be prohibited, the extraction right would go beyond "the appropriation of the results" as TDM is a use of the data included in the database, it does not involve "the reconstitution" of (part of) the database in a competing product. TDM is to be considered as a form of "consultation" of the content of a database, and as underlined by the CJEU, the protection granted by Article 7 "does not, however, cover consultation of a database".[89] For the Court, "where the maker of a database makes the contents of that database accessible to third parties, even if he does so on a paid basis, his sui generis right does not allow him to prevent such third parties from consulting that database for information purposes".[90]

The notion of "appropriation" used by the CJEU should not lead to a "property" protection: this is a common IP misunderstanding, derived from the wrong analogy with tangible property. "Propertization" of an IP right happens when it is solely conceived as a *right on* a subject matter, while all rights have to be seen and grasped as rights defining the relations between the rights owner and third parties. Those relations are always determined in the infringement test. The sui generis right is based on a misappropriation or unfair competition rationale (Recital 39 of the Database Directive indicates that it "seeks to safeguard the position of makers of databases against misappropriation"). In an interesting opinion, Advocate-General Szpunar has recently reiterated that "the *sui generis* right provided for in Article 7 of Directive 96/9 has as its objective to protect database makers against the creation of parasitical competing products"; in addition, Recital 47 indicates

[88] Case 304/07 *Directmedia Publishing GmbH v. Albert-Ludwigs-Universität Freiburg* [2008] ECLI:EU:C:2008:552, para 33 (emphasis added), relying on the earlier decisions in Case 46/02 *Fixtures Marketing v Oy Veikkaus Ab* [2004] ECLI:EU:C:2004:694, para 35; Case 203/02 *The British Horseracing Board and Others v William Hill Organisation Ltd* [2004] ECLI:EU:C:2004:695, paras 32, 45, 46 and 51; Case 338/02 *Fixtures Marketing Ltd v Svenska Spel* AB [2004] ECLI:EU:C:2004:696, para 25; and Case 444/02 *Fixtures Marketing v Organismos prognostikon agonon podosfairou* [2004] ECLI:EU:C:2004:697, para 41.

[89] *Directmedia Publishing GmbH v. Albert-Ludwigs-Universität Freiburg* (n 100), para 51 relying on *The British Horseracing Board and Others v William Hill Organisation Ltd* (n 100), para 54.

[90] *Directmedia Publishing GmbH v. Albert-Ludwigs-Universität Freiburg* (n 100), para 53.

that this right "must not at the same time have the effect of preventing the creation of innovative products which have added value".[91] Therefore, according to the AG opinion, to conclude that there is an infringement, "the national courts should therefore verify not only whether the extraction or reutilisation of the whole or a substantial part of the contents of a database has taken place and whether it is shown that there has been a substantial investment in either the obtaining, verification or presentation of those contents, but also whether the extraction or reutilisation in question constitutes a risk to the possibilities of recouping that investment".[92] In other words, "the condition that there be an adverse effect on the investment of the maker of a database" should be factor in the infringement analysis.

In the case of a database, the propertization of the sui generis right resulting from the absence of a correct analysis of the infringement conditions must be opposed more cogently than in other cases as it leads to the creation of a property right on data. And this is precisely what the Directive seeks to avoid: "the right to prevent unauthorized extraction and/or re-utilisation does not in any way constitute an extension of copyright protection to mere facts or data" (Recital 45). If TDM applied to databases could be prevented by the extraction right, it would amount to a form of appropriation of mere facts and data, contrary to the Directive's objective which, according to Recital 46, "should not give rise to the creation of a new right in the works, data or materials themselves".

In addition, one has to consider whether TDM could fall among the lawful uses recognized by the Database Directive at Articles 6(1) and 8.[93] These rights are expressly protected against conflicting contractual provisions (Article 15 Database Directive). Article 6(1) allows the lawful user to make a copy of the database in order to access the contents or to allow a "normal use" of the same. It would be coherent with the rationale of the Directive to consider TDM as performing a normal use of the database. Interestingly, the "normal use" argument was followed in the national track leading to the CJEU *Ryanair* case.[94] The Dutch appeal court considered that the comparison of the flights prices by an online intermediary involved a reproduction of the information from the Ryanair website, but it amounted to a "normal use" of the database protected by copyright, thus any contrary contractual provision was deemed unenforceable.[95]

---

[91] Opinion of Advocate General Szpunar, 14 Jan. 2021 in Case C-762/19 *SIA 'CV-Online Latvia' v. SIA 'Melons'*, para 40.

[92]Idem, para 47 (see also para 59).

[93] See, Hilty and Richter (n 47).

[94] Case 30/14, *Ryanair Ltd v PR Aviation BV* (n 29), para 21, referring specifically to the Dutch Copyright Law.

[95] However, the CJEU noted that the existence of the sui generis right was not proven in the case: as a consequence, the prohibition of contractual overriding did not apply.

Moreover, in many cases, the TDM tool would likely extract and mine information from an insubstantial part of a database, without conflicting with the normal exploitation of the database or unreasonably prejudicing the legitimate interests of the maker or the author of the works or subject-matter contained in the database. Article 9(b) Database Directive also shows that the extraction right is not intended to control the TDM or other data analysis processes, only the output in the form of a publication, as the exception "in the case of extraction for the purposes of [...] scientific research" requires that "the source is indicated" (this condition arguably will only apply to the publication resulting from the extraction). Similarly, the way the research exception is designed for copyright under Article 5(3)(a) InfoSoc Directive shows that the EU legislator has not considered that copyright could apply to the copies made during the process of consultation. The exception under Article 5(3)(a) only targets the output-copies, those that are addressed to the public as it provides with a special requirement to indicate "the source, including the author's name" (vice versa, process-copies such as those during the TDM activity are purely internal, and thus cannot be subject to a recognition by a human and to an obligation to credit the source).

In sum, the Database Directive leaves some room to mine text and data, as TDM does not constitute a misappropriation of the contents leading "to reconstituting a substantial part of the content". It is a form of consultation that cannot, and should not, be prohibited or impeded..

5. TDM exceptions and the purposive interpretation of the rights of reproduction and extraction in context: remaining issues

Whether we follow the purposive interpretations of the rights of reproduction/extraction or we embrace the (limited) TDM exceptions, some inherent obstacles remain. First of all, the doctrinal and legislative solutions work as long as the object of the mining is protected by copyright or the sui generis right. As established in *Ryanair*, if a database is not protected according to the Database Directive, the owner can contractually limit the use of the database (restricting the "legitimate uses" that the beneficiary could have enjoyed under the IP regime).[96] Therefore, from a practical perspective, it might even be counterproductive for the producer having acquired the copyright or for the maker of a database to claim their exclusive rights: the contract offers a more flexible and efficient solution to enforce their economic interests.

Private autonomy, however, is not absolute and can be limited, especially when it results in the abuse of rights.[97] For example, the principle of transparency in consumer and data protection law might prevent a "dictatorship" of contract, and legitimize TDM in smart

[96] *Ryanair Ltd v PR Aviation BV* (n 29), para 39.

[97] For a comparative overview of the doctrine of 'abus du droit' see Jan Peter Schmidt, 'Good Faith and Fair Dealing' in Nils Jansen, Reinhard Zimmermann (eds) *Commentaries on European Contract Laws* (Oxford University Press, 2018).

disclosure systems.[98] If the latter are meant to automatize the reading and analysis of contractual terms to spot potential risks or unfair conditions, there is no reason to impede the crawling of the webpage hosting the contract. However, this kind of analysis entails a case-by-case evaluation and, in most of the circumstances, it will be a matter of national contract law. As a result, solutions may vary and a TDM conducted over cross-border resources would entail huge analysis and transaction costs.

A second order of limitations comes from trade secret protection. The owner of the AI tool might refuse access to the training set, output data, and/or the algorithm, thus impeding the mining of text and data (including the code). This issue is of course not addressed in the CDSMD, but it is closely linked with the TDM debate and its possible use to verify and to explain the 'black box' nested in many AI tools relying on machine learning. TDM can help build an AI system, but some form of data retrieval and analysis can as well work as an ex-post tool to implement the explainability requirement.[99] However the data used to train algorithms will often: (i) be secret, (ii) have an economic value because it is secret, and, (iii) be the subject of reasonable steps to keep it secret. As a result, such data may qualify for trade secret protection under Article 2 Trade Secrets Directive.[100] Sometimes, private developers of AI tools, such as the COMPAS system used to assess the likelihood of recidivism in criminal cases, claim that the details on how their tools work are trade secrets and refuse granting access to the confidential information.[101] This access issue might mirror the problem of lawful access to copyright works for TDM, such as in the case of the BlueDot project.[102] Further research is needed to conclude that a carve-out in the trade secrets protection should be added in order to allow some data analysis and reduce the risk of bias or discrimination. This could be an opportunity to align the TDM and research exceptions in different IP fields.

---

[98] Ducato and Strowel (n **Error! Bookmark not defined.**).

[99] This requirement that the processes for designing the AI tool need to be transparent and its decisions explainable to the extent possible is underlined in most ethical guidelines on AI. See, for example, High-Level Expert Group on AI, 'Ethics Guidelines for Trustworthy Artificial Intelligence' (2019) <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai> accessed 31 July 2020.

[100] Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure, OJ L 157, 15.6.2016, pp. 1–18.

[101] Rebecca Wexler, 'Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System' (2018) 70 *Stanford Law Rev*iew 1343, referring among other to the case State v. Loomis, 881 W.W.2d 749, 760 (Wis. 2016), cert. denied, 137 St. Ct. 2290 (2017).

[102] See also Flynn, Geiger, Quintais, Margoni, Sag, Guibault, Carroll (n 89).

6. Conclusions

The main reason behind the introduction of the TDM exceptions in the CDSMD was the reduction of legal uncertainties and the diverging national implementations of existing exceptions (including those for research, private copying, temporary reproduction), in order to remove possible "interpretative" obstacles for European research organizations. In this paper, we have raised several doubts about the potential of the newly crafted TDM exceptions to reach a fair balance between the promotion of innovation in research (including outside research organizations) and the interests of some rightholders or "owners" of datasets.

As seen above, a purposive analysis of the conditions for infringing the reproduction and extraction rights might actually solve many of the shortcomings and loopholes we have identified. Such an interpretation can still be followed and applied by courts even when the TDM exceptions are implemented in national copyright and database laws.

All this said, we do not propose throwing the baby out with the bathwater. The phase of national transposition might indeed represent a valuable occasion for Member States to create a more favourable TDM environment, to improve the CDSMD exceptions (and potentially to induce more competition that would attract researchers and businesses relying on the new data analysis methods). In the context of scientific research, it would be desirable that this discussion on the implementation of the exceptions be coupled with a critical assessment of TDM in the context of Open Access for scientific publications and Open Data for Open Science. Such policies, in particular, should expressly prevent private ordering measures (contracts and TPMs, for instance) restricting TDM on the content of databases and TDM-generated corpora.

Bibliography

Akaichi J, Dhouioui Z and Pérez MJL-H, *Text mining facebook status updates for sentiment classification* (IEEE 2013)

Cabay J, *L'objet de la protection du droit d'auteur: Contribution à l'étude de la liberté de création* (2016) PhD dissertation, Université Libre de Bruxelles, Bruxelles, unpublished.

Cahan EM and others, 'Putting the data before the algorithm in big data addressing personalized healthcare' (2019) 2 npj Digital Medicine 78

Caso R, 'Il conflitto tra diritto d'autore e ricerca scientifica nella disciplina del text and data mining della direttiva sul mercato unico digitale' (2020) Trento LawTech Research Paper nr. 38, available at https://ssrn.com/abstract=3533401.

Caspers M and Guibault L, *A right to 'read' for machines: Assessing a black-box analysis exception for data mining* (2016)

Caspers M and Guibault L, *Baseline report of policies and barriers of TDM in Europe*, 2016

Derclaye E, *The legal protection of databases: a comparative analysis* (Edward Elgar Publishing, 2008)

Ducato R, 'Data protection, scientific research, and the role of information' (2020) 37 Computer Law and Security Review 105412

Ducato R and Strowel AM, 'Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to Machine Legibility' (2019) 50 IIC - International Review of Intellectual Property and Competition Law 649

Dusollier S, 'Electrifying the Fence: The Legal Protection of Technological Measures for Protecting Copyright' (1999) 3 European Intellectual Property Review 285

Dusollier S, 'Exceptions and technological measures in the European Copyright Directive of 2001' (2003) 34 International Review of Industrial Property and Copyright Law 62

Eickhoff C, Yubin K, White R, 'Overview of the Health Search and Data Mining (2020) Proceedings of the 13th International Conference on Web Search and Data Mining 901

Feldman R and Sanger J, *The text mining handbook: advanced approaches in analyzing unstructured data* (Cambridge University Press 2007)

Flynn S, Geiger C, Quintais JP, 'Implementing User Rights for Research in the Field of Artificial Intelligence: A Call for Action at International Level' (Kluwer Copyright Blog 2020)

Flynn S and others, 'Implementing User Rights for Research in the Field of Artificial Intelligence: A Call for International Action' (2020) 7 European Intellectual Property Review 393

Geiger C, Frosio GF and Bulayenko O, 'The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects' (2018) Centre for International Intellectual Property Studies (CEIPI) Research Paper No 2018-02

Geiger C, Gervais D and Senftleben M, 'The three-step test revisited: How to use the test's flexibility in national copyright law' (2013) 29 Am U Int'l L Rev 581

Geiger C and others, 'Declaration on a balanced interpretation of the "three-step test" in copyright law' (2008) 39 IIC 707

Geiger C, Frosio G and Bulayenko O, *The exception for text and data mining (TDM) in the proposed Directive on Copyright in the Digital Single Market: legal aspects : in-depth analysis* (European Parliament 2018)

Gervais D, 'Exploring the Interfaces between Big Data and Intellectual Property Law' (2019) 10 *JIPITEC* <https://www.jipitec.eu/issues/jipitec-10-1-2019/4875>

Griffiths J, 'The 'Three-Step Test'in European Copyright Law-Problems and Solutions' (2009) Queen Mary School of Law Legal Studies Research Paper

Goldstein P, 'Copyright and Its Substitutes' (1997) 865 Wisconsin Law Review

Goldstein P and Hugenholtz P, *International Copyright. Principles, Law, and Practice* (Oxford University Press 2013)

Gonzalez Otero B, 'Las excepciones de minería de textos y datos más allá de los derechos de autor: La ordenación privada contraataca' in Saiz García C and Evangelio Llorca R (eds), *Propiedad intelectual y mercado único digital europeo* (Tirant lo Blanc 2019)

Guarino S, Trino N, Chessa A, Riotta G, 'Beyond Fact-Checking: Network Analysis Tools for Monitoring Disinformation in Social Media' in H. Cherifi et al. (eds) *Complex Networks and Their Applications VIII. COMPLEX NETWORKS 2019. Studies in Computational Intelligence* (Springer 2019), 436

Hearst MA, 'Text Data Mining' in Mitkov R (ed), *The Oxford Handbook of Computational Linguistics* (Oxford University Press 2003)

Hilty R and Richter H, 'Position Statement of the Max Planck Institute for Innovation and Competition on the Proposed Modernisation of European Copyright Rules Part B Exceptions and Limitations (Art. 3–Text and Data Mining)' (2017)

Hilty RM, 'Declaration on the Three-Step Test: Where Do We Go from Here' (2010) 1 J Intell Prop Info Tech & Elec Com L 83

Hugenholtz PB, 'Copyright, contract and code: What will remain of public domain?' (2000) 26 Brooklyn Journal of International Law 77

Hugenholtz PB, 'The new database right: Early case law from Europe' (2002) Int'l Intell. Prop. L. & Pol'y 70- 7: 1

Hugenholtz PB, *The New Copyright Directive: Text and Data Mining (Articles 3 and 4)* (Kluwer Copyright Blog 2019)

Jothi N, Nur'Aini A, Wahidah H, 'Data Mining in Healthcare – A Review', Procedia Computer Science, Volume 72, 2015, p. 306-31

Kaplan B, 'Selling health data: de-identification, privacy, and speech' (2015) 24 Cambridge Quarterly of Healthcare Ethics 256

Ledford H, 'Millions of black people affected by racial bias in health-care algorithms' (2019) 574 Nature 608

Lemley M. Casey B, 'Fair Learning' (forthcoming 2021) Texas Law Review, available at SSRN at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3528447.

Levendowski A, 'How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem' (2018) 93 Wash L Rev 579

Lo K, 'What is Technological Neutrality (in Copyright) Anyway? Revisiting CBC v. SODRAC' (CanLII 2017)

Margoni T and Dore G, 'Why We Need a Text and Data Mining Exception (But it is Not Enough)' (Zenodo 2016)

Margoni T and Kretschmer M, 'The Text and Data Mining exception in the Proposal for a Directive on Copyright in the Digital Single Market: Why it is not what EU copyright law needs' CREATe Blog

Montagnani ML and Aime G, 'Il text and data mining e il diritto d'autore' (2017) AIDA 376

Poort J, 'Borderlines of Copyright Protection: An Economic Analysis' in Hugenholtz PB (ed), *Copyright Reconstructed: Rethinking Copyright's Economic Rights in a Time of Highly Dynamic Technological and Economic Change* (Wolters Kluwer 2018)

Prosser M, 'How AI Helped Predict the Coronavirus Outbreak Before It Happened' (SingularityHub 2020).

Quintais JP, 'Rethinking normal exploitation: enabling online limitations in EU copyright law' (2017) 6 AMI-tijdschrift v oor auteurs-, media-en informatierecht

Quintais JP, The New Copyright Directive: A tour d'horizon – Part II (of press publishers, upload filters and the real value gap) (KluwerCopyrightBlog 2019)

Quintais JP, 'The New Copyright in the Digital Single Market Directive: A Critical Look ' (2020) 1 European Intellectual Property Review 28

Raghupathi W, 'Data mining in health care' in S P Kudyba (ed), *Healthcare informatics: improving efficiency and productivity* (CRC Press 2016), 353

Rosati E, 'An EU Text and Data Mining Exception for the Few: Would it Make Sense?' (2018) 6 *Journal of Intellectual Property Law & Practice* 429

Rosati E, 'Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and its Role in the Development of AI Creativity' (2019) 2 *Asia Pacific Law Review* 198

Schmidt JP, 'Good Faith and Fair Dealing' in Nils Jansen, Reinhard Zimmermann (eds) *Commentaries on European Contract Laws* (Oxford University Press, 2018).

Senftleben M, 'The international three-step test: a model provision for EC fair use legislation' (2010) 1 J Intell Prop Info Tech & Elec Com L 67

Shapiro T and Hansson S, 'The DSM copyright directive : EU copyright will indeed never be the same ' (2019) 41 European Intellectual Property Review 404

Shivhare SN and Saritha SK, 'Emotion detection from text documents' (2014) 4 International Journal of Data Mining & Knowledge Management Process 51

Stamatoudi IA, 'Text and data mining' in Stamatoudi IA (ed), *New Developments in EU and International Copyright Law* (Wolters Kluwer 2016)

Strowel A, *Droit d'auteur et copyright, Divergences et convergences* (Bruylant et Paris, L.G.D.J. 1993)

Strowel A, 'Droit d'auteur et copyright. Convergences des droits, régulation différente des contrats', *Mélanges en l'honneur du Professeur André Lucas* (LexisNexis 2014)

Strowel A, 'Reconstructing the Reproduction and Communication to the Public Rights: How to Align Copyright with Its Fundamentals' in Hugenholtz PB (ed), *Copyright Reconstructed Rethinking Copyright's Economic Rights in a Time of Highly Dynamic Technological and Economic Change* (Wolters Kluwer 2018)

Strowel A, 'Intellectual Property Strengthened by the Court of Justice Interpretation of Article 17(2) of the EU Charter of Fundamental Rights' in Pollicino O, Riccio GM, Bassini M (eds), *Copyright and Fundamental Rights in the Digital Age* (Edward Elgar 2020).

Tsao W-K and others, 'A data mining approach to face detection' (2010) 43 Pattern Recognition 1039

Triaille JP, de Meeûs d'Argenteuil J and de Francquen A, *Study on the legal framework of text and data mining (TDM)*, March 2014

Wachter S and Mittelstadt B, 'A right to reasonable inferences: re-thinking data protection law in the age of big data and AI' (2019) Columbia Business Law Review, https://osf.io/preprints/lawarxiv/mu2kf/

Wexler R, 'Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System' (2018) 70 *Stanford Law Rev*iew 1343