International Journal of Legal Information 2025, pp. 1–12. © The Author(s), 2025. Published by International Association of Law Libraries
doi:10.1017/jli.2024.46

Development of Large Language Models: Copyright Law Perspectives for Research Institutions and Research Libraries

Inger Berg Ørstavik*

Abstract

This article discusses European copyright law as applied to the development and training of generative AI and natural language processing in public interest research institutions and libraries. The article focuses on the scope of the new exceptions from copyright law for text and data mining (TDM) for research purposes and discusses them from the perspective of research ethics and principles of open science in publicly financed research. The public interest mission of research institutions and libraries includes the open dissemination of research results but the exceptions from copyright are focused only on the training phase in AI development. Regulation on data transparency is fragmented. The article finds that while new exceptions open for developing language models under research institutions and libraries' public interest mission to preserve national languages, the regulation is not adapted to principles of research ethics and open science, and legal uncertainty remains.

Keywords: copyright, generative AI, research institutions, open science, EU law

Introduction

The digital processing of data opens new, ground-breaking possibilities in research. Data analysis techniques, generally classified as text and data mining (TDM) techniques, make it possible to draw new information from existing research and data. TDM techniques are also used to develop generative artificial intelligence (AI) systems, including large language models (LLMs). Such models are likely to become standard working tools for efficient text production in the future, and research institutions and libraries are engaging to develop models in national languages as part of their public interest mission to preserve and manage national cultural heritage and languages.[1]

For researchers, TDM methods are powerful tools for understanding and making use of existing data and information. The combination of unlimited "reading" capability and the ability to analyze huge amounts of data using statistical and mathematical methods yields new research results in the forms of new learning and understanding as well as the ability to make more accurate predictions using automated tools. With generative AI, research activities are turning towards the development of AI models, including language models. Research institution libraries are in a unique position to train and develop national language LLMs due to their access to large collections and repositories of literature and textual materials. The public interest mission of publicly financed research institutions and libraries, however, demands that research activities adhere to principles of research ethics and that results are made available to benefit society.

Editor's Note: This article is based on the author's presentation given at the 42nd Annual Course of the International Association of Law Libraries held in Oslo, Norway, 14 – 20 June 2024.

*Department of Private Law, University of Oslo, Oslo, Norway. Email: i.b.orstavik@jus.uio.no.

[1] See, for instance, the models under development by the National Library of Norway, accessed Oct. 8, 2024, https://ai.nb.no/models.

The recognition of public interest in such research activities and the need for research institutions and cultural heritage institutions to engage in TDM activities without legal uncertainty have been recognized by the European Commission. In response, an exception from copyright—that is, database rights and press publishers' rights in works and data for TDM for scientific research—has been included in EU copyright law with the DSM Directive.[2] The DSM Directive was drafted and enacted before the potential of generative AI was generally recognized. However, with the enactment of the AI Act on May 21, 2024,[3] it was made clear that, for the use of copyright-protected works in AI training, the AI Act relies on the TDM provisions in the DSM Directive.[4]

This paper discusses the scope and application of Article 3 of the DSM Directive, which addresses LLM training in research institutions. The discussion will also apply to other generative AI using copyright-protected works for training, such as AI tools for generating music or visual art. The article considers these questions from the perspective of research ethics and principles of open science in publicly financed research. The legal framework considered is European copyright law in the European Union (EU) and European Economic Area (EEA).

The Role of Research Institutions and National Libraries in Training LLMs

Training advanced AI models requires access to vast amounts of data. In natural language processing (NLP), the training materials are texts in the relevant language. National libraries and research libraries are in a unique position to develop innovative and high-quality LLMs due to their access to vast collections of texts and literature. Access to collections and repositories of research articles and materials enables the development of models applying professional language and qualitative research methods. Access to newspapers has proven valuable for training models for sentiment analysis and general knowledge.[5] Many of these collections are already digitized, although investment may be required for data cleaning and processing for machine-learning purposes. Because of the vast amount of training materials and the complex training required, the costs of developing LLMs for lesser-used national languages may be too high to motivate commercial for-profit developers. For these reasons, there is an important role for research institutions and libraries to play in developing and training LLMs as part of their public interest mission to preserve cultural heritage and national languages.

The public interest mission of research institutions and libraries includes the dissemination of research results for the benefit of society. Research results should be made openly available for private and commercial actors as bases for further research, including for-profit innovation.[6] An assessment of the legal framework applicable to libraries and research institutions when training LLMs should therefore include the conditions for making the model openly accessible. This adds to the complexity of applying EU copyright law, which is fragmented and mainly focuses on the training phase. Furthermore, principles of research ethics and open science go beyond the narrow

[2] 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 (DSM Directive). Art. 3(1): "Member States shall provide for an exception to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, and Article 15(1) of this Directive for 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." See also recitals (8) and (10).

[3] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence (AI Act).

[4] Recital (105), AI Act: "Any use of copyright protected content requires the authorization of the rightsholder concerned unless relevant copyright exceptions and limitations apply. Directive (EU) 2019/790 introduced exceptions and limitations allowing reproductions and extractions of works or other subject matter, for the purpose of text and data mining, under certain conditions. Under these rules, rightsholders may choose to reserve their rights over their works or other subject matter to prevent text and data mining, unless this is done for the purposes of scientific research."

[5] *Cf.* Report from the Norwegian National Library, *Evaluating the effect of copyright protected materials in generative large language models for Norwegian languages* (2024), https://www.nb.no/content/uploads/2024/08/Mimirprosjektet_teknisk-rapport.pdf. Norwegian language.

[6] *Cf.* Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (Open Data Directive), Art. 10. From national law, the Norwegian Universities and Colleges Act § 2-1 requires institutions to contribute to innovation and value creation based on research results. See also the EU Commission's open science policy, https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en.

https://doi.org/10.1017/jli.2024.46

scope of exception from copyright for scientific research in Article 3 of the DSM Directive, challenging the scope of the exception.

SCOPE OF THE EXCEPTION IN ARTICLE 3 OF THE DSM DIRECTIVE

TDM and the Development of AI

Machine-learning systems use algorithms and statistical models to draw inferences from patterns in data to make predictions or decisions without being explicitly programmed to do so.[7] In generative AI, the model will continue its training and development when applied to new data through self-assessment and feedback methods. The information extracted during training is stored in a separate file and added to the model's set of "rules," which it will use to make more accurate predictions when applied to new data.[8] TDM uses similar statistical and mathematical techniques, but the objective is to present the extracted information. Since the definition of TDM in the DSM Directive only covers the process of extracting information and not the use of the information generated, machine-learning processes for AI will mostly be covered by the definition of TDM in the DSM Directive.[9]

Under Article 3 of the DSM Directive, EU Member States are obliged to provide for an exception to copyright[10]—that is, copyright, database rights,[11] and press publishers' rights.[12] In theory, the exception should pave the way through the layers of exclusive rights in collections and repositories of works for the purpose of TDM in research. The objective of the exception was to permit usage types that were not covered clearly enough by existing EU rules on exceptions and limitations and to harmonize variations in national legislation regarding TDM for research resulting from the optional character of other exceptions and limitations.[13]

Copyright: The Right of Reproduction

Books, articles, papers, poems, and other literary expressions are subject to copyright under European copyright law.[14] Only the simplest, most elemental texts will not qualify for protection.[15] Due to the massive amount of text required to train a language model, it can safely be assumed that access to a large amount of copyright-

[7] See, e.g., Josef Drexl et al., "Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective," *Max Planck Institute for Innovation & Competition Research Paper no. 19-13* (Oct. 2019): 4, https://ssrn.com/abstract=3465577.

[8] See, further, Inger B. Ørstavik, "Access to data for training algorithms in machine learning: copyright law and 'right-stacking,'" in *Artificial Intelligence and the Media*," eds. Taina Pihlajarinne and Anette Alén-Savikko (Cheltenham, UK: Edward Elgar Publishing, 2022): 272, 276–78; David Lehr and Paul Ohm, "Playing with the Data: What Legal Scholars Should Learn About Machine Learning," *UC Davis Law Review* 51 (2017): 653, 655–717; Thomas Margoni, "Artificial Intelligence, Machine learning and EU copyright law: Who owns AI?," *CREATe Working Paper* 12 (2018): 4–5, DOI:10.5281/zenodo.2001763; Mauritz Kop, "Machine Learning & EU Data Sharing Practices," *Stanford-Vienna Transatlantic Technology Law Forum*, 1 (2020): 7.

[9] Art. 2(2) of the DSM Directive defines text and data mining as "any automated analytical technique aimed at analyzing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations," ref. Recital (18). See Romain Meys, "Data Mining Under the Directive on Copyright and Related Rights in the Digital Single Market: Are European Database Protection Rules Still Threatening the Development of Artificial Intelligence?," *GRUR Int*. 69, no. 5 (2020): 457, 464–65. DOI: 10.1093/grurint/ikaa046.

[10] Art. 2 of Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonization of certain aspects of copyright and related rights in the information society (Infosoc Directive).

[11] Arts. 5(a) and 7(1) of Directive 96/9/EC of the European Parliament and the Council of 11 March 1996 on the legal protection of databases (Database Directive).

[12] Art. 15(1), DSM Directive.

[13] Recital (5), DSM Directive.

[14] Art. 2, Infosoc Directive.

[15] The CJEU has found that a text of only eleven words could enjoy copyright protection if it is the expression of the intellectual creation of the author, Case C-05/08, *Infopaq I*, ECLI:EU:C:2009:465, 39, 47–48.

protected works is necessary.[16] The introduction of the exceptions for TDM in the DSM Directive confirmed the presumption that machine learning will infringe copyright in the training materials.[17] To assess the legal framework applicable to publicly financed research institutions, it is useful to take a brief look at which phases of development of a natural language model could infringe EU copyright law.

Under EU copyright law, the "temporary or permanent reproduction by any means and in any form, in whole or in part" is reserved for the copyright holder.[18] The Court of Justice of the European Union (CJEU) has construed the provision broadly, extending the right to every act of reproduction.[19] All digital copies of a work are considered an act of reproduction, including copies in the RAM or cache memory of a computer, even if such copies may be intrinsic to (lawfully) accessing a work by computer, such as online browsing.[20] This very broad and formal construction of the right of reproduction has been criticized for going beyond the fundamentals of copyright.[21] Also, such use does not interfere with the author-audience nexus of free speech and enlightened human communication at copyright's core.[22]

Already, the compilation of a centralized training corpus would entail several acts of reproduction infringing Article 2 of the Infosoc Directive. To train an LLM on works in a library collection, texts may have to be digitized or transformed from digital human-readable formats such as PDF or similar machine-readable formats.[23] Pre-processing the data (that is, sorting out outliers and other cleaning of the training data, as well as adding metadata and annotations for training) likely requires temporary copying in the computer's cache memory, which is also covered by the right of reproduction in Article 2 of the Infosoc Directive.[24]

In the training process, the model goes back and forth between the training data, testing, and modifying its "rules," yielding statistical information about correlations, trends, differences, and the like in the training data.[25] This information enables a model in basic form to predict the most likely next word and finally produce large amounts of text. The original works are not recognizable in the model's stored files, even if the model can be prompted to produce text that is identical or very similar to specific training materials. During the training process, it is therefore likely that copies are created, which are considered reproductions under Article 2 of the Infosoc Directive, even if no human-readable copies are made.[26] The stored files in the model itself, however, do not necessarily include reproductions of

[16] *That is*, provided that the model is intended to apply a reasonably modern language and not train only on older works for which the copyright has expired.

[17] Recital (8), DSM Directive, and recital (105), AI Act.

[18] Art. 2, Infosoc Directive.

[19] Michel M. Walter and Silke von Lewinsky, eds., *European Copyright Law: A Commentary* (Oxford: Oxford University Press, 2010), 968; recital (21), Infosoc Directive.

[20] Ibid. See also Inger B. Ørstavik (n 8).

[21] See, in general, P. Bernt Hugenholtz, ed., *Copyright Reconstructed: Rethinking Copyright's Economic Rights in a Time of Highly Dynamic Technological and Economic Change* (Alphen aan den Rijn, The Netherlands: Wolters Kluwer, 2018). Under US law, this reasoning has led to the application of the exception of "fair use" to machine learning albeit not without friction; see, e.g., Mark A. Lemley and Bryan Casey, "Fair Learning," *Texas Law Review* 99, no. 4 (Mar. 2021): 743–86.

[22] Alain Strowel, "Reconstructing the Reproduction and Communication to the Public Rights: How to Align Copyright with its Fundamentals," in ed. P. Bernt Hugenholtz, *Copyright Reconstructed: Rethinking Copyright's Economic Rights in a Time of Highly Dynamic Technological and Economic Change* (Alphen aan den Rijn, The Netherlands: Wolters Kluwer, 2018), 206–09.

[23] See Jean-Paul Trialle, *Study on the legal framework of text and data mining*, EC Commission (2014), 32, 29–31, https://op.europa.eu/en/publication-detail/-/publication/074ddf78-01e9-4a1d-9895-65290705e2a5/language-en. It does not matter if the materials are deleted after completion of the training process; *cf.* Art. 5(1), Infosoc Directive.

[24] Christophe Geiger et al., "Text and data mining in the proposed copyright reform: making the EU ready for an age of big data? Legal analysis and policy recommendations," *IIC* 49, no. 7 (2018): 814–44, 818, https://doi.org/10.1007/s40319-018-0722-2.

[25] Some doubt has been expressed in the literature as to whether a model makes copyright-relevant copies when running on training materials; see Meys 460 (n 9); and Geiger (n 24). This likely depends on the model in question; *cf.* also the statements in recitals 8 and 9 of the DSM Directive.

[26] Case C-360/13, *Meltwater*, ECLI:EU:C:2014, 1195. See also Rossana Ducato and Alain Strowel, "Limitations to text and data mining and consumer empowerment: making the case for a right to 'machine legibility,'" *IIC* 50, no. 6 (2019): 649–84, 658; Christoph Geiger et al., "Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU," *CEIPI Studies Research Paper no. 2019-08*, https://ssrn.com/abstract=3470653, 8. The author's right is infringed if so much of the work is copied that it includes subject matter that is "original in the sense that it is its author's own intellectual creation," Case C-05/08, *Infosoc I*, ECLI:EU:C:2009:465, 37.

works, and when the model is applied to new data or to produce text, it would in most cases not entail the reproduction of training materials. There may be some variation between models.

Database Rights

Research institutions and libraries also have access to works through subscriptions to other collections. Collections may be protected databases under Article 7(1) of the Database Directive, protecting databases for which "obtaining, verification or presentation of the contents" requires "a substantial investment." The *sui generis* right to databases gives the right holder an exclusive right of extraction as defined in the Database Directive, Article 7(2)(a), as "the permanent or temporary transfer of all or a substantial part of the contents of a database to another medium by any means or in any form."[27] Using a protected collection for training AI would most likely infringe this right of extraction, as it also covers the "repeated and systematic" extraction of insubstantial parts of the database.[28] Whereas searching the database may entail digitally copying contents, this does not infringe the extraction right if the consultation is lawful.[29]

The *sui generis* right is infringed if the investment in making the database is harmed, which it is if the right holder is deprived of revenue that should have enabled the holder to redeem the cost of investment.[30] Machine learning reveals new knowledge and facilitates new and innovative services, activities that would not likely be characterized as "parasitical competing" activities infringing the *sui generis* right.[31] The methods used for machine learning, however, make use of the economic value associated with the database containing a large collection of works. It is therefore unlikely that machine learning would be considered a "normal exploitation" of a database under Article 8(2) of the Database Directive if not explicitly included in a license.[32] For activities that appropriate the value inherent in the database, the CJEU has found it legitimate for the database holder to reserve a fee in consideration for such use.[33] It therefore seems likely that the CJEU would find machine learning and AI training to be harmful to the investment in the database that the *sui generis* right protects.[34]

Research institutions and libraries also hold repositories of works, and such repositories could be subject to database rights. A recent EU regulation points in the direction that, for public institutions, the objective of data transparency would supersede that of protecting database rights in collections of data.[35] Under Article 1, nr. 6 of the Open Data Directive, public sector bodies are prohibited from exercising their database rights to prevent the reuse of documents or to restrict reuse beyond the limits set by the Open Data Directive.[36] The Open Data Directive applies to universities as well as bodies governed by public law.[37] It requires that research data be made openly available, defining research data as digital documents "collected or produced in the course of scientific research activities."[38] The definition excludes scientific publications. However, as pointed out in the preamble to the Open Data Directive, research output should already be available under open-access policies. Open access is understood as "the practice of providing online access to research outputs free of charge for the end user and without restrictions on use and reuse

[27] Recital (44), Database Directive.

[28] Art. 7(5), Database Directive. Ref. discussion in Inger B. Ørstavik (n 8).

[29] Case C-203/02, *William Hill*, ECLI:EU:C:2004:695, 54; Case C-304/07, *Directmedia*, ECLI:EU:C:2008:552, 51.

[30] Recital (49), Database Directive, and Case C-203/02, *William Hill*, ECLI:EU:C:2004:695, ¶ 51; Case C-202/12, *Innoweb*, ECLI:EU:C:2013:850, 37; Case C-304/07, *Directmedia*, ECLI:EU:C:2008:552, 33 and 35.

[31] Recital (42), Database Directive; Case C-203/02, *William Hill*, ECLI:EU:C:2004:695, 47.

[32] Recital (24), Database Directive.

[33] Case C-203/02, *William Hill*, ECLI:EU:C:2004:695, 57.

[34] See Case C-490/14, *Verlag Esterbauer*, ECLI:EU:C:2015:735, ¶ 16; Case C-202/12, *Innoweb*, ECLI:EU:C:2013:850, 46–48. See Lemley and Casey: 127 (n 21); Geiger et al. (2018), 823–24 (n 24).

[35] *Cf.* also Thomas Margoni and Martin Kretschmer, "A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology," *GRUR International* 71, no. 8 (2022): 685–701, https://doi.org/10.1093/grurint/ikac054. See also recital (107), AI Act, where it is assumed that public databases may be used for AI development based on the public interest in data transparency and data accountability.

[36] Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information; recital (107), AI Act.

[37] Arts. 1(2) and (4), recital (4), Open Data Directive.

[38] Art. 10 ref. Art. 2(9), and recital 27–28, Open Data Directive.

https://doi.org/10.1017/jli.2024.46

beyond the possibility to require acknowledgement of authorship."[39] For public research institutions, it could be contrary to their public interest mission of open access to research and research output as expressed in the Open Data Directive to invoke *sui generis* database rights to restrict the use of open-access repositories for training LLMs, even for commercial purposes. Research institutions engaging in the development of AI or LLMs would also be able to use the repositories of other public research institutions to which they have lawful access for natural language processing activities. Letting the public interest in data transparency prevail over the economic interests of data holders as protected under the *sui generis* database right is consistent with other recent EU regulations.[40]

Press Publishers' Rights

The exception in Article 3 also covers the newly introduced press publisher's right under Article 15 of the DSM Directive. Under Article 15, press publishers are granted full exclusive rights as in Articles 2 and 3(2) of the Infosoc Directive, but only against "the online use of their press publications by information society service providers" and only for a period of two years from publication. An information society service is "any service normally provided for remuneration, at a distance, by electronic means and at the individual request of a recipient of services"—that is, any service provided individually over the internet.[41] Many research institutions and libraries subscribe to various press publications. While it could be questioned whether the use of press publications by public interest research institutions and libraries for the purpose of scientific research would infringe Article 15, the inclusion of the press publisher's right in the exception in Article 3 removes any legal uncertainty as to whether these collections are accessible for the development of LLMs within the scope of the exception.

The Condition of "Lawful Access"

The exception in Article 3 of the DSM Directive applies only to works or other subject matter to which the institutions have "lawful access." This includes the institutions' own collections, such as repositories of research output and data, as well as digitized collections of lawfully acquired physical copies. Lawful access can be based on a license or a subscription to a collection of works or through an open-access license.[42] The exception also applies to content that has been made available to the public online without reservation for TDM.[43] In addition, the exception will cover works that have been donated to the institution as well as works made available to the institution through licensing arrangements required under legislation or as a result of the implementation of limitations to copyright in national law.[44]

The exception covers the research activities of persons "attached" to the organization. In the case of subscriptions, this will also depend on the terms of the subscription agreements.[45]

Finally, the condition of "lawful access" must be distinguished from the concept of "lawful use" in Article 5 (1) of the Infosoc Directive. Under this article, a use is lawful if it is either authorized by the right holder or falls outside the scope of the author's exclusive right.[46] The CJEU has applied the exception in Article 5(1) to services that

[39] Recital (27), Open Data Directive.

[40] Ref. Art. 43 of Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonized rules on fair access to and use of data (Data Act), Article 50 and Article 10 (data quality) AI Act, Article 27 and recital (70) of Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services (Digital Services Act). *Cf.* also Thomas Margoni and Martin Kretschmer (2022), 699–700 (n 35).

[41] Directive (EU) 2015/1535 of the European Parliament and of the Council of 9 September 2015 laying down a procedure for the provision of information in the field of technical regulations and of rules on Information Society services Article 1(1)(b).

[42] Recital (10), DSM Directive.

[43] Recital (14), DSM Directive.

[44] A curious example is the Norwegian Act of 9 June 1989 no. 32 relating to the Legal Deposit of Generally Available Documents, under which publishers are obligated to deposit copies of all published works with the National Library. It is not entirely clear whether the deposited works may lawfully be used for training AI under Art. 3 of the DSM Directive.

[45] Recital (14), DSM Directive.

[46] Case C-302/10, *Infopaq II*, ¶ 42, ECLI:EU:C:2012:16, and Case C-403/08, *Premier League*, 168, ECLI:EU:C:2011:631, and recital (33), Infosoc Directive.

https://doi.org/10.1017/jli.2024.46

have as their output excerpts of the works so small that they do not reproduce the "expression of the intellectual creation of the author."[47] A language model trained to produce text will not aim to reproduce the text of the works on which it has been trained but may still happen to do so. For the exceptions in Articles 3 and 4 of the DSM Directive, to fulfill their objective of balancing the interests of right holders against those of users,[48] the exception should be applied to any acts of reproduction during the development and training of LLMs, but without regard to the output from using the model. Considering the model's output lawful because the training of the model was lawful would tip the scales. A separate assessment of the output allows for the application of Article 5(1) of the Infosoc Directive when appropriate and a finding of copyright infringement in cases of memorization or the override of paywalls, etc.

BENEFICIARIES OF THE EXCEPTION IN ARTICLE 3 OF THE DSM DIRECTIVE

"Research organization" is broadly defined in Article 2(1) of the DSM Directive and includes universities, libraries, research institutes, and hospitals that carry out research.[49] The scope of the exception in Article 3 hinges upon a functional definition of research organizations to allow for the different legal forms and structures of research organizations in EU Member States.[50] The organization must have as its primary goal to conduct scientific research or to carry out educational activities also involving scientific research.[51] The organization must be involved in scientific research either "on a not-for-profit basis or by reinvesting all the profits in its scientific research," or "pursuant to a public interest mission recognized by a Member State" per Article 2(1) of the DSM Directive. A public interest mission will often be reflected in the public funding of universities and their libraries, but it could also be reflected in provisions in national laws or in public contracts.[52]

"Cultural heritage institution" is defined in Article 2(4) as "a publicly accessible library or museum, an archive or a film and audio heritage institution." It does not matter what types of works or data that the institution holds in its permanent collection. The definition also includes educational establishments, research organizations, and public sector broadcasting organizations.[53]

The exception aims to enhance legal certainty to the benefit of the research community but only insofar as the research is conducted in a way that also adheres to the values of independent and open research. In organizations where commercial undertakings have a decisive influence, allowing such undertakings to exercise control through their shareholders or members, which could result in preferential access to the research results, falls outside the definition of research organizations in the directive.[54] Some private universities and research organizations are owned by foundations that are either not-for-profit or reinvest profits in research. To benefit from the exception, their statutes or boards should be set up with explicit guarantees against preferential access to research results. The directive also encourages public-private partnerships (PPPs) in research and for public research organizations to use private partners to carry out TDM, including using their technological tools.[55]

Using a broad definition of research organizations provides flexibility for large research initiatives that rely on funding from a combination of sources. The DSM Directive does not require that the organizations be established in the EU. As long as an organization is covered by the definition in Article 2(1), the directive covers research activities taking place within the EU, in line with the stated objective to ensure the EU's competitive position as a research area.[56] The DSM Directive has been criticized for excluding startups and individual researchers.[57]

---

[47] Case C-05/08, *Infopaq I*, ¶ 39, ECLI:EU:C:2009:465; Art. 3(1), Infosoc Directive.
[48] Recital (6). DSM Directive.
[49] Recital (12), DSM Directive.
[50] Ibid.
[51] Art. 1(2) and recital (12), DSM Directive.
[52] Recital (12), DSM Directive.
[53] Recital (13), DSM Directive.
[54] Art. 1(2) and recital (12), DSM Directive.
[55] Recital (11), DSM Directive.
[56] Recitals (8) and (10), DSM Directive.
[57] A. Dermawan, "Text and data mining exceptions in the development of generative AI models: What the EU Member States could learn from the Japanese 'nonenjoyment' purposes?" *Journal of World Intellectual Property* 27, no. 1 (2023): 44, 52–53, https://doi.org/10.1111/jwip.12285; Christophe Geiger, "The Missing Goal-Scorers in the Artificial Intelligence Team: of Big Data, the Fundamental Right to Research and the failed Text and Data Mining Limitations in the CSDM Directive," *PIJIP/TLS Research Paper Series no. 66* (2021), in *Intellectual Property and Sports, Essays in Honour of P. Bernt Hugenholtz*, eds. Martin Senftleben et al. (Alphen aan den Rijn, The Netherlands: Kluwer Law International, 2021), 383, 388.

https://doi.org/10.1017/jli.2024.46

Commercial services may serve the important public interest of information integrity, such as fact-checking services.[58] However, both startups and commercial service development go beyond the purpose of publicly financed scientific research, and they will not always be able to guarantee adherence to principles of independent and open research. When interpreting Article 3, it should be taken into account that it is part of the EU research policy as anchored in the Treaty on the Functioning of the European Union (TFEU), Article 179.[59] Integral to EU research policy are principles of independent and open research.[60] Whether the organization is structured and operates in a way that guarantees these principles, especially regarding the dissemination of research results, could guide the interpretation of the exception in Article 3.

Text and Data Mining "For the Purposes of Scientific Research"

The exception in Article 3 of the DSM Directive applies only to TDM "for the purposes of scientific research." This includes both the natural sciences and human sciences.[61] Research activities are characterized by their aim to yield new knowledge. There is no clear definition of when research is scientific, but a basic condition would be that the activity adheres to general principles of methodology and ethics relevant to the subject.[62] For AI development, like research activities in general, ethical research must respect all applicable national, EU, and international laws.[63] Ethical guidelines for AI development particularly point to intellectual property rights and the protection of personal data.[64] Furthermore, the training and development of the model would have to comply with principles of fairness, representativity, transparency, and accountability.[65] In practice, this means that it should be open to what materials have been used for training and that the researchers are aware of the impact that different types of training materials could have on the model.[66] Since it is part of the public interest mission of research institutions to disseminate research results to benefit society, the fact that a language model is developed with the goal of releasing it into society should not exclude the application of the exception in Article 3 of the DSM Directive.

Actions Covered by the Exception in Article 3 of the DSM Directive

The exception in Article 3 is mandatory and cannot be overridden by contract.[67] The exception is also made without payment to right holders.[68] This means that Member States may not envisage a compensation requirement when implementing the exception.[69] However, there is some concern that the prices for subscriptions and licenses will increase or that publishers and database owners will employ licensing strategies with differentiated prices for

[58] Recital (18), DSM Directive. These entities may benefit from the exception in Art. 4 of the DSM Directive, which, requires that right holders are given the opportunity to "opt-out."

[59] Treaty on the Functioning of the European Union, signed on 13 December 2007 (TFEU); *cf.* recital (12), DSM Directive.

[60] Recital (27), Open Data Directive; European Code of Conduct for Research Integrity (rev. ed. 2023), https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/european-code-of-conduct-for-research-integrity_horizon_en.pdf.

[61] Recital (12), DSM Directive.

[62] European Code of Conduct for Research Integrity, 3 (n 60).

[63] EU Commission, *Living Guidelines on the Responsible use of Generative AI in Research* (Mar. 2024), https://research-and-innovation.ec.europa.eu/document/download/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en?filename=ec_rtd_ai-guidelines.pdf.

[64] Ibid.

[65] Ibid.; EU Commission, High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI* (2019), https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai; recital (27), AI Act.

[66] *Cf.* Report from the Norwegian National Library (n 5).

[67] Recitals (10), (17), and (18), DSM Directive.

[68] Recital (17), DSM Directive.

[69] Eleonora Rosati, *Copyright in the Digital Single Market: Article-by-Article Commentary to the Provisions of Directive 2019/790* (Oxford: Oxford Academic, 2021), "Article 3 - Text and Data Mining for the Purposes of Scientific Research," 41, https://doi.org/10.1093/oso/9780198858591.003.0004.

TDM.[70] In addition to driving up the price of research for public interest, this may also negatively affect the quality of AI models.[71] Price and license restrictions may lead institutions to train their models with materials of poorer quality or a smaller data set. This could be problematic when developing AI models in research, as it could challenge research standards and methodology. The general principle of EU law effectiveness could probably be invoked to support enforcement action against contract practices that deprive the Article 3 exception of its effectiveness.[72] However, such litigation would likely be time-consuming and costly due to the legal as well as factual uncertainty.

Right holders are allowed to apply measures to ensure the security and integrity of networks and databases where the works or data are stored per Article 3(3) of the DSM Directive. Such measures may include Internet Protocol (IP) address validation or user authentication to ensure that only persons having lawful access to the works can access them.[73] Measures must be proportionate to the risks involved and not exceed what is necessary to ensure the system's security and integrity.[74] These measures must not undermine the effective application of the exception.[75] The DSM Directive is not very specific in terms of what measures might be applied or how to address overly restrictive measures or the circumvention of such measures. Protection measures should be limited to mitigating the security risks connected with the lawful use of the works.[76] It is likely that protective measures will be defined in best practices or similar measures as encouraged in Article 3(4) of the DSM Directive.

Under Article 3(2), the works may be stored and retained for "the purposes of scientific research." This includes verification of research results. In practice, this would indicate that downloading and compiling a training corpus specifically for AI training is lawful and that it could be stored for some time, perhaps as long as the research output (that is, the trained AI) is openly accessible and can be used as a basis for further research and therefore needs to be verified. This means that more permanent copying is covered by this exception than by the exception for temporary reproductions in Article 5(1) of the Infosoc Directive.

A further question is whether a training corpus could be reused for developing other models once it is processed, annotated, and stored. It is costly to process a training corpus even if a research institution or library already has access to large collections of works and data. The wording of the DSM Directive does not exclude reusing data if the new research activity is covered by the exception in Article 3. If it would be lawful to process and store the training corpus for the new activity under Article 3, an existing corpus may be reused at least within the same organization. Whether a processed training corpus could be licensed or sold to other research organizations for scientific research covered by Article 3 is doubtful, as licensing or selling a corpus could go beyond the normal exploitation of works and prejudice the interest of database right holders or press publishers' rights in the corpus.[77] Article 3 would not cover the licensing of a training corpus to commercial actors. Such use would have to be considered under Article 4 of the DSM Directive; see later in this paper for a more in-depth discussion of this topic.

If copies are kept, they should be stored in a secured environment, and the DSM Directive nudges Member States to appoint trusted bodies for managing repositories.[78] However, any such requirements should be proportionate and not go beyond what is needed for retaining the copies in a safe manner and preventing unauthorized use.[79]

---

[70] Geiger et al. (2019), 36–37 (n 26).

[71] Thomas Margoni and Martin Kretschmer (2022), 700 (n 35).

[72] In this direction, with regard to technical protection measures, see recital (16) DSM Directive. See also Case C-403/08, *Premier League*, ECLI:EU:C:2011:631, 163.

[73] Recital (16), DSM Directive.

[74] Art. 3(3) and recital (16), DSM Directive.

[75] Recital (16), DSM Directive.

[76] The provision in Art. 3(3) of the DSM Directive is related to the provision of technical protection measures in Art. 6(3) of the Infosoc Directive. However, technical protection measures under the Infosoc Directive should mitigate the risk of unlawful use of works, and the balancing of interests under this provision is, therefore, fundamentally different from that under Art. 3 (3) of the DSM Directive. Christoph Geiger et al. (2019) (n 26), 34, points to the regulation of traffic management measures in Art. 3(3) of Regulation (EU) 2015/2120 of the European Parliament and of the Council of 25 November 2015, laying down measures concerning open internet access as a better guide for how to interpret Art. 3(3) of the DSM Directive.

[77] Art. 5(5), Infosoc Directive, and recital (6), DSM Directive.

[78] Recital (15), DSM Directive.

[79] Ibid.

https://doi.org/10.1017/jli.2024.46

Finally, Article 7(2) of the DSM Directive prescribes that the three-step test in Article 5(5) of the Infosoc Directive shall apply.[80] Accordingly, the exceptions in the DSM Directive may only apply to special cases that don't conflict with the normal exploitation of the work and do not unreasonably prejudice the right holder's legitimate interests.[81] The CJEU has repeatedly stressed that the exceptions from copyright must be interpreted strictly.[82] The Court has also stated that the conditions for exception must be interpreted so as to "enable the effectiveness of the exception thereby established to be safeguarded and permit observance of the exception's purpose."[83] Copyright should not be detrimental to the development of new technologies, and the Court has emphasized the need to strike a fair balance between the right holder's interests and the users of works that implement new technologies.[84] In recent case law, the CJEU has pointed to online media users' right to information and characterized this right as a fundamental right enshrined in the EU Charter of Fundamental Rights, Article 11.[85] This means that internet users must be able to trust that website publishers have fulfilled their obligation to obtain sufficient consent to online use from individual right holders.[86]

Applied to Article 3 of the DSM Directive, there is, on the one hand, the public interest in innovation based on access to information that would call for a broad interpretation of the exception. On the other hand, for the public interest in the right to information and information integrity, represented by LLM users, the exception must be interpreted in a way that ensures that the model can be lawfully used without further rights clearance. In theory, this may include users' interests in the interpretation of Article 3 of the DSM Directive. However, it remains to be seen how the CJEU will approach the three-step test under Article 3, especially since the balancing of interests is more complex with the inclusion of database rights and press publishers' rights under the exception. These rights protect interests in the investments in the collection, which differ from the right holders' interests in individual works.[87]

Open Access to Research Output and Article 4 of the DSM Directive

When the exception in Article 3 of the DSM Directive is directed at research organizations and libraries carrying out scientific research pursuant to a public interest mission, the institutions follow EU principles on open science and dissemination of research results, which are anchored in the TFEU, Articles 179 through 183. The principles of open science and open-access dissemination of research output are mandatory under the EU Horizon programs,[88] a practice that national research councils have also adopted. Open science and open dissemination of results is implemented in the strategy of most publicly financed research organizations.[89] The relevant organizations

[80] See the discussion in Inger B. Ørstavik, 287–89 (n 8).

[81] Recital (6), DSM Directive.

[82] Case C-05/08, *Infopaq I*, ¶ 56, ECLI:EU:C:2009:465; case C-302/10, *Infopaq II*, para 27, ECLI:EU: C:2012:16; Case C-360/13, *Meltwater*, ECLI:EU:C:2014:1195, 23

[83] Case C-403/08, *Premier League*, 163, ECLI:EU:C:2011:631. Taina Pihlajarinne, "Copyright exceptions and limitations – is the principle of narrow interpretation gradually fading away?" *NIR – Nordiskt Immateriellt Rättsskydd* 89, no. 1 (2020): 117, 121; P. Bernt Hugenholtz, "Flexible Copyright: Can the EU Author's Rights Accommodate Fair Use?," in *Copyright Law in an Age of Limitations and Exceptions*, ed. Ruth L. Okediji (Cambridge University Press, 2017), 275, 286, has argued that there is room for a broader weighing of interests under Art. 5(5) of the Infosoc Directive.

[84] With reference to Art. 5(1) of the Infosoc Directive, *cf.* Case C-403/08, *Premier League*, 164 and 179, ECLI:EU: C:2011:631; Case C-360/13, *Meltwater*, ECLI:EU:C:2014:1195, 24; recital (31), Infosoc Directive.

[85] Case C-516/17, *Spiegel Online*, ECLI:EU:C:2019:625, 54 and 57.

[86] Case C-360/13, *Meltwater*, ECLI:EU:C:2014:1195, 57–59. See recital (3), DSM Directive, and discussion by Taina Pihlajarinne (2020) (n 83): 122; Geiger et al. (2018) p. 282 (n 24).

[87] The three-step test does not apply to *sui generis* database rights, as the wording of the relevant Art. 8(2) in the Database Directive differs from Art. 5(5) in the Infosoc Directive. Including the three-step test in the DSM Directive might be a step towards a better balancing of interests in the scope of the *sui generis* database right. In this direction, see Meys, 469 (n 23); DG CONNECT, "*Study in support of the evaluation of Directive 96/9/EC on the legal protection of databases*" (2018): 25, https://op.europa.eu/en/publication-detail/-/publication/5e9c7a51-597c-11e8-ab41-01aa75ed71a1.

[88] Art. 14 of Regulation (EU) 2021/695 of the European Parliament and of the Council of 28 April 2021, establishing Horizon Europe – the Framework Programme for Research and Innovation, laying down its rules for participation and dissemination.

[89] For example, the Norwegian Universities and Colleges Act § 2-1.

https://doi.org/10.1017/jli.2024.46

will also be held accountable for adhering to general principles on research integrity and research ethics.[90] Specialized integrity principles have been developed for the use and development of AI in the research process.[91]

This section looks at how language models developed by research institutions may be disseminated in line with the above-mentioned principles, while not extending beyond the scope of Article 3 of the DSM Directive. Activities that go beyond the scope of Article 3 must comply with Article 4 of the DSM Directive, most notably that right holders must have a real opportunity to reserve the use of their works in the development of the AI model.

An initial question is whether the model, once trained, could be made openly available for the public to use, including for commercial purposes. For example, a model trained on the works available in a national law library could be very attractive not only to other researchers but to courts, ministries, and commercial players, such as law firms and industry, along with the general public. Regarding the individual works on which the model has been trained, this hinges on whether the model itself contains reproductions of sufficient parts of works.[92] This comes down to the model's technical features and whether the file containing the experiences from the training process contains parts of the training materials.[93] If the model does not contain parts of the training materials amounting to a work's reproduction, making it openly accessible for public use, including for commercial purposes, should be lawful under Article 3 of the DSM Directive even without complying with the obligation to allow right holders to "opt out" per Article 4 of the DSM Directive. When the lawfulness of the model's further use hinges on whether a reproduction is made, the model may also be further trained and developed for commercial purposes if this process is compliant with Article 4. "Article 3-materials" may not be used for training for commercial purposes without the right holders having had the opportunity to "opt out."

For database rights, problems arise for those that are privately held, as public research institutions' obligations vis-à-vis open access and open data likely prevent them from invoking database rights against commercial developers of AI.[94]

For privately held database rights and press publishers' rights, infringement hinges upon whether commercial use of the model entails copying to such an extent that right holders' investments in the production of the content are undermined.[95] While the CJEU has been less concerned with the amount of copying when assessing infringement of database rights, there is still a basic requirement that the model's use also entails accessing the collections or that the model has stored sufficient information during training to amount to infringement under Article 7 of the Database Directive or Article 15 of the DSM Directive. This seems unlikely.

The press publisher's right is limited to two years, and the training of the model could therefore possibly be designed to avoid infringement. Database rights last for fifteen years from the database's completion, and since many press publishers are likely to have overlapping database rights in their collections,[96] adherence to Article 4 of the DSM Directive might still be necessary.

Problems related to the model's output text fall outside the scope of Articles 3 and 4 of the DSM Directive. For the press publisher's right, the EU Commission has explicitly stated that this right does not extend to facts.[97] Output from fact-finding services or information service LLMs could therefore present facts from press publications. LLMs may be able to produce text that infringes copyright by "memorization" or by circumventing paywalls. Such output is more natural to assess individually for copyright infringement. As AI models producing "creative" content become more sophisticated, it may be called into question whether right holders' interests are unreasonably prejudiced if the model produces text and creative content that might not infringe individual rights but erode right

[90] European Code of Conduct for Research Integrity (n 60).

[91] EU Commission (n 63), and EU Commission, High-Level Expert Group on Artificial Intelligence (n 65).

[92] An extract of as little as eleven consecutive words could infringe copyright, Case C-05/08, *Infopaq I*, 39, ECLI:EU:C:2009:465.

[93] *Cf.* secs. 3 a and b above. See also Thomas Margoni and Martin Kretschmer (2022): 693–94 (n 35).

[94] *Cf.* sec. 3 c, and Art. 1, nr. 6, Open Data Directive.

[95] Recital (58), DSM Directive, with regard to the press publisher's rights. For database rights, the investment must relate to the obtainment, verification, or presentation of the database contents, Art. 7(1), Database Directive.

[96] Lionel Bentley et al., "Strengthening the Position of Press Publishers and Authors and Performers in the Copyright Directive," study for DG IPOL, 23.

[97] Recital (57), DSM Directive.

https://doi.org/10.1017/jli.2024.46

holders' economic interests.[98] These questions point to the importance of research involving AI to adhere to ethical principles. Under the EU Digital Strategy, the development of trustworthy AI requires that the model is lawful, respecting all applicable laws and regulations, including in the development phase.[99] Responsible use of AI in research requires the respect of rights (that is, intellectual property rights) as well as accountability for the whole research process.[100]

FINDINGS

While standards for ethical research demand respect for intellectual property rights, the discussion in this article has shown that research activities involving the training of language models in compliance with copyright law present new challenges to research ethics and methodology.

First, ethically sound information management must consider technology's social effects, such as bias, diversity, non-discrimination, and fairness, thereby respecting fundamental rights.[101] Compiling a training corpus must be done with regard to these standards, an assessment that might be quite complicated.[102]

Second, there is still uncertainty as to whether the DSM Directive will efficiently guarantee access to materials protected by intellectual property rights for training language models in public interest research. Data transparency prerogatives in recent EU legislation[103] support better access to databases and collections of works. However, the DSM Directive does not foresee specific enforcement action against contractual or commercial practices that restrict access to protected materials for research purposes, such as prohibitively expensive licensing or differentiated subscriptions. While data transparency and accountability are values generally promoted in legislation, regulation is fragmented and serves different objectives, and legal uncertainty remains.

Finally, there is still legal uncertainty regarding the dissemination of results (that is, the use of a model for different societal purposes). The EU Commission did not consider the inclusion of the public interest mission of public research institutions and libraries—that is, making research output openly available in society when drafting the exceptions for TDM in the DSM Directive. The narrow focus of the exception on the training phase causes legal uncertainty regarding the dissemination phase.[104] The mere use of a developed model to produce new text rarely requires revisiting and the (digital) copying of the training materials. If someone wants to develop and train the model further, a more nuanced assessment is necessary. That a model can be used for further training by commercial undertakings is consistent with the public interest mission of public research institutions but falls outside the scope of the exception in Article 3 of the DSM Directive. The lawfulness of such training depends on whether new (digital) copies of "Article 3-training materials" are made, which again depends on the model's technical specifications. Retraining the model, including on the original "Article 3-training materials" would require that the conditions in Article 4 of the DSM Directive are met, notably that right holders have been given the opportunity to "opt out." It could be questioned whether pinning open science to the choice of machine-learning methods could potentially lead to an imbalance when weighing interests while choosing methods for designing and training AI models. If copyright law is given too much weight, methodologies may be chosen that have less consideration for other important values, such as bias, non-discrimination, and fairness.

[98] In such cases, it is possible that the application of the three-step test in Art. 5(5), Infosoc Directive, ref. Recital (6), the DSM Directive would provide a basis for finding the scope of Art. 3 of the DSM Directive is overreached.

[99] EU Commission (n 65).

[100] EU Commission (n 63); recital (107), AI Act.

[101] EU Commission 5 (n 63).

[102] *Cf.* report from the Norwegian National Library (2024) (n 5).

[103] Notably, in the Open Data Directive; see the section on "Database Rights" in this paper.

[104] More generally, see Sean M. Fiil-Flynn et al., "Legal reform to enhance global text and data mining research," *Science* 378, no. 6623 (2022): 951–53.

https://doi.org/10.1017/jli.2024.46