The Insufficiency of the EU's Text and Data Mining Exceptions for Using Artificial Intelligence

Maryna Manteghi[1]

Abstract

This article examines how the EU legal frameworks on copyright, database, and computer programs restrict the application of text and data mining technology (TDM) in the European Union (the EU). The paper analyzes the sufficiency of TDM exceptions of the EU Directive on copyright and related rights in the Digital Single Market (the DSM Directive) in the context of the development of Artificial Intelligence (AI). The paper indicates that national transposition is one of the possible ways to improve the TDM exceptions. The EU Member States had until 7 June 2021 to transpose the DSM Directive into their national laws. However, the implementation process appears to be far from complete. Therefore, the article considers the positions of the EU Member States, which have already implemented the new legal framework into their national laws, regarding the scope of TDM exceptions. In particular, the paper analyses whether issues that have been unclear since the adoption of the DSM Directive have been covered adequately through the transposition process. Finally, this paper puts forward recommendations on how it could be implemented to unlock the maximum potential of TDM and AI research in the EU.

Keywords: Text and data mining, copyright, database right, DSM Directive, TDM exceptions, AI development

[1]Maryna Manteghi
PhD Candidate, Faculty of Law, University of Turku, Turku, Finland; e-mail: maryna.manteghi@utu.fi

1. Introduction

Nowadays, access to information has become a greater public demand. Although the potential value of Big Data[2] is colossal, it has become too much to manage and analyze it effectively. In this regard, AI can be seen as a powerful instrument, which enables faster access to and processing of a huge amount of digital data.[3] Practically, the development and functioning of AI rely on machine learning techniques,[4] such as TDM which is understood as an automatic analytical process aimed at analyzing large quantities of digital data in order to generate new information such as patterns, trends, and correlations.[5] AI offers significant opportunities in a wide range of areas[6] and yet at the same time, raises many concerns due to its potential to undermine and violate fundamental rights protections.[7] For instance, the use of AI may threaten the right to privacy and data protection[8] as the analysis of data using AI technologies may reveal private (sensitive) information about individuals. Moreover, AI-powered surveillance, such as facial recognition, if used in public spaces to identify, for instance, participants of a demonstration may have an ill effect on the right to freedom of expression and the freedom of assembly.[9] To address the challenges of AI and alleviate or prevent harms caused by certain uses of innovative technologies, the European

[2] Big Data can be defined as an "extremely large data sets that have grown beyond the ability to manage and analyze them with traditional data processing tools." Ohlhorst F, "Big Data Analytics: Turning Big Data into Big Money" (2012) Wiley & Sons, New Jersey, p. 10.

[3] AI is defined as a "software that is developed with one or more of the techniques and approaches…and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with." *See* Art. 3 (1) of the Artificial Intelligence Act (AIA)

[4] Surden H, "Machine learning and law: An overview" in: Vogl R (ed), Research Handbook on Big Data law, Edward Elgar Publishing (2021), p. 171.

[5] Art. 2 (2) of the DSM Directive. On the technical aspects of TDM *see*, for example, Jusoh S and Alfawareh H "Techniques, Applications and Challenging Issue in Text Mining" International Journal of Computer Science Issues (2012) 9(6):431-436.

[6] AI has a great potential for promoting and protecting human rights such as the right to information (Art.11 of the EUCFR) and the freedom of research (Art. 13) by facilitating an access to a vast amount of data. For example, the analysis of a huge number of videos is important for research in meteorology and police forensics; an automatic summarization of courts' decisions and a document automation (contracts drafting) are important in the field of legal research; an analysis of the content made available through social platforms may be useful for anthropologists in their research of cultural phenomena (See Hargreaves I et al. "Standardization in the area of innovation and technological development, notably in the field of Text and Data Mining: report from the expert group" Luxembourg: Publications Office of the European Union (2014), p.10.

[7] Rodrigues R "Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. Journal of Responsible Technology (2020) 4: 1-12, p.1.

[8] Art. 8 of the EUCFR. See also the EU Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, OJ L 119, 4.5.2016, p. 1–88.

[9] Art. 12 of the EUCFR.

Commission has recently proposed a set of regulations related to the Artificial Intelligence Act[10] (AIA).

The long-awaited draft regulation on AI uses a risk-based approach and sets a range of obligations depending on whether the AI systems pose a low, medium, or high risk while establishing a list of completely prohibited AI.[11] The European Commission attempts to present a balancing act by normatively capturing rapidly developing technologies and at the same time not constraining the development of innovation in the EU.[12] The AIA aims to ensure more legal certainty around AI to facilitate investment and innovation in AI.[13] Although the European Commission extensively addresses the regulation and classifications of AI technologies, the proposed regulation completely disregards issues related to another area of AI's concerns, in particular the current EU intellectual property (IP) law. Whilst the European Commission acknowledges that the development of AI technologies should not be unreasonably constrained or impeded, the AIA ignores the apparent obstacles that copyright or other IP rights may pose to such development. It appears to contradict the EU Parliament's intention to establish and maintain an effective IP system, which would fit for the digital age.[14]

As previously noted, the development and functioning of AI depend on data that is usually extracted by means of TDM. In this regard, TDM may involve acts protected by copyright or the *sui generis* database right. Therefore, if there are no appropriate exceptions or limitations, the authorization from the rightholders may be required. Although the DSM Directive[15] somewhat addresses this issue by introducing two mandatory TDM exceptions to copyright under Arts. 3 and 4, the efficiency of these provisions when applying to start-ups and SMEs in the context of AI development is questionable. The TDM exception for scientific research will only cover a limited group of beneficiaries such as research organizations and cultural heritage institutions.[16] The

[10] Proposal for a Regulation of the European parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM/2021/206 final
[11] See AIA
[12] AIA 1
[13] AIA 1 (1.1)
[14] European Parliament, Resolution of 20 October 2020 (2020/2015(INI)).
[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.
[16] Art.3 of the DSM Directive

second TDM exception that applies to many actors in the private sector is accompanied by an "opt-out" mechanism, which allows rightholders to "reserve" the use of their works for TDM.[17] This limitation to TDM could seriously reduce the opportunity for digital start-ups and SMEs, which do their business with data analytics, as it would be too complicated, costly, and time-consuming for them to negotiate with all rightholders over protected works or databases. That would mean that only very large commercial companies or Internet giants would benefit from this TDM exception. Therefore, to facilitate rapid technological development and promote innovation in the EU, it is required to establish a more favorable legal environment for TDM users that would help the EU data economy grow.[18] This could be achieved through *inter alia* the proper implementation of the DSM Directive into national laws of the EU Member States. Although the deadline for national transposition was 7 June 2021, the implementation process has not been completed yet due to several reasons such as COVID-19 and the delay of the European Commission's guidance on Art. 17 of the DSM Directive.

To address the abovementioned issues this article will, in the first place, analyze the existing EU frameworks on copyright, database, and computer programs to determine the legal constraints on TDM (Part II). Next, the DSM Directive's TDM exceptions will be evaluated to ascertain whether they are effective and adequate to secure sufficiently AI development. Further, the paper will briefly analyze the positions of the EU Member States, which have already implemented the new legal framework into their national laws, regarding the scope of TDM exceptions. In particular, the article will analyze whether issues that have been unclear since the adoption of the DSM Directive have been covered adequately through the transposition process. In addition, this paper will put forward recommendations on how it could be implemented to unlock the maximum potential of TDM and AI research in the EU (Part III). Finally, the concluding remarks on the issues at stake will be provided (Part IV).

[17] Art. 4 of the DSM Directive

[18] For more discussions on the crucial role of the start-ups and SMEs in the EU economy *see*, for instance, Margoni T, Kretschmer M "A deeper look into the EU Text and Data Mining exceptions: Harmonisation, data ownership, and the future of technology" CREATe Working Paper (2021).

2. The EU Restrictions on TDM in the Context of AI Development

As previously mentioned, the functioning of AI technologies is based on TDM that employs techniques from "natural language processing, machine learning, information retrieval, and knowledge management for the automated analysis of digital content"[19] to extract new knowledge and insights.[20] In the light of recent development and emergence of cutting-edge AI applications, it is apparent that TDM will become an essential technique allowing new types of information-based services and applications.[21] Therefore, to train algorithms and improve AI applications, broader access to datasets, containing digital materials and computer programs, is required.[22] Although TDM appears to be indispensable for the development of innovative technologies, current tensions between copyright and database holders' rights and interests and those of TDM users create barriers to TDM.[23] The following subchapters will discuss the challenges arising from the existing EU legal frameworks that seem to be not very favorable towards the development of AI technologies.

2.1 The InfoSoc Directive and AI

After explaining the functioning of AI technologies and their relationship to TDM, it is crucial to analyze whether the use of copyrighted works in TDM may constitute copyright infringement in

[19] Ducato R, and Strowel A "Limitations to Text and data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility" (2019) IIC 50(6), p.651.
[20] For a technical definition of TDM see Hearst M "Text and Data Mining" in: Mitkov R (ed), The Oxford handbook of computational linguistics. Oxford University Press. Oxford (2005), pp.616-662.
[21] Ducato R, and Strowel A "Limitations to Text and data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility" (2019) IIC 50(6), p.651.
[22] For more discussions on the technical aspects of TDM *see*, for instance, Pesenti H and Hall W (2017) Growing the artificial intelligence in the UK. Independent Report. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf. Accessed 3 March 2022.
[23] For more discussions on copyright reform and TDM *see*, for example, Triaille J "Study on the legal framework of text and data mining (TDM)" (2014) https://www.fosteropenscience.eu/sites/default/files/pdf/3476.pdf. Accessed 7 January 2022; Caspers M, Guibault L et al. "A Baseline report of policies and barriers of TDM in Europe" (2016) https://project.futuretdm.eu/wp-content/uploads/2017/05/FutureTDM_D3.3-Baseline-Report-of-Policies-and-Barriers-of-TDM-in-Europe.pdf. Accessed 10 January 2022; Margoni T and Giulia D "Why We Need a Text and Data Mining Exception (But it is Not Enough)" Working paper (2016) https://interop2016.github.io/pdf/INTEROP-13.pdf. Accessed 23 October 2021. Accessed 12 February 2022; Geiger C "The exception for TDM in the proposed directive in the digital single market-legal aspects" Central of international IP studies Research paper n 2018-02: 1-34 https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/604941/IPOL_IDA(2018)604941_EN.pdf. Accessed 13 February 2022; Rosati E "National and EU TDM exceptions: room for coexistence?" (2018) https://ipkitten.blogspot.com/2018/03/national-and-eu-text-and-data-mining.html. Accessed 20 January 2022.

the EU. In some cases, the developers of AI use protected data lawfully. For example, large commercial organizations or internet giants such as Facebook and Google may have licenses obtained from right holders to use protected content for TDM purposes.[24] Sometimes, AI developers may access data licensed under a Creative Commons license that enables the free distribution of copyrighted content.[25] However, most of the time AI developers may not have the authorization to use works for TDM objectives. The unauthorized use of protected subject matters may potentially lead to copyright infringement under the EU copyright law.[26]

In the EU, rightholders are granted a set of exclusive rights including a broad right to reproduction.[27] Therefore, if a whole work or a substantial part of it is reproduced by any means and in any form, there will be a violation of the right to reproduction. In the course of a TDM process, a computer makes digital copies of the extracted content a couple of times to produce intended outputs thus potentially infringing the reproduction right.[28] Moreover, on some occasions, the outputs or final results that the AI module generates may constitute an infringement too. It may happen, for example, when a small amount of data is loaded into the AI system and therefore the generated output may be similar to the training data or even contain entire or significant parts of protected content. The repeated reproduction of data that may be explicitly regarded as a copyright-

[24] See, for instance, very broad data policy presented by Facebook that may involve the use of users' content for TDM processes with the reason to improve services. See https://www.facebook.com/about/privacy/update. Accessed 18 December 2021.

[25] See https://creativecommons.org/licenses/. Accessed 18 December 2021.

[26] For more discussions on TDM and possible copyright infringement see, for example, Triaille J "Study on the legal framework of text and data mining (TDM)" (2014) https://www.fosteropenscience.eu/sites/default/files/pdf/3476.pdf. Accessed 7 January 2022; Margoni T and Kretchmer M "The text and data exception in the proposal for a DSM: why it is not what EU copyright law needs" (2018) https://www.create.ac.uk/blog/2018/04/25/why-tdm-exception-copyright-directive-digital-single-market-not-what-eu-copyright-needs/. Accessed 10 February 2022; Stamatoudi I "Text and Data Mining" in: Stamatoudi I (ed), New developments in EU and International copyright law. Wolters Kluwer. Alphen aan den Rijn (2016), pp.251-282; Strowel A "Reconstracting 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, Alphen aan den Rijn (2018), pp 203-240.

[27] Art. 2 of the InfoSoc Directive. For more discussions on a broad definition of protected rights see, for example, Simopoulou A "Text and Data Mining under EU Copyright Law" Thessaloniki (2020) https://repository.ihu.edu.gr/xmlui/bitstream/handle/11544/29743/Text%20and%20Data%20Mining%20under%20EU%20Copyright%20Law.pdf?sequence=1 Accessed 1 February 2022, p.6; Margoni T and Kretchmer M "The text and data exception in the proposal for a DSM: why it is not what EU copyright law needs" (2018) https://www.create.ac.uk/blog/2018/04/25/why-tdm-exception-copyright-directive-digital-single-market-not-what-eu-copyright-needs/. Accessed 10 February 2022, p.4; Margoni T, Kretschmer M "A deeper look into the EU Text and Data Mining exceptions: Harmonisation, data ownership, and the future of technology" CREATe Working Paper (2021), p.11.

[28] Sobel "Artificial Intelligence's fair use crisis" Columbia Journal of Law & the Arts 41 (1) (2017), pp.62-63.

relevant act is crucial for the application of TDM techniques during the AI development process.[29] The potential copyright infringement could be exempted based on the statutory exceptions and limitations provided under Article 5 of the InfoSoc Directive.[30] However, as is well known, the EU legal system of exceptions and limitations is based on a closed number of provisions, which were designed during the initial phase of the Digital Era. Therefore, the adoption of these provisions to advanced digital technologies and techniques such as TDM could be problematic as there may be a concern about their ability to balance copyright holders' exclusive rights and users' interests.[31]

For instance, one of the exceptions that could apply to TDM is a compulsory exception for temporary acts of reproduction.[32] This exemption is composed as a set of cumulative conditions, implying that a disregard of any of those conditions will prevent users of copyrighted works benefit from the exception.[33] Art. 5 (1) of the InfoSoc Directive requires that the reproduction of copyrighted works be: (1) temporary; (2) transient or incidental; (3) an integral and essential part of a technological process; (4) the sole purpose of which is to enable ... a lawful use of a work; and (5) the act has no independent economic significance.[34] The European Court of Justice (ECJ) has indicated that those conditions must be interpreted strictly because they derogate from a general principle established by the InfoSoc Directive.[35] However, those conditions are often difficult to meet in TDM processes and their interpretation could not be always viewed as straightforward.[36] For instance, the exception covers only transient and incidental acts of reproduction, which constitute an essential part of a technological process.[37]In this regard, the ECJ has clarified that the duration of a transient copy is limited to what is necessary for the proper completion of the

[29] Borghi M, Karapapa S "Copyright and mass digitization: a cross-jurisdictional perspective" (2013) Oxford University Press, p.51.
[30] Art. 5 of the InfoSoc Directive. For more discussions see, for instance, Geiger C and Schonherr F "The Information Society Directive (Articles 5 and 6(4))" in: Stamatoudi I, Torremans P (eds) EU Copyright Law. Edward Elgar Publishing, Cheltenham (2014), pp.1-67 and Geiger C et al. "Understanding the 'three-step test'" n: Gervais D (ed) Research handbook on international intellectual property law. Edward Elgar Publishing, Cheltenham, pp.167-189.
[31] Margoni T and Kretchmer M "The text and data exception in the proposal for a DSM: why it is not what EU copyright law needs" (2018) https://www.create.ac.uk/blog/2018/04/25/why-tdm-exception-copyright-directive-digital-single-market-not-what-eu-copyright-needs/. Accessed 10 February 2022, p.5.
[32] Art. 5 (1) of the InfoSoc Directive
[33] Infopaq International A/S v Danske Dagblades Forening, C-302/10, EU:C:2012:16
[34] Art. 5 (1) of the InfoSoc Directive
[35] Infopaq International A/S v Danske Dagblades Forening, C-5/08, EU:C:2009:465, paras.56 and 57; Public Relations Consultants Association Ltd v. Newspaper Licensing Agency Ltd and Others C-360/13, EU:C:2014:1195, para. 23
[36] Margoni T, Kretschmer M "A deeper look into the EU Text and Data Mining exceptions: Harmonisation, data ownership, and the future of technology" CREATe Working Paper (2021), p.17.
[37] Art. 5 (1) of the InfoSoc directive.

technological process that enables the browsing of protected content.[38] Moreover, the ECJ has established that the transient copies have to be created and deleted automatically and without human intervention.[39] However, the ECJ reconsidered its position towards human intervention in subsequent rulings by indicating that a technological process that involves human interaction could result in transient copies.[40] In regards to TDM, copies created during an automatic analytical process would likely not qualify as "transient or incidental" but rather permanent since they do not merely enable data browsing. The copies are kept for a longer period in order to be processed, uploaded and finally mined for the purpose of the training of AI system.[41] Moreover, the created copies are not removed automatically when data analysis is completed; the deletion process depends on the will of AI developers.[42]

Further, while the copying process underlying the TDM practices may be referred as an integral and essential part of this technological process,[43] the requirement of enabling a "lawful use" of protected content appears to be problematic.[44] The InfoSoc Directive clarifies the notion of a "lawful use" as a use that is authorized by the rightholder or not restricted by law.[45] As discussed above, TDM may infringe the right to reproduction, which is an act restricted by law, therefore, the authorization must be obtained for each copy made during data analysis that is very complicated or even impossible, bearing in mind that a huge number of works is processed. The

[38] Infopaq International A/S v Danske Dagblades Forening, C-5/08, EU:C:2009:465, para. 64.
[39] Infopaq International A/S v Danske Dagblades Forening, C-5/08, EU:C:2009:465, para. 63.
[40] Infopaq International A/S v Danske Dagblades Forening, C-5/08, EU:C:2009:465, paras. 32, 36, 39.
[41] Public Relations Consultants Association Ltd v. Newspaper Licensing Agency Ltd and Others C-360/13, EU:C:2014:1195, where the ECJ considered relation between 'transient' and 'incidental' paras. 40-50. See also Kovalev S "The EU Commission's proposed TDM exception: unlocking research or innovation?" Thesis (2017) https://www.duo.uio.no/bitstream/handle/10852/58663/Thesis.pdf?sequence=1 Accessed 22 January 2022, p.23.
[42] Public Relations Consultants Association Ltd v. Newspaper Licensing Agency Ltd and Others C-360/13, EU:C:2014:1195, paras. 40-50. Kovalev S "The EU Commission's proposed TDM exception: unlocking research or innovation?" Thesis (2017) https://www.duo.uio.no/bitstream/handle/10852/58663/Thesis.pdf?sequence=1 Accessed 22 January 2022, p.23; Triaille J "Study on the legal framework of text and data mining (TDM)" (2014) https://www.fosteropenscience.eu/sites/default/files/pdf/3476.pdf. Accessed 7 January 2022, p. 46; Lemley M and Casey B (2020) "Fair Learning" (2022) Available at SSRN: https://ssrn.com/abstract=3528447 Accessed 30 December 2021, pp.120-121.
[43] Art. 5 (1) of the InfoSoc Directive.
[44]Triaille J "Study on the legal framework of text and data mining (TDM)" (2014) https://www.fosteropenscience.eu/sites/default/files/pdf/3476.pdf. Accessed 7 January 2022, pp. 44-46.
[45] Recital 33 of the InfoSoc Directive. Infopaq International A/S v Danske Dagblades Forening, C-302/10, EU:C:2012:16, para 44, where the ECJ stated that where an act of reproduction is not restricted by law the authorization from the rightsholder is not required; See also joined cases Football Association Premier League, C-403/08 and 429/08, EU:C:2011:631, para 169-171; C-527/15, Stichting Brein v Jack Frederik Wullems EU:C:2017:300, para 65-68.

last criterion concerning the economic significance of the copies created during the mining process raises some questions as well, especially when discussing a private sector. The ultimate goal of start-ups and SMEs involved in AI creativity is to obtain a financial reward, therefore, the reproduction, as one of the steps leading to economic benefits, during TDM may potentially have an independent economic significance.[46] Failing to meet the abovementioned conditions of Art. 5 (1) of the InfoSoc Directive invalidates the application of this exception due to its cumulative nature.

In fact, the ECJ has acknowledged that Art. 5 (1) serves as a facilitator of technological development when it specifies the main function of that provision as to "allow and ensure the development and operation of new technologies, and safeguard a fair balance between the rights and interests of rights holders and of users of protected works who wish to avail themselves of those technologies."[47]Although the ECJ's statement seems to benefit modern TDM practices and AI development, it is difficult to imagine how the users' interests can be safeguarded by a strict interpretation and the cumulative nature of the already narrowly defined conditions of Art. 5 (1).[48] The fact that the exception only embraces acts of temporary reproduction does not offer a comfortable "safe harbor" for users employing TDM analytical techniques.[49]

The exception for scientific research is another copyright limitation that is relevant for AI's data analysis processes.[50] Art. 5 (3) (a) of the InfoSoc Directive reads as "use for the sole purpose of illustration for teaching or scientific research, as long as the source, including the author's name, is indicated, unless this turns out to be impossible and to the extent justified by the non-commercial purpose to be achieved."[51] It is possible to identify a number of obstacles in applying the scientific research exception to AI development. Specifically, the exception has not been implemented uniformly across the EU due to its optional nature, resulting in different interpretations and

[46]Triaille J "Study on the legal framework of text and data mining (TDM)" (2014) https://www.fosteropenscience.eu/sites/default/files/pdf/3476.pdf. Accessed 7 January 2022, p. 47.
[47] Case C-360/13, Public Relations Consultants Association Ltd v. Newspaper Licensing Agency Ltd and Others (2014), EU:C:2014:1195, para. 24.
[48] Margoni T, Kretschmer M "A deeper look into the EU Text and Data Mining exceptions: Harmonisation, data ownership, and the future of technology" CREATe Working Paper (2021), p.17.
[49] Margoni T, Kretschmer M "A deeper look into the EU Text and Data Mining exceptions: Harmonisation, data ownership, and the future of technology" CREATe Working Paper (2021), p.21
[50] Art. 5 (3) (a) of the InfoSoc Directive.
[51] Art. 5 (3) (a) of the InfoSoc directive.

applications through the EU Member States.[52] The provision covers the purpose of "illustration" of copyrighted works which is understood as "to clarify something by giving or serving as, an example of a comparison"[53]. During a TDM process, extracted data is not used as an example ("illustration") but rather as a "material" of analysis. Moreover, the provision requires that the use would cover the purpose of "scientific research" that is not clear whether the process of extracting new ideas and insights by means of TDM would fit that purpose, meaning that all TDM projects in the context of AI development will add something new to the "state of science"[54] as they may only pursue another, a profit-gain purpose.

Another difficulty in applying this exception to TDM appears from the requirement of the non-commercial nature of the activity.[55] The wording of the provision is likely to exclude private organizations such as start-ups and SMEs from the scope of the exception for scientific research as they usually pursue commercial goals when doing business based on data analysis. Further, the provision requires that the source of protected content must be indicated unless this turns out to be impossible.[56] In the context of TDM that requires the processing of a huge amount of data to generate desired outputs, it could be difficult or even impossible to attribute all extracted works. Finally, after considering two exceptions that might apply to the acts of reproduction of works that are related to AI analysis, it becomes obvious that they cannot be viewed as a sufficient legal foundation for TDM due to their inapplicability or legal uncertainty.

2.2 The Database Directive and AI

Since a TDM process may involve data that is obtained from databases, the EU database legal framework comes to be applicable to AI technologies. In the EU, therefore, the use of data may be

[52] Dusollier S "The limitations and exceptions to copyright and related rights for libraries, research and teaching uses" (2013) in Study on the Application of Directive 2001/29/EC on Copyright and Related Rights in the Information Society (the "InfoSoc Directive"), European Union., p. 61.
[53] Since the InfoSoc directive does not define illustration I am indicating its ordinary meaning by referring to the online Merriam-Webster dictionary.
[54] Triaille J "Study on the legal framework of text and data mining (TDM)" (2014) https://www.fosteropenscience.eu/sites/default/files/pdf/3476.pdf. Accessed 7 January 2022, p. 60; In this regard, the High Court of Justice in UK have ruled that the purpose of research must be narrowly interpreted, see Forensic Telecommunications Services Ltd v Chief Constable of West Yorkshire Police, High Court of Justice, Chancery Division 9 November 2011, [2011] EWHC 2892 (Ch), [2012] F.S.R. 15, § 109.
[55] Recital 42 of the InfoSoc Directive.
[56] Art. 5 (3) (a) of the InfoSoc Directive.

limited based on the EU Directive on the legal protection of databases[57] (the Database Directive). In the case when a dataset qualifies as a "database,"[58] a dataset would be eligible for protection by copyright, *sui generis* right, or even a combination of them under certain circumstances.[59] Specifically, the copyright protection covers the structure of the database and thus, would not affect protected works that it includes,[60] whilst the *sui generis* protection would rather expand to the content of that database.[61] The copyright protection of the database intends "databases which, by reason of the selection or arrangement of their contents, constitute the author's own intellectual creation."[62] Art. 5 of the Database Directive prescribes a number of exclusive rights granted to the author of a database,[63] including the right to reproduction that is relevant to data analysis.[64] The right to reproduction could be infringed if any copies of the original structure of a database were created during the stage of extraction of protected content.[65]

The maker of a database[66] would benefit from the *sui generis* protection if "there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents."[67] The "investment" could be understood as any financial, material, or human efforts.[68] The author enjoys two transferable rights: the right to extraction and the right to re-utilization.[69] In respect to AI, the training of the algorithms on datasets would violate the right to extraction if there were a transfer of a substantial part of the database to another medium by any means or in any form.[70] However, the maker of a database cannot prohibit the use of

[57] 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*
[58] Art. 1 (2) of the Database Directive defines "database" as "a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means."
[59] Art. 7(4) of the Database Directive. See, for instance, Lewinski S "Database Directive" (2010) In: Walter M, Lewinski S (eds), European copyright law. A commentary, Oxford university press. Oxford, p.763.
[60] Art. 3 (2) of the Database Directive.
[61] Art. 7 of the Database Directive.
[62] Recital 16 and Art. 3 (1) of the Database directive.
[63] Art. 5 of the Database Directive.
[64] Art. 5 (a) of the Database Directive.
[65] Geiger C et al. (2018) (a) Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data? IIC 49 (7), p.118.
[66] The Database Directive defines the maker of a database as "the person who takes the initiative and the risk of investing" and can be a natural or a legal person. Art. 11 and Recital 41 of the Database Directive.
[67] Art. 7 (1) of the Database Directive.
[68] Recitals 7, 39 and 40 of the Database Directive; C-203/02 the British horseracing board LTD; ECR-I 76=GRUR Int 2005; Fixtures Marketing Ltd v Organismos Prognostikon Agnon Podosfairou, C-444/02, ECR-I 28, 2005.
[69] Art. 7 of the Database Directive.
[70] Art. 7 (2) (a) of the Database Directive. See also Caspers M, Guibault L et al. "A Baseline report of policies and barriers of TDM in Europe" (2016) https://project.futuretdm.eu/wp-content/uploads/2017/05/FutureTDM_D3.3-Baseline-Report-of-Policies-and-Barriers-of-TDM-in-Europe.pdf. Accessed 20 January 2022, p.22.

insubstantial parts of the content of a database, unless such activities involve "the repeated and systematic extraction and/or re-utilization of insubstantial parts of the contents of the database implying acts which conflict with a normal exploitation of that database or which unreasonably prejudice the legitimate interests of the maker of the database."[71]During the course of AI development, the extracted content is mined a couple of times and kept for a moment in the computer allowing it to learn from past experience.[72] It is not quite clear whether this analytical technique may violate the *sui generis* right as the algorithms can only crawl through the datasets to discover new insights and patterns, thereby extracting merely insubstantial parts of the content of a database.[73] However, such extractions, once repeated and systematic, may end up amounting to the extraction of a substantial part of the content of a database.[74] To sum up, much like in the case of copyright protection, the development of AI technologies may cover acts that violate the *sui generis* right; therefore, in the absence of any relevant exceptions the right holders' authorization is required.[75]

Although the Database Directive introduces a number of optional exceptions,[76] their relevance to start-ups and SMEs that are involved in the development of the new forms of AI is questionable. For instance, one limitation that may cover TDM processes would be an exception for scientific research, providing that the source is indicated and the purpose is non-commercial.[77] The Database Directive indicates that the term "scientific research" covers both the natural sciences and the

[71] Art. 7 (5) of the Database Directive.

[72] For more discussions on the technical aspects of TDM *see*, for instance, Pesenti H and Hall W " Growing the artificial intelligence in the UK" (2017) Independent Report. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf. Accessed 23 December 2021.

[73]Triaille J "Study on the legal framework of text and data mining (TDM)" (2014) https://www.fosteropenscience.eu/sites/default/files/pdf/3476.pdf. Accessed 7 January 2022; Lewinski S "Database Directive" (2010) in: Walter M, Lewinski S (eds), European copyright law . A commentary, Oxford university press. Oxford.

[74] For more discussions *see*, for example, Triaille J "Study on the legal framework of text and data mining (TDM)" (2014) https://www.fosteropenscience.eu/sites/default/files/pdf/3476.pdf. Accessed 7 January 2022; Lewinski S "Database Directive" (2010) in: Walter M, Lewinski S (eds), European copyright law. A commentary, Oxford university press. Oxford.

[75] Lewinski S "Database Directive" (2010) in: Walter M, Lewinski S (eds), European copyright law. A commentary, Oxford university press. Oxford, p.763. Triaille J "Study on the legal framework of text and data mining (TDM)" (2014) https://www.fosteropenscience.eu/sites/default/files/pdf/3476.pdf. Accessed 7 January 2022, p.32.

[76] Art. 6 (2) of the Database Directive.

[77] Art. 6 (2) (b) and 9 (b) of the Database Directive.

human sciences[78] and implies that there would be some contribution to the field of science.[79] However, in the case of AI, this aspect is not absolutely clear. Mostly, the development of AI technologies results in scientific progress, but sometimes the developers of innovative technologies merely employ scientific methods and thus raise a question of the applicability of the notion of "scientific research" to their processes. Moreover, the requirement to indicate the source of protected databases appears to be an insurmountable technical problem for developers of AI as TDM processes include the processing of a huge number of protected works.[80] Considering a "non-commercial purpose" requirement, it is reasonable to assume that start-ups and SMEs would be deprived of the scope of the exception as their activities imply direct or indirect commercial objective.[81] To summarize, the exception for scientific research appears to be irrelevant for businesses in the context of development of AI.

The Database Directive also provides mandatory exceptions to ensure a balance of interests between database holders and users of the protected content.[82] The relevant provisions stipulate that a lawful user of a database may not be prevented from access to and use of the contents of a database.[83] The ECJ has pointed out that the notion of "lawful user" should be understood as covering users "whose access to the contents of a database for the purpose of consultation results from the direct or indirect consent of the right holder."[84] Further, the ECJ has also identified that lawful access to content is an essential condition for the lawful use of that content.[85]Accordingly, in the case when a content was made available to the public with the rightholder's consent, such access should be regarded as lawful.[86] Art. 6 (1) of the Database Directive also indicates that a "lawful user" is required to use the content of a database in a "normal" way.[87] The notion of

[78] Recital 36 of the Database Directive.
[79] See for instance, Stamatoudi I "Text and Data Mining" (2016) In: Stamatoudi I (ed), New developments in EU and International copyright law. Wolters Kluwer. Alphen aan den Rijn, p.262.
[80] Kovalev S "The EU Commission's proposed TDM exception: unlocking research or innovation?" Thesis (2017) https://www.duo.uio.no/bitstream/handle/10852/58663/Thesis.pdf?sequence=1 Accessed 22 January 2022, p.26.
[81] Ducato R, and Strowel A "Limitations to Text and data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility" (2019) IIC 50(6), p.661.
[82] Art. 6 (1) and 8 (1) of the Database Directive.
[83] Art. 6 (1) and 8 (1) of the Database Directive.
[84] C-203/02 The British horseracing board LTD and others v William Hill Organization Ltd, ECR-I 76=GRUR Int 2005, para.58.
[85] C 435/12 ACI Adam BV and Others v Stichting de Thuiskopie and Stichting Onderhandelingen Thuiskopie vergoeding, EU:C:2014:254, para. 39.
[86] C 463/12 Copydan Båndkopi v Nokia Danmark A/S, EU:C:2015:144, para.74.
[87] Art. 6 (1) of the Database Directive.

"normal use" refers to the use based on the relevant exceptions or licensing.[88] Since the existing exceptions to copyright protection of databases appear to be irrelevant for start-ups and SMEs, they are always obliged to acquire the right holder's consent to qualify as "lawful users" to mine protected content. Accordingly, it can be assumed that a "normal use" covers the purpose and the way of access and use specified in the licensing agreement.[89] The fact that TDM can be limited or disregarded through the agreement for the benefit of the rightholders[90] also limits the application of Art. 6 (1) to data analysis.

Indeed, the requirement that the reproduction must be necessary "for the purposes of access to the contents of the databases and normal use of the contents"[91] appears to have no application for TDM. On this point, since the main purpose of TDM in AI development is to extract new patterns, correlations and insights from a huge number of existing databases, it is unlikely that the mining process would be viewed as "necessary" to access the content of databases in the context of the "normal use" of databases.[92] In respect to the *sui generis* protection, the Database Directive, as has been discussed above, allows the use of insubstantial parts of the content of a database.[93] In the context of the development of AI, start-ups and SMEs practically need to extract whole or substantial parts of accessed datasets to conduct mining processes. Limiting this exception to "insubstantial parts" makes it ineffective and worthless for successful training of algorithms. Eventually, it can be argued that there are no relevant exceptions in the Database Directive that would be applicable to and provide more legal certainty for start-ups and SMEs that intend to use databases for AI development.[94] Accordingly, the EU legal framework on database protection may constrain the creation of new innovative products in a private sector.

---

[88] Triaille J "Study on the legal framework of text and data mining (TDM)" (2014) https://www.fosteropenscience.eu/sites/default/files/pdf/3476.pdf. Accessed 7 January 2022, p.75.

[89] C-30/14 Ryanair Ltd v PR Aviation BV, EU:C:2015:10, para. 21; Triaille J "Study on the legal framework of text and data mining (TDM)" (2014) https://www.fosteropenscience.eu/sites/default/files/pdf/3476.pdf. Accessed 7 January 2022, pp. 72-73. Recital 34 of the Database Directive.

[90] Caspers M, Guibault L et al. "A Baseline report of policies and barriers of TDM in Europe" (2016) https://project.futuretdm.eu/wp-content/uploads/2017/05/FutureTDM_D3.3-Baseline-Report-of-Policies-and-Barriers-of-TDM-in-Europe.pdf. Accessed 20 January 2022, p. 33.

[91] Art. 6 (1) of the Database Directive.

[92] Triaille J "Study on the legal framework of text and data mining (TDM)" (2014) https://www.fosteropenscience.eu/sites/default/files/pdf/3476.pdf. Accessed 7 January 2022, pp.75-76; Geiger C et al. "Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?" (2018) IIC 49 (7), p.824.

[93] Art. 8 (1) of the Database Directive.

[94] Margoni T and Kretchmer M "The text and data exception in the proposal for a DSM: why it is not what EU copyright law needs" (2018) https://www.create.ac.uk/blog/2018/04/25/why-tdm-exception-copyright-directive-digital-single-market-not-what-eu-copyright-needs/. Accessed 10 February 2022, p.5.

2.3 The Software Directive and AI

Computer programs can be viewed as an important source of mining in the course of AI development. However, when start-ups and SMEs perform TDM on software, their interests may clash with the interests of the owner of software protected under the EU Directive on the legal protection of computer programs[95] (the Software Directive). When TDM is employed for training algorithms during the AI development process with the aim to generate new knowledge by using software,[96] any acts of reproduction of that software require the rights holder's authorization.[97] In this respect, the Software Directive introduces two mandatory exceptions to the exclusive rights of the right holder of a computer program under Art. 5, which permits the appropriate use of computer programs by a lawful acquirer,[98] and Art. 6, which allows decompilation of the program to achieve interoperability of computer programs if it is indispensable to obtain the interface information.[99]

In respect to TDM, the exception introduced under Art. 5 is of special importance. Art. 5 (1) of the Software Directive indicates that the authorization of the rightsholder is not required where it is "necessary for the use of the computer program by the lawful acquirer in accordance with its intended purpose, including for error correction".[100] As already explained concerning Art. 6 (1) of the Database Directive, the notion of a lawful acquirer (user) implies that only access to computer programs by means of licensing mechanism or relevant exceptions is covered by this exception.[101] The Software Directive does not define the concept of "intended purpose" but it most likely refers to the purpose for which software was obtained and the normal use for that purpose.[102] As it was discussed in respect to the Database Directive, the purpose of TDM in developing AI is to extract

[95] 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*.
[96] Triaille J "Study on the legal framework of text and data mining (TDM)" (2014) https://www.fosteropenscience.eu/sites/default/files/pdf/3476.pdf. Accessed 7 January 2022, p. 75.
[97] The right to reproduction protected under Art. 4 (1) (a) of the Software Directive.
[98] Art. 5 (1) of the InfoSoc Directive.
[99] Art. 6 of the Software Directive.
[100] Art. 5 (1) of the Software Directive.
[101] Triaille J "Study on the legal framework of text and data mining (TDM)" (2014) https://www.fosteropenscience.eu/sites/default/files/pdf/3476.pdf. Accessed 7 January 2022, p.75.
[102] C-406/10 SAS Institute Inc. vastaan World Programming Ltd, EU:C:2012:259, para. 57.

new knowledge and it can therefore be supposed that the reproduction of computer programs in this sense is not such as to constitute a necessary act for normal use of databases.[103]

Art. 5 (3) of the Software Directive introduces another exception, the so-called "black box analysis," that could be potentially applicable to TDM.[104] Therefore, this exception permits "the person having the right to use a copy of a computer program…to 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."[105] In fact, the extraction of ideas contained in protected works is external to copyright protection.[106] The principle of idea/expression dichotomy implies that copyright only protects original expressions, whereas ideas, procedures, facts and principles are exempted from protection.[107] Accordingly, in order to "constrain" the broadly defined right to reproduction from restricting access to the underlying ideas and principles of the computer programs unprotected by copyright, Art. 5 (3) becomes effective and appropriate.[108] In this regard, the purpose of TDM is not to copy the expression of computer programs but rather to extract new ideas from software and reproduce it to proceed with data analysis for the training of the AI system.[109] With this in view, TDM can fall within the scope of Art. 5 (3) without infringing the copyright in software as such.

---

[103] See part 2.2 of this article. Triaille J "Study on the legal framework of text and data mining (TDM)" (2014) https://www.fosteropenscience.eu/sites/default/files/pdf/3476.pdf. Accessed 7 January 2022, pp.75-76. Geiger C et al. "Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?" (2018) IIC 49 (7), p. 824.

[104] Ducato R and Strowel A "Ensuring Text and Data Mining: Remaining issues with the EU copyright exceptions and possible ways out" (2021) CRIDES Working Paper Series No. 1/2021, p.14.

[105] Art. 5 (3) of the Software Directive. Recitals 14-15 of the Software Directive.

[106] Margoni T and Kretchmer M "The text and data exception in the proposal for a DSM: why it is not what EU copyright law needs" (2018) https://www.create.ac.uk/blog/2018/04/25/why-tdm-exception-copyright-directive-digital-single-market-not-what-eu-copyright-needs/. Accessed 10 February 2022, p.11.

[107] The recognition of the idea-expression doctrine can be found in the TRIPS Agreement (Art. 9 (2)), in the Software Directive (Recital 11 and Arts. 1 and 5 (3)), Recital 45 of the Database Directive and in Recital 9 of the EU DSM Directive. See for example Case C-833/18 SI and Brompton Bicycle Ltd v Chedech / Get2Get (2020), EU:C:2020:461, para. 27; Case C-683/17 Cofemel – Sociedade de Vestua´rio SA v. G-Star Raw CV (2019), ECLI:EU:C:2019:721, para 29; Case C-393/09 Bezpečnostní softwarová asociace - Svaz softwarové ochrany v Ministerstvo kultury (2010), EU:C:2010:816, para. 49; Case C-306/05 Sociedad General de Autores y Editores de España (SGAE) v Rafael Hoteles SA (2006), para 35.

[108] Triaille J "Study on the legal framework of text and data mining (TDM)" (2014) https://www.fosteropenscience.eu/sites/default/files/pdf/3476.pdf. Accessed 7 January 2022, p.109.

[109] Triaille J "Study on the legal framework of text and data mining (TDM)" (2014) https://www.fosteropenscience.eu/sites/default/files/pdf/3476.pdf. Accessed 7 January 2022, p.109; See for further discussion Caspers M, Guibault L et al. "A Baseline report of policies and barriers of TDM in Europe" (2016) https://project.futuretdm.eu/wp-content/uploads/2017/05/FutureTDM_D3.3-Baseline-Report-of-Policies-and-Barriers-of-TDM-in-Europe.pdf. Accessed 20 January 2022.

However, the ECJ has pointed out that the determination of the ideas and principles, which underline any element of the computer program, may be carried out within the framework of the acts permitted by the license.[110] Although Art. 8 of the Software Directive provides that any contractual provisions contrary to the exception in Art. 5 (3) must be null and void,[111] the righholders may define the permitted use within narrow margins, thus limiting the possibility for TDM users, such as start-ups and SMEs, to involve in TDM for AI development while enjoying their contractual rights. In addition, the provision of the exception postulates that only "the person having a right to use a copy of a computer program"[112] is entitled to benefit from this exception, therefore further limiting its application. Furthermore, the exception in Art. 5 (3) of the Software Directive is likely to cover TDM of computer programs only for research purposes.[113] Although Art. 5 (3) does not discern the difference between commercial and non-commercial activities,[114] the acts aimed at "scientific research" normally imply non-commercial actions. To conclude, the exception does not provide much legal certainty for start-ups and SMEs engaged in developing AI technologies as its applicability to TDM is questionable due to a number of limiting criteria.

3. The DSM Directive's TDM Exceptions: a Sufficient Balancing Instrument?

3.1 Introducing Arts. 3 and 4 of the DSM Directive

As it has been indicated earlier, the EU system of exceptions and limitations before the DSM Directive was highly criticized for its inability to ensure an adequate balance between copyright and the public interests provoked by the limited adaptability of the closed and exhaustive list of exceptions to rapid technological development.[115] The DSM Directive has introduced two

[110] C-406/10, SAS Institute Inc., (2012), EU:C:2012:259, paras. 54-55.
[111] Art. 8 of the Software Directive.
[112] Art. 5 (3) of the Software Directive.
[113] Ducato R and Strowel A "Ensuring Text and Data Mining: Remaining issues with the EU copyright exceptions and possible ways out" (2021) CRIDES Working Paper Series No. 1/2021, p. 15
[114] Guibault L "Blogpoll: towards a Text & Data Mining exception in EU copyright law?" (2015) Kluwer Copyright Blog. Accessed 10 March 2022.
[115] Sganga C "A New Era for EU Copyright Exceptions and Limitations? Judicial Flexibility and Legislative Discretion in the Aftermath of the CDSM Directive and the Trio of the Grand Chamber of the CJEU" (2020) ERA Forum, vol.21 (2), Springer Verlag, p.314; Geiger C " Flexibilising copyright – remedies to the privatisation of information by copyright law" (2008) IIC 39(2), p.178; Guibault L "Why cherry-picking never leads to harmonisation: the case of the limitations of copyright under directive 2001/29/EC" (2010) JIPITEC 55(1). https://www.jipitec.eu/issues/jipitec-1-2-2010/2603/JIPITEC%202%20-%20Guibault-Cherrypicking.pdf.pdf. Accessed 27 March 2022, p.53; Hugenholtz B and Senftleben M "Fair use in Europe. In search of flexibilities" (2011)

mandatory TDM exceptions to copyright and database protection that can be viewed as a significant step forward in regulating TDM in the EU. The new TDM exceptions seek to achieve a "fair balance" between the rights and interests of authors and users.[116] Art. 3 of the DSM Directive introduces a mandatory exception to copyright, *sui generis* database right, and new press publishers' right under article 15 of the DSM Directive for the purpose of TDM excluding the rights protected under the Software Directive.[117] The provision exempts acts of "reproductions and extractions made by research organizations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access."[118]

The DSM Directive defines "research organizations" and "cultural heritage institutions" as universities, research institutions, libraries, museums or any other entity conducting scientific research on a nonprofit basis or by reinvesting all the profits in their scientific research or pursuant to a public interest mission.[119] The DSM Directive also provides that research organizations and cultural heritage institutions should also benefit from this exception when their research activities are carried out in the framework of public-private partnerships when they rely on their private partners and their technological tools for carrying out TDM.[120] In this way, the existing Union research policy encourages research organizations to collaborate with the private sector.[121] Art. 3 (2) of the DSM Directive also indicates that copies made "shall be stored with an appropriate level of security and may be retained for the purposes of scientific research, including for the verification of research results."[122] Additionally, Art. 3 (3) allows the rightsholders to apply measures to ensure the security and integrity of the networks and databases, which are proportionate to achieve the objective,[123] whereas according to Art. 3 (4) the application of these measures should be the result

Amsterdam Law School Legal Studies Research Paper No. 2012-39. Institute for Information Law Research Paper No. 2012-33, p.9.

[116] Recital 6 of the DSM Directive.

[117] Art.3 of the DSM Directive.

[118] Art.3 (1) of the DSM Directive.

[119] Art. 2 (1) and (3) of the DSM Directive.

[120] Recital 11 of the DSM Directive.

[121] Recital 11 of the DSM Directive. 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) Asia Pacific Law Review 27 (2), p. 212.

[122] Art. 3 (2) of the DSM Directive.

[123] Art 3 (3) of the DSM Directive. See additionally Recital 16 of the DSM Directive.

of commonly agreed best practice when applying the exceptions.[124] The exception is unwaivable by a contract according to Art. 7 (1) of the DSM Directive.[125]

Art. 4 of the DSM Directive introduces the second mandatory exception that would apply to a broader group of beneficiaries and allow TDM for all types of purposes so as not limiting to non-commercial activities.[126] As stated in Art. 4 (1), Member States are required to provide exceptions or limitations to copyright, *sui generis* database right, press publishers' right under Art. 15 of the DSM Directive and, on the contrary to Art. 3 (1), to rights protected under the Software Directive, permitting acts of "reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining."[127] Contrary to Art. 3 (2) of the DSM Directive, Art. 4 (2) indicates that "reproductions and extractions made…may be retained for as long as is necessary for the purposes of text and data mining."[128] Moreover, Art. 4 (3) of the DSM Directive stipulates that the rightholder may reserve in an appropriate manner the use of protected works for the purpose of TDM.[129] Although the provision does not clarify how the act of reservation can be conducted, it is likely that the rightholder will use, in the case where the work has been publicly available online, technical protection measures (TPMs) such as "machine-readable means, including metadata and terms and conditions of a website or a service."[130] The Recital 18 of the DSM Directive also provides that "in other cases, it can be appropriate to reserve the rights by other means, such as contractual agreements or a unilateral declaration."[131] In fact, the exception is not unwaivable by contract,[132] the rightholder, therefore, may limit its application for commercial use by including "robot.txt" metadata in their online works or imposing relevant restrictions through licensing mechanism.[133]

[124] Art. 3 (4) of the DSM Directive.
[125] Art. 7 (1) of the DSM Directive.
[126] Art. 4 (1) of the DSM Directive.
[127] Art. 4 (1) of the DSM Directive.
[128] Art. 4 (2) of the Database Directive.
[129] Art. 4 (3) of the DSM Directive.
[130] Recital 18 of the DSM Directive.
[131] Recital 18 of the DSM Directive.
[132] Art. 7 (1) of the DSM Directive.
[133] Hugenholtz B (2019) The New Copyright Directive: Text and Data Mining (Articles 3 and 4). http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/ Accessed 11 March 2022. Ducato R, and Strowel A "Limitations to Text and data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility" (2019) IIC 50(6), pp.670-671.

3.2 The DSM Directive's TDM exceptions: challenges and solutions

After the initial review of the TDM exceptions, it seems reasonable to consider the legal framework introduced by the DSM Directive from a critical point of view. This subchapter will discuss the potential shortcomings of the TDM exceptions and their national transposition. In addition, the paper will briefly analyze the positions of the EU Member States which have already implemented the new legal framework into their national laws. As of now (21 June 2022), fourteen of 27 EU Member States have fully transposed the DSM Directive into their national law.[134] The article will refer to a national transposition of some of those countries to provide a comprehensive analysis of the issue at stake.

As has been indicated above, the exception in Art. 3 of the DSM Directive is limited in scope as it only benefits research organizations, which "have to operate either on a not-for-profit basis or by reinvesting all the profits in their scientific research "or "pursuant to a public interest mission recognized by the Member State."[135] Although Recital 11 of the DSM Directive introduces the possibility for public-private partnership,[136] Art. 2 (1) stipulates 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 organization."[137] The narrow scope of the beneficiaries of Art. 3 limits the exercising of the exception by research organizations in the case where their TDM projects are supported by private funding or unaffiliated researchers, which often rely on private organizations that usually have control over the TDM projects.[138]

This provision can potentially restrain the development of new applications in the private sphere for the reason that startups and SMEs may not qualify as research organizations. Moreover, it appears unlikely that startups and SMEs will qualify as "cultural heritage institutions,"[139] which form the second group of beneficiaries. As a result, many organizations would be discouraged from conducting AI research based on TDM due to the large licensing costs that they would be

[134] DSM Directive Implementation Tracker. Available at https://www.notion.so/DSM-Directive-Implementation-Tracker-361cfae48e814440b353b32692bba879 . Accessed 21 June 2022.
[135] Art. 2 (1) (a) and (b) of the DSM Directive.
[136] Recital 11 of the DSM Directive.
[137] Art. 2 (1) of the DSM Directive.
[138] Ducato R and Strowel A "Ensuring Text and Data Mining: Remaining issues with the EU copyright exceptions and possible ways out" (2021) CRIDES Working Paper Series No. 1/2021, p.19; Samuelson P "The EU's Controversial Digital Single Market Directive" 92018) Communications of the ACM 61 (11), p.23.
[139] For more discussions *see*, for instance, Geiger C et al. "Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?" (2018) IIC 49 (7): 814-844.

required to pay for mined content. In this regard, there is a possibility for the EU Member States to extend the scope of the TDM exception as to cover additional groups of beneficiaries. For instance, the German legislator has already taken "timid" steps toward expansion of the scope of this TDM exception by explicitly permitting individual researchers that do not pursue commercial purposes to perform TDM for the purpose of scientific research.[140] The Italian legislator has allowed researchers using TDM to not only make copies of the mined content but also perform the communication to the public of the results of the research.[141]

Moreover, Art. 3 of the DSM Directive also limits the purpose of TDM activities applying only to "scientific research" that "covers both the natural sciences and the human sciences."[142] Such limitation may be problematic for many actors involved in TDM activities as sometimes it can be difficult to classify research as both natural science and human science, in particular in the context of AI development. In this regard, the researcher would be required to obtain licenses to comply with the provision of Art. 3 of the DSM Directive that would definitely impede the AI innovative processes.[143] The notion of "scientific research" could be better formulated through the process of national implementation of the DSM Directive. It could be interpreted in a broad manner as to involve any research activity conducted in accordance with relevant methodological standards, for the purpose of generating new knowledge and insights in the relevant field of study. Moreover, it is unclear whether the purpose limitation of Art. 3 will heavily restrict research organizations from using TDM for other institutional objectives such as educational or administrative purposes including teaching illustration. This appears to be a critical issue which should also be clarified during the period of national implementation.

Further, Art. 3 (2) of the DSM Directive also indicates that the beneficiaries of this exception are permitted to retain the copies of the mined content with an appropriate level of security, for instance for the verification of research results. [144] However, the status of those copies is not clear

[140] See the German Implementation Law (Gesetz zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarktes). Available at https://www.notion.so/Germany-d305b35bb3b24bffb909f00592da791e , para 60 d (3) 2. Accessed 12 April 2022.
[141] See The Italian legislative Decree Art. 70-ter. (DECRETO LEGISLATIVO 8 novembre 2021, n. 177). Available at https://www.gazzettaufficiale.it/atto/serie_generale/caricaDettaglioAtto/originario?atto.dataPubblicazioneGazzetta=2021-11-27&atto.codiceRedazionale=21G00192&elenco30giorni=true. Accessed 12 April 2022.
[142] Recital 12 of the DSM Directive.
[143] Ducato R and Strowel A "Ensuring Text and Data Mining: Remaining issues with the EU copyright exceptions and possible ways out" (2021) CRIDES Working Paper Series No. 1/2021, p.11.
[144] Art. 3 (2) of the DSM Directive.

as it is not specified whether third parties are allowed to use such data for the purpose of scientific research such as scientific peer review, joint research, and also further verification of the research outputs.[145] To bring more certainty and reflect the demands of research work, the issue could be clarified during the phase of national implementation. For instance, the German Implementation Act explicitly establishes that a group of actors involved in joint scientific research or third parties which are engaged in the process of verification of the quality of scientific research are allowed to access the "collection" of copies of the protected content generated during TDM. However, once the joint scientific research or verification of the quality of scientific research has been completed, public disclosure must be terminated.[146]

As previously mentioned, right holders are allowed to establish security measures "to ensure the security and integrity of the networks and databases where the works or other subject matter are hosted."[147] Member States should encourage the relevant stakeholders to define best practices concerning the application of the relevant security measures required under Art. 3 (2) and Art. 3 (3).[148] The Irish transposition regulation obliges beneficiaries of the TDM exception to "ensure that only persons who have lawful access to the data contained in that copy shall be permitted to access those data, including through IP address validation or user authentication."[149] Moreover, they should inform authors of the making of the copy,[150] provide information on the steps taken to ensure the security and integrity of the networks[151] and consider additional security measures to be applied on author's request.[152] Therefore, the Irish transposition regulation places additional burdens on researchers and cultural heritage institutions that could be viewed as not very favorable adaptation of Art. 3 of the DSM Directive. What would be more important is to clarify the nature of those measures during the period of national implementation.

---

[145] Ducato R and Strowel A "Ensuring Text and Data Mining: Remaining issues with the EU copyright exceptions and possible ways out" (2021) CRIDES Working Paper Series No. 1/2021, p.11.
[146] Para 60 d (4) 1 and 2 of the German Implementation Act (Gesetz zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarktes) Available at https://www.notion.so/Germany-d305b35bb3b24bffb909f00592da791e , Accessed 12 March 2022 .
[147] Art. 3 (3) of the DSM Directive.
[148] Art. 3 (4) of the DSM Directive.
[149] The Irish transposition law of the DSM Directive, Part 3, 3 (1), (3B) Available at https://www.irishstatutebook.ie/eli/2021/si/567/made/en/pdf. Accessed 10 March 2022.
[150] Part 3, 3 (1), (3C) a of the Irish transposition law of the DSM Directive.
[151] Part 3, 3 (1), (3C) b of the Irish transposition law of the DSM Directive.
[152] Part 3, 3 (1), (3C) c of the Irish transposition law of the DSM Directive.

Art. 4 of the DSM Directive provides more limited in scope TDM exception. In fact, Art. 4 (3) introduces an "opt-out" mechanism implying that the rightholder may "reserve in an appropriate manner" the use of protected works for TDM purposes.[153] For instance, in the case when the content is freely available online, the rightholder may limit or entirely prohibit TDM through the terms and conditions of the website or by employing machine-readable means.[154] Moreover, the provision of Art. 4 can be overridden by a contract. Although this exception intends to benefit the private sector by providing the possibility of reserving use for the purpose of TDM to the right holders, it "legitimizes a derivative market in TDM, enabling right holders to control, license or entirely prohibit such activities".[155] Giving too much control to the rightholders over their intellectual "products" may suppress the ability of users to access, share, or express protected content across the public and private spheres.[156] In particular, it can be too complicated, costly, and time-consuming for start-ups and SMEs involved in AI development to negotiate with all rightholders in the case of the "reserved use". That would mean that only very large commercial companies or Internet giants would benefit from this TDM exception.

Therefore, it would be crucial, during the phase of national implementation, to explicitly specify how this "opt-out" mechanism has to be performed to avoid unnecessary restrictions of TDM. However, we already have a "bad" example of transposition of Art. 4 of the DSM Directive. The Italian Legislative Decree has omitted to implement an important part of this provision. Art. 70-*quarter* of the Legislative Decree simply permits rightholders to reserve the use of their works, without the need to indicate that such reservation should be done "in an appropriate manner".[157] This narrow reproduction of the wording of the DSM Directive can significantly limit the application of Art. 4. In contrast, the Irish transposition regulation explicitly clarifies what the reserving rights for TDM are, stating that such reservation is concerned (a) *"machine-readable in the case of content made publicly available online, including metadata and terms and conditions*

[153] Art. 4 (3) of the DSM Directive; Recital 18 of the DSM Directive.
[154] The use of the robot.txt can be sufficient to prohibit TDM on online content. For instance, rightholders may employ the exclusion protocol to prevent crawling and indexing on their websites. See more technical aspects at http://www.robotstxt.org/robotstxt.html. Accessed 12 March 2022.
[155] Hugenholtz B (2019) The New Copyright Directive: Text and Data Mining (Articles 3 and 4). http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/. Accessed 12 April 2022.
[156] Tyner A "The EU Copyright Directive: "Fit for the digital age or finishing it?" (2020) Journal of Intellectual Property Law 26 (2), p. 281.
[157] See Art. 70 of the Italian Legislative Decree.

of a website or a service,"[158] and (b) "*in case of content not made publicly available online, is clearly communicated to all persons who have lawful access to it."*[159] These details are definitely indicate positive aspects of national implementation of the DSM Directive.

Another important issue with regards to the TDM exceptions in Art. 3 and Art. 4 of the DSM Directive is that the provisions stipulate that both groups of beneficiaries are required to have "lawful access" to the mined works.[160] The "lawful access" is understood as covering access to data through for instance, an open access policy, subscriptions, and other lawful means, including access to content that is freely available online.[161] The requirement of "lawful access" presents a strict approach to the application of the exceptions as it would only depend on the will of rightholders that can limit or completely deny access to content.[162] The "lawful access" requirement may create difficulties for TDM projects, as additional costs will restrict the scope of research.[163] Therefore, it could be complicated for research organizations that cannot afford paid subscriptions and for other actors that do not have permission to access protected materials. For instance, it could be burdensome for start-ups and SMEs to obtain licenses from the rightholders, especially on reasonable terms that may undermine their innovative power.[164] As discussed above, only a limited number of actors will be able to obtain licenses for all works related to the particular TDM project.[165] AI development relies basically on large amounts of high-quality data used to train the algorithms. If the mined content is of low quality and incomplete, the outcomes of TDM will be scarce and unreliable.[166] Undoubtedly, it would be important to clarify the concept of "lawful access" through national transposition of the DSM Directive.

---

[158] Part 2, 4 (a) 53B (3) (a) of the Irish transposition law of the DSM Directive.
[159] Part 2, 4 (a) 53B (3) (b) of the Irish transposition law of the DSM Directive.
[160] Art. 3 (1) and 4 (1) of the DSM Directive.
[161] Recital 14 of the DSM Directive.
[162] European Copyright Society "General opinion on the EU copyright reform package" (2017) Available at https://europeancopyrightsocietydotorg.files.wordpress.com/2015/12/ecs-opinion-on-eu-copyright-reform-def.pdf Accessed 20 February 2022, p. 4.
[163] Geiger C "The exception for TDM in the proposed directive in the digital single market-legal aspects" (2018) Central of international IP studies Research paper n 2018-02: 1-34 https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/604941/IPOL_IDA(2018)604941_EN.pdf. Accessed 13 April 2022, p.22.
[164] Samuelson P "The EU's Controversial Digital Single Market Directive" (2018) Communications of the ACM 61 (11), p. 23.
[165] European Copyright Society "General opinion on the EU copyright reform package" (2017) Available at https://europeancopyrightsocietydotorg.files.wordpress.com/2015/12/ecs-opinion-on-eu-copyright-reform-def.pdf Accessed 20 February 2022, p. 3.
[166] European Copyright Society "General opinion on the EU copyright reform package" (2017) Available at https://europeancopyrightsocietydotorg.files.wordpress.com/2015/12/ecs-opinion-on-eu-copyright-reform-def.pdf Accessed 20 February 2022, p.3.

Another issue arises from the fact that only the provision of Art. 4 of the DSM Directive applies to the economic rights granted by the Software Directive[167], while research organisations and cultural heritage institutions are excluded from this privilege.[168] In this regard, it appears to be unclear why commercial entities should only be allowed to carry out TDM on software.[169] Is it an error in drafting the DSM Directive or there are other reasons for such exclusion? A possible answer could be that Art. 5 (3) of the Software Directive already provides for the "black box analysis" exception by allowing the lawful users to study, observe or test the functions of the program without the authorization of the rightholder.[170] As the research exception under Art. 5 (3) of the Software Directive already covers non-commercial reproduction of computer programs,[171] there was no need for Art. 3 of the DSM Directive to refer to the relevant provisions of the Software Directive.[172] In contrast, there was a necessity to clarify, by including the Software Directive in Art. 4 of the DSM Directive, that reproduction and adoption of computer programs outside the research context are allowed.[173]

Alternatively, research organizations and cultural heritage institutions may potentially rely on Art. 4 of the DSM Directive when performing TDM research on software. However, the exception in Art. 4 of the DSM Directive is likely no to be a perfect solution for research purposes as there could be a risk for the beneficiaries of Art. 3 to be blocked by the rightholders by a contract or TPMs.[174] Further, the interrelation between Arts. 4 of the DSM Directive and 5 (3) of the Software Directive can provoke legal uncertainty and create a paradox for actors involved in AI development since the activity of observing, studying, and testing the functioning of a program could be reserved by the rightholder if we apply Art. 4 of the DSM Directive, but it cannot be limited by a contract if we consider the Software Directive.[175] Consequently, if start-ups and SMEs performing TDM

[167] Art. 4 (a) and (b) of the DSM Directive.
[168] Art. 3 of the DSM Directive.
[169] Ducato R and Strowel A "Ensuring Text and Data Mining: Remaining issues with the EU copyright exceptions and possible ways out" (2021) CRIDES Working Paper Series No. 1/2021, p.14.
[170] Art. 5 (3) of the Software Directive.
[171] Art. 5 (3) of the Software Directive.
[172] Ducato R and Strowel A "Ensuring Text and Data Mining: Remaining issues with the EU copyright exceptions and possible ways out" (2021) CRIDES Working Paper Series No. 1/2021, p.14.
[173] Ducato R and Strowel A "Ensuring Text and Data Mining: Remaining issues with the EU copyright exceptions and possible ways out" (2021) CRIDES Working Paper Series No. 1/2021, p.15.
[174] Ducato R and Strowel A "Ensuring Text and Data Mining: Remaining issues with the EU copyright exceptions and possible ways out" (2021) CRIDES Working Paper Series No. 1/2021, p.15.
[175] Ducato R and Strowel A "Ensuring Text and Data Mining: Remaining issues with the EU copyright exceptions and possible ways out" (2021) CRIDES Working Paper Series No. 1/2021, p.15.

for training AI system cannot rely on the exceptions provided under Art. 5 (3) of the Software Directive and Art. 4 of the DSM Directive, they will always be obliged to enter into contractual relationships with the rightholder.[176] Therefore, it would be crucial to specifically allow TDM on software for the purpose of scientific research.

Based on the above discussions, one may come to a reasonable conclusion that even though the mandatory TDM exceptions introduced under Art. 3 and Art. 4 of the DSM Directive could be viewed as a significant step forward toward efficient regulation of TDM in the EU, the number of shortcomings shows that the legislative changes are insufficient to adequately cover the use of TDM in the context of AI development.

4. Conclusion

The legality of access to and usage of data for the purpose of TDM is becoming more and more important as an increasing number of businesses wish to incorporate AI technologies, which are based on TDM practices, into their business models. Therefore, there is a need for an open and technology-neutral legal regime in the EU that would successfully address technological development and innovation. However, due to the broad definition of the right to reproduction and the *sui generis* database right, TDM may potentially infringe the exclusive rights of copyright and database holders. In this regard, the exceptions and limitations that existed prior to the implementation of the DSM Directive failed to create legal certainty for start-ups and SMEs performing TDM for training AI systems, as they are limited in scope, strictly interpreted by the ECJ, and mostly optional that lead to the fragmentation in their application within the EU Member States.

To reflect those problems, the EU legislator introduced two mandatory exceptions under Arts. 3 and 4 of the DSM Directive in order to facilitate TDM research in both public and private spheres. However, both exceptions pose a number of drawbacks that may, in principle, impede the development of AI, including a narrow scope of Art. 3 of the DSM Directive limiting a group of beneficiaries and the purpose of TDM; an "opt-out" mechanism introduced in Art. 4 of the DSM Directive allowing rightholders to "reserve" the use of their content for the purpose of TDM; the difficulties in defining the requirement of "lawful access"; an ambiguity of the relationships

[176] Ducato R and Strowel A "Ensuring Text and Data Mining: Remaining issues with the EU copyright exceptions and possible ways out" (2021) CRIDES Working Paper Series No. 1/2021, p.15.

between the DSM Directive and the Software Directive and so on. Indeed, the wording of Art. 3 and Art. 4 of the DSM Directive reduces the positive effect of the legislative changes on the regulation of TDM by mostly favoring rightholders interests and empowering them to strictly control the use of their works. The mandatory exceptions appear to be insufficient to regulate TDM effectively so as to ensure a balance between the rightholders' rights and interests and those of TDM users.

Therefore, currently, start-ups and SMEs involved in TDM activities are not in a very favorable position that may undermine the development of AI technologies in the EU. In this regard, the national implementation of the DSM Directive through the EU Member States may be viewed as a good opportunity to create a more favorable legal framework for TDM. The DSM Directive indicates that the EU Member States are allowed to 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"[177] "as long as they do not limit the scope of the mandatory exceptions or limitations"[178] introduced with the DSM Directive.

The EU Member States may, for instance, extend the scope of the TDM exception for scientific research to cover additional groups of beneficiaries, such as individual researchers and entities that do not pursue commercial purposes. The national laws of Member States may introduce provisions that could clarify the notion of "scientific research" that could be interpreted in a broad manner as to involve any research activity conducted in accordance with relevant methodological standards for the purpose of generating new knowledge and insights in the relevant field of study. The EU Member States could also explicitly permit TDM for other institutional objectives such as educational or administrative purposes, including teaching illustration. Moreover, to bring more certainty and highlight the necessity of research work, it seems relevant to allow third parties that are involved in the verification process of the quality of scientific research to freely access the "collection" of the copies of the protected content generated during TDM. The national laws of Member States may introduce provisions that could clarify the notion of the "opt-out" mechanism to avoid unnecessary restrictions of TDM through, for instance the proper interpretation of the reserved use that should be done "in an appropriate manner."

[177] Art. 25 of the DSM Directive.
[178] Recital 5 of the DSM Directive.

However, the overwhelming majority of the EU Member States that have already implemented the DSM Directive in their national law have transposed the TDM exceptions in their literal sense by barely copy-pasting the wording of the DSM Directive. If the EU Member States do not take any decisive steps to eliminate the legal uncertainty of TDM exceptions, the national implementation is unlikely to introduce meaningful and substantial improvements to the TDM exceptions in a way that would secure sufficiently AI development.