Martin Senftleben*

Copyright Data Improvement for AI Licensing – The Role of Content Moderation and Text and Data Mining Rules

1. Introduction

To enable European authors, performers and creative industries to benefit from licensing opportunities in the field of new technologies, such as AI training,[1] it is important to establish a comprehensive metadata infrastructure that ensures the visibility and accessibility of European work repertoires in digital and algorithmic environments, and facilitates rights clearance. Recognising the need for metadata improvement, various European initiatives aim to increase awareness among artists and rightholders, and build bridges between existing metadata collections and infrastructures.[2] One central factor in the equation, however, has remained underexplored to this day: copyright norms may serve as legal vehicles to encourage rightholders to constantly provide updated metadata in standardised form. By strategically using copyright provisions as statutory incentive schemes for data improvement, metadata creation and updating could become a task which rightholders perform routinely. Copyright

* Ph.D.; Professor of Intellectual Property Law and Director, Institute for Information Law (IViR), Amsterdam Law School, University of Amsterdam; Of Counsel, Bird & Bird, The Hague, The Netherlands.

[1] For a discussion of the need for rights clearance and remuneration systems for AI training, see G. Westkamp, 'Borrowed Plumes: Taking Artists' Interests Seriously in Artificial Intelligence Regulation', 1 (19-26), forthcoming; K. de la Durantaye, 'Nutzung urheberrechtlich geschützter Inhalte zum Training generativer künstlicher Intelligenz – ein Lagebericht', *Archiv für Presserecht* 55 (2024), 9 (21-22); M.R.F. Senftleben, 'AI Act and Author Remuneration – A Model for Other Regions?', 1 (6-23), available at: https://ssrn.com/abstract=4740268; C. Geiger, 'Elaborating a Human Rights Friendly Copyright Framework for Generative AI', *International Review for Intellectual Property and Competition Law* 2024, forthcoming, 1 (29-33), available at: https://ssrn.com/abstract=4634992; D. Friedmann, 'Creation and Generation Copyright Standards', *NYU Journal of Intellectual Property and Entertainment Law* 14 (2024), forthcoming, 1 (7-8); C. Geiger, 'When the Robots (Try to) Take Over: Of Artificial Intelligence, Authors, Creativity and Copyright Protection', in: F. Thouvenin/A. Peukert et al. (eds.), *Innovation – Creation – Markets, Festschrift für Reto M. Hilty*, Berlin: Springer 2024, 67 (67-87); M.R.F. Senftleben, 'Generative AI and Author Remuneration', *International Review of Intellectual Property and Competition Law* 54 (2023), 1535 (1542-1556); G. Frosio, 'Should We Ban Generative AI, Incentivise It or Make It a Medium for Inclusive Creativity?', in: E. Bonadio, C. Sganga (eds.), *A Research Agenda for EU Copyright Law*, Cheltenham: Edward Elgar 2024, 1 (19-21), available at: https://ssrn.com/abstract=4527461; C. Geiger/V. Iaia, 'The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative AI', *Computer Law and Security Review* 52 (2024), forthcoming, 1 (10-16), available at: https://ssrn.com/abstract=4594873.

[2] Cf. P. Rixhon/A. Strowel et al., *Study on Copyright and New Technologies: Copyright Data Management and Artificial Intelligence – SMART 2019/0038*, Brussels: Publications Office of the European Union 2022, 53-55 and Annex 5.3, available at: https://data.europa.eu/doi/10.2759/570559, 53-55 and Annex 5.3; M.R.F. Senftleben/T. Margoni et al., 'Ensuring the Visibility and Accessibility of European Creative Content on the World Market: The Need for Copyright Data Improvement in the Light of New Technologies and the Opportunity Arising from Article 17 of the CDSM Directive', *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* 13 (2022), 67 (74-77); N. Gronau/M. Schaefer, 'Why Metadata Matters for the Future of Copyright', *European Intellectual Property Review* 43 (2021), 488 (490-494).

rules that already require the transmission of work-related information could be transformed into data improvement instruments that contribute to the evolution of accurate, harmonised and interoperable metadata. For instance, it seems possible to transform the notification of work-related information under Article 17 of the 2019 Directive on Copyright in the Digital Single Market (CDSMD)[3] and the opt-out mechanism relating to text and data mining (TDM) under Article 4 of the same Directive into regulatory frameworks that generate a broader spectrum of descriptive and ownership data. If information stemming from these metadata engines is pooled in a central European copyright data repository, the accumulation of copyright data could lead to a metadata reservoir that is capable of enhancing licensing and remuneration opportunities in digital and algorithmic contexts.

The following discussion of this metadata mainstreaming strategy first describes problems arising from inadequate copyright metadata and past initiatives that sought to improve the copyright data infrastructure (section 2). It then turns to a new music licensing hub that has emerged in the US as a result of a legislative intervention and outlines new licensing opportunities that arise in the area of AI training (section 3). Exploring existing data-related rules in the EU copyright acquis against this background, it will become apparent that work notifications for content blocking purposes under Article 17 CDSMD and the reservation of copyright with regard to text and data mining under Article 4 CDSMD have a remarkable potential to foster copyright data improvement projects. However, it will be necessary to bundle data streams flowing from the application of these provisions in a central EU copyright data repository (section 4). The international prohibition of formalities in copyright law need not thwart metadata mainstreaming initiatives based on data-related provisions of EU copyright law (section 5). Hence, it is worthwhile to explore options for transforming copyright rules into legal tools to encourage metadata creation and improvement (concluding section 6).

2. Inadequate data infrastructure

The problem of insufficient metadata quality in the field of literary and artistic works is not new. More precisely, missing, inaccurate or non-interoperable metadata lead to a lack of adequate information on the work and its production (title, genre, content, year of creation, etc.), individual contributions enjoying protection, current rightholders etc. With the evolution of new licensing opportunities and rights clearance tasks in the area of AI training,[4] the issue of metadata quality is high on the agenda of both rightholders in the creative industries and AI entrepreneurs in the high tech sector. If copyright data for identifying training resources and concluding licensing deals are unavailable, unreliable or non-interoperable, the development of high quality AI products and services may be frustrated. Authors, performers and creative industry rightholders may miss important licensing opportunities.

[3] 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, *Official Journal of the European Communities* 2019 L 130, 92. For an overview of relevant copyright exceptions in EU copyright law, see M.R.F. Senftleben, *Study on EU Copyright and Related Rights and Access to and Reuse of Data*, European Commission, Directorate-General for Research and Innovation (DG RTD), Brussels: Publications Office of the European Union 2022, 27-28 (temporary copying) and 36-37 (provisions for (scientific) TDM), available at: https://data.europa.eu/doi/10.2777/78973.

[4] Cf. the literature references supra note 1. As to issues arising from the practical implementation of Article 4 CDSMD, see P. Mezei, 'A Saviour or a Dead End? Reservation of Rights in the Age of Generative AI', 1 (12-13), available at: https://ssrn.com/abstract=4695119.

Initiatives to improve copyright metadata have a long tradition in Europe. After several unsuccessful attempts to establish a comprehensive data infrastructure,[5] the strategic use of data-related copyright provisions could lead to metadata "mainstreaming": it could make metadata creation and updating a routine task of rightholders and offer important new impulses for enhancing data quality and licensing opportunities. The advantages of a comprehensive data infrastructure are evident: it could make the complete work repertoire visible and accessible for rights clearance purposes. Even smaller (country) repertoires, which might otherwise remain unconsidered due to access and language barriers, could be brought to light. Increased visibility of works and work catalogues, in turn, can lead to broader licensing opportunities. If a copyright data infrastructure also serves as a one-stop shop for acquiring rights and paying royalties, it can boost the exploitation of protected material in unexpected ways and bring increased revenue to authors, performers and the creative industry.[6]

In the music segment of the creative industries, there are several well-known examples of existing data infrastructures, such as the Common Information System (CIS) of the International Confederation of Societies of Authors and Composers (CISAC). With its various nodes in several regions of the world, the CIS-Net system and its associated standards represent a global tool to facilitate music licensing and revenue distribution.[7] In terms of music data standardisation, the music publishing industry's *International Standard Work Code* (ISWC),[8] the recording industry's *International Standard Recording Code* (ISRC), the *Interested Party Information Number* (IPI) and the *International Standard Name Identifier* (ISNI) continue to be prime examples of existing initiatives aimed at enabling the exchange of accurate data to identify repertoire or reduce transaction costs associated with the processing of licensing agreements.

At the same time, these examples highlight data deficiencies and interoperability problems resulting from different metadata sets and different approaches to identifying and verifying data. To date, initiatives to harmonise ISWC and ISRC metadata and integrate them into an overarching, comprehensive database have failed. In the EU, former Commissioner Neelie Kroes set up a working group in 2008 to explore the possibilities of establishing a global repertoire database (GRD). The participants of the working group, which included producers, collecting societies and distribution platforms, did come up with recommendations on how to proceed.[9] Ultimately, however, the project was buried in 2014.[10] Other unsuccessful attempts were the *International Music Joint Venture* from 2000, which was founded by several collecting

[5] Cf. Senftleben/Margoni et al., supra note 2, 74-77.
[6] Senftleben/Margoni et al., supra note 2, 70-74.
[7] See https://www.cisac.org/What-We-Do/Information-Services/CIS-Net.
[8] The ISWC was developed by CISAC in collaboration with the International Organisation for Standardisation (ISO) as 'a unique, permanent, and internally recognised reference number for the identification of musical works'. Another example of an identification system is the GRiD (Global Release Identifier) developed by IFPI. See https://www.ifpi.org/resource/grid/.
[9] Cf. M. Isherwood, 'Global Repertoire Database', presented at: World Intellectual Property Organization, *Enabling Creativity in the Digital Environment: Copyright Documentation and Infrastructure*, WIPO Meeting wipo_cr_doc_ge_11, 13-14 October 2011, Geneva: WIPO 2011, available at: https://www.wipo.int/meetings/en/2011/wipo_cr_doc_ge_11/prov_program.html.
[10] See P. Resnikoff, 'Global Repertoire Database Declared a Global Failure', Digital Music News, 10 July 2014, available at: https://www.digitalmusicnews.com/2014/07/10/global-repertoire-database-declared-global-failure/; S.F. Schwemer, *Licensing and Access to Content in the European Union. In Licensing and Access to Content in the European Union: Regulation between Copyright and Competition Law*, Cambridge: Cambridge University Press 2019, 68-73.

societies in Europe and North America, and a project initiated by the World Intellectual Property Organisation (WIPO) in 2011, which aimed to establish a joint rights database and also did not produce any concrete results.[11] Interestingly, WIPO has continued to explore options for data improvement, in particular with regard to the digital revolution. In 2020, the WIPO PROOF service was launched – adding date- and time-stamped digital fingerprints to files to provide proof of its existence at a specific point in time.[12] However, the service was discontinued on 1 February 2022.[13]

This series of unsuccessful attempts already shows that – despite existing metadata infrastructures such as the CIS-Net system and the ISWC/ISRC standards – there is a need in the European music industry to combine work- and rights-based databases to a greater extent in order to ultimately create overarching licensing platforms.[14] Recent initiatives clearly point in this direction. The *Technical Online Working Group Europe* (TOWGE), for example, brought together a large group of European collecting societies, music publishers and rights agencies to develop a digital royalty processing system.[15] An initiative with similar goals was taken by the Finnish collecting society Teosto. The collaboration between Teosto and the start-up *Mind Your Rights* resulted in the platform "Concertify" aiming at providing an efficient and transparent system for cross-border copyright licensing in addition to existing industry infrastructures. Concertify enables artists, rightholders, including collecting societies, music publishers and event organisers, to collaborate and transmit information directly using specific modules, such as a module for the transmission of setlists.[16] With the support of the Slovak Arts Council, a collaboration between the collecting society SOZA and various stakeholders from the music industry led to the creation of a prototype for a comprehensive database, including metadata, of the Slovak music repertoire. This initiative led to the development of a "Listen Local" recommendation system, which meets the requirements for trustworthy AI that have been formulated by the European *High-Level Expert Group on Artificial Intelligence*.[17] The accompanying feasibility study showed and quantified the problems arising from incomplete copyright data in existing databases and commercial AI solutions. It estimated that at least 15% of Slovak, Estonian, Hungarian and Dutch works were difficult to exploit due to data problems.[18]

An initiative taken by standard-setting organisation Digital Data Exchange LLC offers a further example of recent initiatives in the area of copyright data improvement: the DDEX project seeking to develop standards relating to metadata creation and management in the overall digital music supply chain. One of the DDEX standards, DDEX RDR, describes in detail the required metadata for managing international performance rights. The data exchange engine RDx (Repertoire Data Exchange) serves as a tool to implement the DDEX RDR standard in practice. It became fully operational in 2020. RDx is a data exchange hub affording participating entities

[11] Schwemer, id., 69-70.
[12] See https://www.wipo.int/wipoproof/en.
[13] See https://www.wipo.int/wipoproof/en/news/2021/news_0003.html.
[14] See Gronau/Schaefer, supra note 2, 488-494. See also F. Lyons/H. Sun et al., *Music 2025 – The Music Data Dilemma: Issues Facing the Music Industry in Improving Data Management*, Newport: UK Intellectual Property Office 2019, 34, available at: https://www.gov.uk/government/publications/music-2025-the-music-data-dilemma.
[15] See https://www.digitalmusicnews.com/2019/07/26/towge-digital-royalty-group/.
[16] See https://www.mindyourrights.fi/.
[17] See https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai and https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
[18] D. Antal, *Feasibility Study On Promoting Slovak Music in Slovakia and Abroad*, The Hague: Reprex 2020, available at: https://reprex.nl/publication/listen_local_2020/.

the opportunity to send and receive DDEX RDR data.[19] The copyright platform "Cube" offers a further example of an initiative seeking to provide an automated copyright data exchange system. The platform was commissioned by ICE Services, a joint venture of German, Swedish and UK collecting societies in the field of performing rights. Employing cloud computing and machine learning technologies, Cube seeks to increase the speed and accuracy of data consolidation for multi-territorial copyright administration.[20]

Other sectors of the creative industries face similar data issues and have also taken initiatives to improve, harmonise and merge data. In the field of book publishing, industry initiatives such as the establishment of various e-book platforms and catalogues play an important role. Another example is the *Entertainment Identifier Registry* (EIDR): a universal labelling system for film and television data based on DOI technology.[21] Examples of data standardisation initiatives include the International Standard Book Number (ISBN), the International Standard Serial Number (ISSN) for periodicals, the International Standard Music Number (ISMN) for notated music and the *International Standard* Audiovisual *Number* (ISAN) for audiovisual works. Furthermore, in the field of books, e-books and serial volumes, the standardisation work of the international EDItEUR group is important.[22] This initiative has led to the ONIX family of standards.[23] With regard to the digital environment, the International DOI Foundation offers the *Digital Object Identifier* (DOI) services and related registration facilities already mentioned: a technical and social infrastructure for the registration and use of persistent interoperable identifiers for use on digital networks, including identifiers for literary and artistic works.[24]

In the field of visual arts, the *Visual Arts Council of* CISAC has expanded its original work on resale rights and established an online licensing platform under the umbrella of the *International Council of Creators of Graphic, Plastic and Photographic Arts* (CIAGP).[25] OnLineArt (OLA) is a one-stop shop for acquiring licences for the online use of works of visual art worldwide, which currently includes works by 60,000 artists.[26] While existing initiatives in the field of visual arts – particularly digitisation initiatives by museums and other cultural heritage institutions – have significantly expanded data collections of visual art works, the situation in the field of photography and illustration is far less transparent.[27] Commercial owners of large fine art libraries, such as Getty Images, have managed to develop data management tools.[28] However, the cost of documenting a large number of individual works can quickly become prohibitive for smaller providers of photographic works and illustrations, given the low average value of individual works in the overall collection.[29] With regard to photo

[19] See https://www.ifpi.org/rdx-recording-industrys-new-data-exchange-service-now-fully-operational/ and https://www.rdx-portal.org/.
[20] See https://www.iceservices.com/innovation/cube/ and https://www.iceservices.com/about/.
[21] See https://www.eidr.org/.
[22] See https://www.editeur.org/2/About/#Intro.
[23] See https://www.editeur.org/8/ONIX/.
[24] See https://www.doi.org/.
[25] See http://www.ciagp.org/ and https://www.cisac.org/What-We-Do/Creators-Relations/CIAGP.
[26] See https://onlineart.info/.
[27] For a more detailed analysis of the specific situation and dynamics in the field of visual arts, see the study by T. Azzi/Y. El Hage, *Les métadonnées liées aux images fixes*, Paris: CSPLA 2021.
[28] As to the watermarking system of Getty Images and the position of Getty Images in the debate on AI training, see High Court of England and Wales, 1 December 2023, case IL-2023-000007, Getty Images (US) Inc, Getty Images International UC et al./Stability AI Ltd, [2023] EWHC 3090 (Ch), para. 3-8.
[29] On this investment dilemma, see R.A. Posner, 'Transaction Costs and Antitrust Concerns in the Licensing of Intellectual Property', *John Marshall Review of Intellectual Property Law* 4 (2005), 325.

metadata, the IPTC Photo Metadata Standard provides a widely used and recognised tool for documentation.[30] Nonetheless, challenges for reliably identifying authentic media assets remain in the area of photography.[31] The vast amount of works – and potentially limited commercial exploitation options – pose persisting problems. Compared to the status quo reached in the music sector, the process of creating, harmonising, linking and bundling work- and rights-related data in the visual arts sector still seems to be in its infancy.

Considering the conversion of work representations in the digital environment, it is important to highlight cross-sectoral initiatives, such as the *International Standard Content Code* (ISCC) which the *International Organization for Standardization* (ISO) is currently developing.[32] This initiative aims to provide a universal identification system for digital assets, based on encodings of text, images, audio, video and other content across media sectors and categories. The ISCC is intended to provide similarity-preserving fingerprints designed to identify digital content in decentralised and networked environments across creative industry branches.[33]

3. New licensing opportunities

This brief – by no means exhaustive[34] – outline of data initiatives clearly shows that the problem of inadequate data quality in the creative industries remains complex and unresolved. At the same time, there is an urgent need for improvement. Current developments in the increasingly digital and algorithmic information society make it more important than ever to take a fresh look at the problem and develop stronger regulatory support for copyright data improvement initiatives.

First, there is the issue of worldwide competition for data hegemony. While European initiatives, as described, have not yet led to an overarching, comprehensive data infrastructure for the individual branches of the creative sector, the creation of a comprehensive database has succeeded in the US, at least in the area of music. The US initiative goes back to the *Music Modernisation Act* (MMA) passed in 2018.[35] Title I of the MMA establishes the *Mechanical Licensing Collective* (MLC) as a one-stop shop for music licensing. For the proper functioning of this new licensing body, the creation of a comprehensive database of music rights was indispensable. The MLC ultimately achieved this goal on the basis of close cooperation with major providers of music streaming services, in particular Apple and Spotify.[36] The new licensing hub managed to provide a US-wide platform for royalty administration, enforcement and processing as of 1 January 2021.[37] With regard to the strategic use of legislation to improve

[30] See https://ieeexplore.ieee.org/abstract/document/10353009 and https://iptc.org/standards/photo-metadata/quick-guide-to-iptc-photo-metadata-and-google-images/.
[31] See https://ieeexplore.ieee.org/abstract/document/10353009 and https://link.springer.com/chapter/10.1007/978-3-031-27818-1_14.
[32] See https://www.iso.org/standard/77899.html.
[33] See https://core.iscc.codes/.
[34] Rixhon/Strowel et al., supra note 2, Annex 5.3.
[35] See House Report 1551, Public Law 115-264, dated 11 October 2018.
[36] See https://www.appleworld.today/blog/2019/11/18/apple-spotify-to-fund-new-music-royalties-collective.
[37] See https://www.themlc.com/press/mechanical-licensing-collective-begins-full-operations-envisioned-music-modernization-act. With respect to the underlying preparatory work, see further U.S. Copyright Office Library of Congress, *MLC Comments in Reply to the Designation Proposal of the American Music Licensing Collective, Inc*, Docket No. 2018-11, 21, available at: https://bw-98d8a23fd60826a2a474c5b4f5811707-bwcore.s3.amazonaws.com/photos/Proposed_MLC_-_Reply_Comments.pdf.

the copyright data infrastructure, it is of particular interest that this development in the US can be traced back to a legislative intervention, namely the adoption of the MMA in 2018.

The creation of this licensing infrastructure on the other side of the Atlantic raises the question whether fresh European initiatives could also pave the way for large-scale bundling of copyright metadata. Otherwise, powerful structures, such as the MLC system built on Apple and Spotify data, may become de facto standards and expand their sphere of influence to the European continent.[38] A European initiative could aim at including the full spectrum of cultural diversity in Europe. With appropriate search and recommendation tools, it could make all repertoires – big and small – visible and accessible.

Second, technical developments and related licensing opportunities increase the need to create an overarching data infrastructure. The aforementioned training of generative AI systems, for example, is largely reliant on extensive use of human source material that allows the systems to analyse the parameters of literary and artistic works. Without machine-readable literary and artistic input from flesh-and-blood authors, an AI system has no template for algorithmic processes capable of emulating human creativity. The machine is only capable of mimicking human literary and artistic works after it had the opportunity to deduce patterns from myriad human creations that served as resources for training purposes. Considering the insatiable appetite of AI systems for training material, it is foreseeable that promising mass licensing opportunities will arise in this area. To create licensing opportunities not only for big repertoire holders[39] but also for smaller (country) repertoires, however, it is imperative to enhance the visibility of available work repertoires and reduce search and transaction costs by making work-related metadata – describing content and providing rights clearance information – available in a harmonised and interoperable format.

Modern data-driven AI often uses text and data mining (TDM) techniques to obtain the data needed for machine learning.[40] TDM has emerged as one of the most powerful digital tools in the AI environment for extracting patterns, correlations and hidden knowledge from existing content and data.[41] Techniques currently discussed under the terms machine learning, natural language processing and deep neural networks require the training of AI systems on vast amounts of content and data. Required training information is often extracted by automated machine reading techniques from books, journal articles, musical works or films that enjoy copyright protection.

---

[38] It is important to note that, in the process, the MLC database may be enriched with more and more European copyright data. The MLC system provides for a so-called "matching tool" that "allows Members to match sound recordings to compositions/musical works within their catalog. This ensures you'll collect the royalties you're owed for your work(s)." See https://help.themlc.com/en/support/what-is-the-matching-tool. In practice, this indicates that, to the extent to which European rightholders are or become MLC members, they will be able to add their copyright data to the overarching MLC database.

[39] For an example of an existing licensing success at big repertoire level, see the agreement concluded between Universal Music and Google/YouTube, as described by A. Nicolaou/M. Murgia, 'Google and Universal Music negotiate deal over AI "deepfakes"', *Financial Times*, 8 August 2023, available at: https://www.ft.com/content/6f022306-2f83-4da7-8066-51386e8fe63b.

[40] For a discussion of the question whether the TDM provisions in Articles 3 and 4 CDSMD can be understood to cover the training of generative AI systems, see Senftleben, 'AI Act and Author Remuneration', supra note 1, 9-10.

[41] See T. Margoni, 'Computational Legal Methods: Text and Data Mining in Intellectual Property Research', in: I. Calboli/M.L. Montagnani (eds.), *Handbook on Intellectual Property Research*, Oxford: Oxford University Press 2021, 487 (487-505).

With the adoption of the AI Act (AIA),[42] the EU has substantially enhanced the rules governing the training of AI systems and, more specifically, the interface with copyright protection. The AI Act clarifies that, from an EU perspective, reproductions carried out for AI training purposes have copyright relevance and require the authorization of rightholders unless a copyright exception, such as the specific TDM rule for scientific research in Article 3 CDSMD, exempts the AI training activity from the control of rightholders. The AI Act also confirms the rights reservation system following from Article 4(3) CDSMD with regard to forms of TDM falling outside the scope of the scientific TDM exemption and going beyond mere temporary copying: declaring an "opt-out" in an appropriate – machine-readable – manner, copyright owners seeking to prevent the use of their works for AI training purposes can reserve their rights. To enable rightholders to police AI training processes, the AI Act introduces a new transparency obligation. Developers of generative AI systems must submit sufficiently detailed information on work repertoires that have been used for training purposes.[43]

In the light of these legislative developments, the insatiable appetite of generative AI systems for literary and artistic data input and the ability of these systems to mimic human literary and artistic productions are more than an unprecedented threat to human creativity.[44] With the rights reservation option in Article 4(3) CDSMD and the transparency obligations following from the AI Act, the EU legislator has provided a legal framework that may lead to new licensing opportunities.[45] If rightholders refrain from categorically prohibiting the use of their works for AI training and, instead, enter into licensing agreements,[46] AI training may offer a promising new source of revenue for the creative industries.[47] However, if the creative industries in Europe fail to provide licences for large-scale AI applications at low transaction costs and at the required large scale, the risk arises that AI training projects will take place in other regions, such as the US. If the necessary content and data can be purchased centrally outside the EU on a larger scale and with less administrative effort, the European creative industries will most probably lose attractive licensing revenues from AI training. The creation of an overarching data infrastructure is therefore desirable and urgent in the light of new technical developments.

4. Copyright mechanisms for metadata improvement

Considering this urgent need for copyright data improvement, it is important to identify starting points for regulatory support – in the sense of data-related copyright provisions that could contribute to the creation and maintenance of a comprehensive copyright data infrastructure – in the EU copyright acquis. While an examination of all potential regulatory avenues is beyond

[42] Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, *Official Journal of the European Union* 2024 L, forthcoming.
[43] Recital 107 and Article 53(1)(d) AIA.
[44] Cf. Senftleben, 'Generative AI and Author Remuneration', supra note 1, 1538-1542.
[45] Cf. P. Keller, 'Protecting Creatives or Impeding Progress? Machine Learning and the EU Copyright Framework', *Kluwer Copyright Blog*, 20 February 2023, available at: https://copyrightblog.kluweriplaw.com/2023/02/20/protecting-creatives-or-impeding-progress-machine-learning-and-the-eu-copyright-framework/.
[46] Cf. Mezei, supra note 4, 10-12; Senftleben, 'Generative AI and Author Remuneration', supra note 1, 1546-1549.
[47] Senftleben/Margoni et al., supra note 2, 71-74.

the scope of the current inquiry,[48] an analysis of work notifications for content blocking purposes and a closer examination of the aforementioned TDM provisions reveals the potential of existing rules to foster copyright data improvement projects. As the following analysis will show, an amalgam of metadata mainstreaming in the context of Article 17 CDSMD (section 4.1) and Article 4 CDSMD (section 4.2) may substantially enhance copyright data quality. To unfold these beneficial effects, however, it would be necessary to bundle data streams flowing from the application of these provisions and bring them together in a central EU copyright data repository (section 4.3).

4.1 Work notifications for content blocking

Arguably, the work notification mechanism in Article 17(4)(b) CDSMD offers a promising opportunity for data improvement, especially with regard to categories of literary and artistic works that regularly play an important role on user-generated content platforms. Music, film, photography and other forms of visual art seem particularly relevant in this context.[49]

Article 17(4)(b) requires online content-sharing service providers (OCSSPs)[50] to use their best endeavours to ensure the unavailability of works and other protected subject matter for which rightholders have provided them with "relevant and necessary" information. Thus, it constitutes a legal provision that initiates a flow of data from rightholders to online platforms. The notification of works opens up the possibility of ensuring the application of measures to block and remove infringing content.[51] In this context, it can be assumed that "relevant and necessary" information in the sense of Article 17(4)(b) goes beyond mere work-related data. A copyright owner submitting information must inform the online platform of his identity, address and other contact details, as well as the nature and (territorial) scope of the rights asserted. According to Article 17(8) CDSMD, OCSSPs must provide rightholders, at their request, with appropriate information on the operation of their procedures for cooperation under Article 17(4) CDSMD. Without contact information, this reporting obligation cannot be fulfilled. When it comes to complaint and redress procedures under Article 17(9) CDSMD, rightholders must also "duly justify" the reasons for their request to block content. Obviously, the exchange of information between rightholders and online platform providers is thus not only intended to ensure up-to-date information on works enjoying copyright protection, but also to bring about an accurate and constantly updated collection of rightholder data, including contact information. Otherwise,

[48] For a broader overview of potential regulatory reference points in EU copyright and digital legislation, see M.R.F. Senftleben/T. Margoni et al., 'Music Metadata Mainstreaming and EU Law', OpenMusE Working Paper D5.6, forthcoming, available at: https://www.openmuse.eu/.

[49] See M.R.F. Senftleben, 'User-Generated Content – Towards a New Use Privilege in EU Copyright Law', in: T. Aplin (ed.), *Research Handbook on IP and Digital Technologies*, Cheltenham: Edward Elgar 2020, 136-162; J.-P. Triaille, S. Dusollier et al., *Study on the Application of Directive 2001/29/EC on Copyright and Related Rights in the Information Society*, Study prepared by De Wolf & Partners in collaboration with the Centre de Recherche Information, Droit et Société (CRIDS), University of Namur, on behalf of the European Commission (DG Markt), Brussels: European Union 2013, 457-510; S.D. Jamar, 'Crafting Copyright Law to Encourage and Protect User-Generated Content in the Internet Social Networking Context', *Widener Law Journal* 19 (2010), 843; N. Helberger/L. Guibault et al., *Legal Aspects of User Created Content*, Amsterdam: Institute for Information Law 2009; M.W.S. Wong, 'Transformative User-Generated Content in Copyright Law: Infringing Derivative Works or Fair Use?', *Vanderbilt Journal of Entertainment and Technology Law* 11 (2009), 1075; OECD, 12 April 2007, *Participative Web: User-Created Content*, Doc. DSTI/ICCP/IE(2006)7/Final, available at: https://web-archive.oecd.org/2012-06-15/135484-38393115.pdf.

[50] See the definition of relevant service providers in Article 2(6) CDSMD.

[51] Cf. CJEU, 26 April 2022, case C-401/19, Poland/Parliament and Council, para. 89.

OCSSPs will hardly be able to report on their practice of blocking content and request rightholders to substantiate blocking requests in complaint procedures.

However, before painting an overly positive picture of Article 17(4)(b) as a cure for all kinds of music metadata problems, it is important to point out that the provision is only one building block in a more complex puzzle. The regulatory framework of Article 17 CDSMD – as can be seen from the first two paragraphs of the provision – focuses on the right of communication to the public and making available to the public. Accordingly, the notification mechanism resulting from Article 17(4)(b) also concerns these exclusive rights. The right of communication to the public and the right of making available to the public may be central to OCSSPs and various other forms of digital services. However, the described developments in the area of AI training that offer promising licensing opportunities for literary and artistic works predominantly affect the reproduction right. Under the outlined EU copyright and AI legislation, AI training requires the acquisition of use permissions in the field of reproduction. As the provisions on text and data mining in Articles 3 and 4 CDSMD show, the reproduction right is central in this context.[52]

The question therefore arises whether metadata stemming from Article 17(4)(b) notifications can provide useful information for initiatives aimed at identifying works and clarifying rights in new technology areas, such as the AI sector. The answer to this question depends on the phrase "relevant and necessary" information in Article 17(4)(b). In order to ensure the unavailability of protected works on online platforms, it seems sufficient to know who is entitled to prohibit the sharing of user-generated content. In other words: information about the holders of rights of communication to the public and/or making available is decisive. However, this fact does not preclude the further enrichment of data transfers. As already pointed out, it is the overarching objective of metadata improvement to increase the visibility and accessibility of protected works and create new licensing opportunities. Rightholders who support these objectives – and are, for instance, interested in licensing for AI training purposes – may therefore be willing to go beyond the information necessary for Article 17(4)(b) content blocking and enrich work notifications with additional information covering a broader range of exclusive rights. This wider provision of metadata may include, for example, the reproduction right. Article 17(4)(b) CDSMD may thus have the effect of jump-starting a broader process of aggregating copyright metadata. This broader process may include additional exclusive rights, such as the right of reproduction.

It should also be taken into account that rightholders provide work-related information under Article 17(4)(b) in order to prevent unauthorised user uploads to online platforms. The data provided serve to identify the work and infringing copies. Given this objective, Article 17(4)(b) notifications may not reveal the nature and content of the work itself. A potential user searching for a specific type of work, such as an AI developer searching for a specific category of human expression, may therefore find information derived from Article 17(4)(b) notifications insufficient.

Again, it is important to bear in mind that, in the framework of copyright data improvement initiatives, reliance on Article 17(4)(b) notifications would be an element of a broader strategy to employ copyright norms as vehicles to enhance the visibility and accessibility of the European repertoire for licensing in digital and algorithmic contexts. These benefits can be a

[52] Cf. Senftleben, supra note 3, 36-37.

strong incentive for rightholders to go beyond minimal work identification data and provide additional descriptive metadata reflecting the genre, nature and content of the work. The stakeholder dialogue which the Commission will initiate on the basis of Article 17(10) CDSMD could also address the issue of copyright data. A discussion of "best practices for cooperation" between copyright holders and online platforms could devote attention to metadata improvement and lead to the establishment of appropriate work notification protocols: protocols requiring enriched metadata that go beyond the information that is strictly necessary for content blocking under Article 17(4)(b).

4.2 Rights reservation with regard to text and data mining

Importantly, the content blocking mechanism in Article 17(4)(b) CDSMD is not the only legal tool that can be employed to support initiatives aiming at copyright data improvement. The TDM rule in Article 4 CDSMD – and in particular the aforementioned opt-out mechanism enshrined in the third paragraph of this provision – offers a further opportunity to provide regulatory support. Considering the aforementioned focus of Article 17(4)(b) CDSMD on the right of communication to the public and the right of making available to the public, Article 4(3) CDSMD seems a particularly important counterpart. As already indicated, the harmonised TDM provisions, including Article 4 CDSMD, concern the right of reproduction.

Complementing the exemption of scientific TDM in Article 3 CDSMD, Article 4(1) CDSMD contains a more general exemption. Under this additional provision, anyone may make copies of works or databases for the purposes of TDM and retain them as long as necessary for the TDM process.[53] With regard to this broader category of TDM outside the scope of the scientific research rule in Article 3 CDSMD, Article 4(3) CDSMD adds an important nuance by stipulating that rightholders can reserve their rights. The provision contains the following opt-out mechanism:

> The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.[54]

Before examining how this opt-out mechanism could contribute to data improvement initiatives, it is important to clarify the scope of the provision, in particular with regard to the use of literary and artistic works for the training of generative AI systems. As the CDSM Directive dates back to 2019, it may be argued that rights reservations under Article 4(3) CDSMD do not cover these AI training activities. Arguably, the EU legislator did not have in mind the use of copyrighted material as mere data input for the training of generative AI systems which did not exist in 2019.[55] In the TDM debate, it has been underlined around the globe that TDM copies have a specific nature. They fall outside the concept of reproduction in the

[53] Article 4(1) and (2) CDSMD. As to the relevance of Article 4 CDSMD to generative AI systems, see J.P. Quintais, 'Generative AI, Copyright and the AI Act', *Kluwer Copyright Blog*, 9 May 2023, available at: https://copyrightblog.kluweriplaw.com/2023/05/09/generative-ai-copyright-and-the-ai-act/.

[54] Article 4(3) CDSMD.

[55] For a discussion of this argument with regard to the right of reproduction in international copyright law, see M.R.F. Senftleben, 'Compliance of National TDM Rules With International Copyright Law – An Overrated Nonissue?', *International Review of Intellectual Property and Competition Law* 53 (2022), 1477 (1493-1502).

traditional sense of making copies for the purpose of consulting and enjoying a work.[56] From a US perspective, Michael Carroll has pointed out that in the context of TDM:

> copies are made only for computational research and the durable outputs of any text and data mining analysis would be factual data and would not contain enough of the original expression in the analysed articles to be copies that count.[57]

Explaining the outright exemption of TDM activities in Article 30-4(ii) of the Japanese Copyright Act, Tatsuhiro Ueno has pointed out that:

> if an exploitation of a work is aimed at neither enjoying it nor causing another person to enjoy it (e.g. text-and-data mining, reverse engineering), there is no need to guarantee the opportunity of an author or copyright holder to receive compensation and thus copyright does not need to cover such exploitation. In other words, exploitation of this kind does not prejudice the copyright holder's interests protected by a copyright law.[58]

Criticizing the regulation of TDM in the EU, Rosanna Ducato and Alain Strowel described the following alternative approach:

> when acts of reproduction are carried out for the purpose of search and TDM, the work, although it might be reproduced in part, is not used as a work: the work only serves as a tool or data for deriving other relevant information. The expressive features of the work are not used, and there is no public to enjoy the work, as the work is only an input in a process for searching a corpus and identifying occurrences and possible trends or patterns.[59]

In fact, the distinction between use of "works as works" and use "as data" is not entirely new in the European copyright debate. In 2011, Mauricio Borghi and Stavroula Karapapa already developed the concept of "de-intellectualized use"[60] against the background of mass digitization projects, such as the Google Book Search. As Borghi and Karapapa point out, mass digitisation turns protected content into mere data – with the result that "the expression of the idea embodied in the work is not primarily used to communicate the 'speech' of the author to the public but rather to form the basis of machine-workable algorithms."[61]

In the light of these comments, one may assume that it is an open question whether the use of copyrighted works during AI training falls within the scope of copyright. As use in this specific setting does not constitute use of an author's individual expression for communication purposes, copyright may be inapplicable from the outset and copyright data improvement may be beyond reach. Luckily, the AI Act – adopted after the generative AI revolution and containing provisions on generative AI[62] – clarifies the matter. Recital 105 AIA confirms that the use of literary and artistic works for AI training purposes has copyright relevance and

[56] Cf. Senftleben, id., 1495-1502.
[57] M.W. Carroll, 'Copyright and the Progress of Science: Why Text and Data Mining is Lawful', *U.C. Davis Law Review* 53 (2019), 893 (954).
[58] T. Ueno, 'The Flexible Copyright Exception for "Non-enjoyment" Purposes Recent Amendment in Japan and Its Implication', *Gewerblicher Rechtsschutz und Urheberrecht International* 70 (2021), 145 (150-151).
[59] R. Ducato/A. Strowel, "Ensuring Text and Data Mining: Remaining Issues with the EU Copyright Exceptions and Possible Ways Out", *European Intellectual Property Review* 43 (2021), 322 (334).
[60] M. Borghi/S. Karapapa, "Non-display Uses of Copyright Works: Google Books and Beyond", *Queen Mary Journal of Intellectual Property* 1 (2011), 21 (45).
[61] Borghi/Karapapa, id., 44-45.
[62] For an overview, see Senftleben, 'AI Act and Author Remuneration', supra note 1, 7-14.

involves acts of text and data mining that require the authorisation of rightholders: "[a]ny use of copyright protected content requires the authorisation of the rightholder concerned unless relevant copyright exceptions and limitations apply."[63]

In line with this clarification in Recital 60i AIA, it can be assumed that EU copyright law brings all forms of TDM, including TDM for generative AI training purposes, under the umbrella of the right of reproduction and, accordingly, offers licensing opportunities and incentives for copyright data improvement, as explained above. More concretely, this configuration of the right of reproduction means that EU copyright law brings commercial AI training falling under Article 4(1) CDSMD within the reach of rightholders seeking to receive a remuneration for the use of their works.[64] Referring to the opt-out mechanism in Article 4(3) CDSMD, the AI Act confirms the intention to give rightholders the opportunity to exercise control over the use of their works for AI training purposes in Article 4 CDSMD scenarios:

> Where the rights to opt out has been expressly reserved in an appropriate manner, providers of general-purpose AI models need to obtain an authorisation from rightholders if they want to carry out text and data mining over such works.[65]

As in other cases where copyright holders can refuse the permission for a given form of use, this veto right can pave the way for licensing negotiations.[66] It is conceivable that the rights reservation option in Article 4(3) CDSMD leads to the evolution of machine-readable rights reservation protocols that express different rightholder standpoints. One standpoint could be a machine-readable rights reservation that signals an outright exclusion of any use of the literary and artistic work at issue for AI training purposes. Using this rights reservation option, rightholders can express their preference for an outright prohibition and prevent TDM of their literary and artistic repertoire altogether. An alternative standpoint, however, could be a machine-readable rights reservation that prohibits use for AI training purposes only if the AI trainer behind the crawler is reluctant to pay remuneration. Using this alternative version, rightholders can express their willingness to permit the use against the payment of remuneration. In other words: the rights reservation option in Article 4(3) CDSMD can lead to generally agreed, machine-readable licensing protocols that trigger an automated process for the payment of remuneration.

For this overarching rights clearance mechanism to take shape, however, it is indispensable to enrich the opt-out declaration with relevant metadata. At first glance, Article 4(3) CDSMD does not seem to have much potential in this regard. A mere opt-out statement, for instance in the form of robots.txt embedded in a website, need not offer detailed information on protected works. The mere indication that the website content is not available for TDM seems sufficient to ensure that the AI crawler looks for training material elsewhere.

The moment a rightholder seeks to use the opt-out mechanism as a vehicle to conclude licensing deals, however, the situation becomes markedly different. As a minimum, the rightholder will have to enrich the machine-readable opt out with basic ownership and contact information. Who

[63] Recital 105 AIA.
[64] Cf. Keller, supra note 45.
[65] Recital 60i AIA.
[66] Cf. the positive assessment of the situation by Keller, supra note 45; Communia, "Using Copyrighted Works for Teaching the Machine", *Communia Policy Paper* 15, 26 April 2023, available at: https://communia-association.org/policy-paper/policy-paper-15-on-using-copyrighted-works-for-teaching-the-machine/.

is the rightholder? What is the (territorial) scope of the rights? How can the rightholder be contacted? The moment the opt out is seen as a vehicle to attract the attention of AI developers and facilitate licensing agreements, it also seems advisable to provide descriptive metadata. Which repertoire is available? Which genre, style, period of creation?

Finally, and perhaps most importantly, it must be considered that the website with the opt-out statement will not always contain the works which the rights reservation concerns. If, for instance, a collective management organisation (CMO), such as the French SACEM,[67] exercises the opt-out right following from Article 4(3) CDSMD, the opt-out statement and corresponding robots.txt become elements of the website of the collecting society. However, this website is unlikely to contain all the musical works belonging to the repertoire administered by the CMO. If a crawler searching for AI training resources refrains from including the CMO website in the training dataset, the opt out thus remains ineffective. CMO repertoire on other websites may still find its way into the TDM dataset unless these other websites contain the same machine-readable opt-out information. Hence, it is of particular importance to enrich the opt-out statement with sufficiently detailed information on the works falling under the rights reservation. Opt outs under Article 4(3) CDSMD should go beyond the mere rights reservation statement. Ideally, they offer richer copyright data. For instance, it is conceivable that the opt-out statement contains a link to a database providing information on authors, performers, rightholders, titles, genres etc. of works covered by the opt out. In particular, a database link makes sense when rightholders, including CMOs, seek to use the opt-out mechanism as an invitation to enter into licensing agreements. If the opt-out mechanism is developed in this way, Article 4(3) CDSMD can become a propelling force for metadata creation and improvement.

4.3 Bundling of metadata streams

Considering the described need for up-to-date information on protected works, the nature and scope of rights, the question of ownership and the characteristics and contents of the work itself, the particular opportunity presented by Articles 17(4)(b) and 4(3) CDSMD becomes apparent: if all notifications and reservations of rights under these provisions, including metadata, could be harmonised and merged into a central EU copyright data collection, the accumulation of data could lead to a data reservoir that dwarfs existing European data silos of collecting societies, rightholders and distribution platforms.[68] Moreover, as the described legal mechanisms require continuous updating of rights and ownership information, a centralised EU copyright data repository fed by Article 17 CDSMD notifications and Article 4 CDSMD opt-outs could have a relatively high degree of timeliness and accuracy.

However, in order to launch such an EU copyright data repository, metadata flowing from rights notifications and reservations under Articles 17(4)(b) and 4(3) CDSMD would have to be bundled in a systematic way. Therefore, rightholders providing metadata in these contexts

[67] See https://societe.sacem.fr/en/news/our-society/sacem-favour-virtuous-transparent-and-fair-ai-exercises-its-right-opt-out: "Against a backdrop of increasing development of artificial intelligence (AI) tools, Sacem is exercising its right to opt-out on behalf of its members. From now on, data mining of works in Sacem's repertoire by entities developing artificial intelligence tools will require prior authorisation from Sacem, in order to ensure fair remuneration for the authors, composers and music publishers it represents."

[68] In the legislative process leading to the adoption of Article 17 CDSMD, Germany had already proposed in this sense to introduce "public, transparent notification procedures" to counteract a de facto copyright register in the hands of dominant platforms. See Council of the European Union, Opinion of Germany, 5 April 2019, point 5, p. 4, available at: https://data.consilium.europa.eu/doc/document/ST-7986-2019-ADD-1-REV-2/en/pdf.

should be required to submit this information in parallel to a central body managing the EU copyright data repository. The pooling and harmonisation of copyright data could then be done in an open and interoperable format to ensure general data accessibility and data transparency for all interested users. In this way, an overarching copyright data collection could ensure that licensing deals come within reach for all repertoire holders – regardless of size and market power. An EU copyright data repository could thus reduce the risk of large repertoire owners, which have a wider spectrum of works and metadata, gaining a competitive advantage with regard to new licensing opportunities, for instance in the field of AI training.

A template for legislation that would ensure this redirection of copyright data to a central data collection point can already be found in Article 3(6) of the 2012 Orphan Works Directive[69] (in relation to information on the use of orphan works) and in Article 10(1) CDSMD (in relation to information on out-of-print works). Interestingly, these provisions also mention the institution that could take care of the central EU copyright database: the European Union Intellectual Property Office (EUIPO).

In order to achieve the desired interoperability of data, the legal obligation to transmit metadata to the EUIPO could be supplemented with an obligation to provide the data in a specific, standardised format. Hence, the law could be used as a tool to address not only issues of data accuracy and timeliness, but also the problem of data interoperability and data harmonisation. The obligation to submit data in harmonised and interoperable form to the EUIPO would have the advantage for rightholders that they can create a generally accepted data submission standard. This generally accepted standard could lead to universal applicability of submitted metadata. If other addressees of metadata information, such as OCSSPs and AI trainers in the case of data transmissions under Articles 17(4)(b) and 4(3) CDSMD, would legally be bound to accept the information in this standardised format, copyright holders would no longer have to deal with individual and fragmented data transmission standards, which may differ from one data user to the other and require the submission of the same information in various formats.

5. International ban on formalities

The international prohibition of formalities arising from Article 5(2) of the Berne Convention for the Protection of Literary and Artistic Works (BC) need not pose insurmountable obstacles when developing the outlined system of data transmissions in interoperable form. According to Article 5(2) BC, "the enjoyment and exercise" of the rights granted in Article 5(1) BC are not subject to any formal requirement. Article 5(1) includes the rights which the laws of the Berne Union countries "do now or may hereafter grant to their nationals, as well as the rights specially granted by this Convention." As Stef van Gompel explains in his in-depth analysis of the scope of the prohibition in Article 5(2) BC, the ban on formalities:

> includes formalities relating to the coming into existence, the maintenance and the enforcement of copyright. The Berne prohibition on formalities does not extend to formalities that regulate the extent of protection or the means of redress afforded to authors to protect their rights. This suggests that formalities are allowed if they establish the manner of exercising copyright, but not if their non-compliance renders the exercise of rights completely impossible.[70]

[69] Directive 2012/28/EU of the European Parliament and of the Council of 25 October 2012 on certain permitted uses of orphan works, *Official Journal of the European Communities* 2012 L 299, 5.
[70] Stef van Gompel, *Formalities in Copyright Law: An Analysis of Their History, Rationales and Possible Future*, Alphen aan den Rijn: Kluwer Law International 2011, 212.

Within this matrix, the legal tools discussed above – the notification system following from Article 17(4)(b) CDSMD and the opt-out mechanism under Article 4(3) CDSMD – fall into the category of permissible formalities concerning the manner in which copyright is exercised and the regulation of the scope of protection.

This can hardly be denied in the case of the content moderation rules in Article 17 CDSMD: by providing that platform providers carry out an act of communication to the public or an act of making available to the public when they grant the public access to protected works uploaded by users, Article 17(1) CDSMD establishes direct, primary liability of online platforms[71] in an area that has traditionally been governed by the rules on secondary liability for the uploading of infringing content by users.[72] The specific design of liability rules in Article 17 CDSMD, including the possibility to avoid liability by purchasing licences and applying content filters (Article 17(4)(a) and (b) CDSMD), clearly regulate the scope of protection.[73] The fact that rightholders, as discussed above, are required to provide "relevant and necessary" information under Article 17(4)(b) CDSMD shows that the provision establishes a specific way of exercising copyright.[74]

Regardless of the specific liability system following from Article 17 CDSMD, rightholders can always enforce their rights against individual uploaders in situations where platform providers have not been granted a licence. Thus, rather than making the exercise of copyright dependent on formal requirements and "completely impossible",[75] Article 17(4)(b) provides rightholders with an additional means of ensuring the unavailability of their works on online platforms.

The same conclusion can be drawn with regard to the opt-out mechanism in Article 4(3) CDSMD. The rights reservation option in this provision offers copyright holders an additional, extended scope of protection and an additional means of redress in a situation that, from the perspective of international copyright law, falls outside the scope of the right of reproduction

[71] For a more detailed discussion of the legal nature of the right granted in Article 17 CDSMD, see M. Husovec/J.P. Quintais, 'How to License Article 17? Exploring the Implementation Options for the New EU Rules on Content-Sharing Platforms under the Copyright in the Digital Single Market Directive', *Gewerblicher Rechtsschutz und Urheberrecht International* 70 (2021), 325 (325-348).

[72] See M. Leistner, 'European Copyright Licensing and Infringement Liability Under Art. 17 DSM-Directive Compared to Secondary Liability of Content Platforms in the U.S. – Can We Make the New European System a Global Opportunity Instead of a Local Challenge?', *Zeitschrift für Geistiges Eigentum/Intellectual Property Journal* 26 (2020), 123 (123-214); M. Husovec, *Injunctions Against Intermediaries in the European Union – Accountable But Not Liable?*, Cambridge: Cambridge University Press 2017; C. Angelopoulos, *European Intermediary Liability in Copyright: A Tort-Based Analysis*, Alphen aan den Rijn: Kluwer Law International 2016; M.R.F. Senftleben, "Breathing Space for Cloud-Based Business Models – Exploring the Matrix of Copyright Limitations, Safe Harbours and Injunctions", *Journal of Intellectual Property, Information Technology and E-Commerce Law* 4 (2013), 87 (87-90 and 94-95); T. Hoeren/S. Yankova, 'The Liability of Internet Intermediaries – The German Perspective', *International Review of Intellectual Property and Competition Law* 43 (2012), 501; R. Matulionyte/S. Nérisson, 'The French Route to an ISP Safe Harbour, Compared to German and US Ways', *International Review of Intellectual Property and Competition Law* 42 (2011), 55; M. Peguera, 'The DMCA Safe Harbour and Their European Counterparts: A Comparative Analysis of Some Common Problems', *Columbia Journal of Law and the Arts* 32 (2009), 481.

[73] Cf. van Gompel, id. 212.

[74] Cf. van Gompel, id. 212.

[75] Cf. van Gompel, id., 212.

altogether.[76] Moreover, the Berne Convention itself contains a longstanding opt-out system that concerns the reproduction of press articles. Article 10bis(1) BC reads as follows:

> It shall be a matter for legislation in the countries of the Union to permit the reproduction by the press, the broadcasting or the communication to the public by wire of articles published in newspapers or periodicals on current economic, political or religious topics, and of broadcast works of the same character, in cases in which the reproduction, broadcasting or such communication thereof is not expressly reserved.

The reservation of rights reflected in this provision entered the Berne Convention as an element of the debate on the protection of publications in newspapers and periodicals, and the freedom to use news information and newspaper articles with the exception of serial stories and tales.[77] Article 10bis(1) evolved from industry practice more than a century ago. At the time, newspapers considered the reproduction of their articles in other newspapers as an advertisement and promotion of their activities.[78] In particular, local newspapers with limited financial resources could hardly have satisfied the news demand of their readers without reproductions of newspaper articles taken from bigger newspapers.[79]

The forms of exploitation covered by Article 10bis(1) BC – reproduction, broadcasting and communication to the public – are central modes of exploiting press articles and broadcasts. Moreover, the provision concerns the initial exploitation period. Article 10bis(1) BC explicitly refers to "current economic, political or religious topics." Hence, it exempts the use of articles that are still fresh and have news value. As the rightholder can opt out by reserving copyright, however, the risk of overbroad inroads into the market for original news products is minimised. Considering this prototype for opt-out rules in the Berne Convention itself, it can hardly be concluded that opt-out mechanisms, such as Article 4(3) CDSMD, impose impermissible copyright formalities in the sense of Article 5(2) BC.[80]

All in all, the notification system following from Article 17(4)(b) CDSMD and the rights reservation option provided by Article 4(3) CDSMD are permissible formalities that only increase the scope of protection and regulate the way copyright is exercised in specific contexts which, in the absence of the approach taken in the EU, would be a matter of secondary platform liability for infringement (Article 17(4)(b) CDSMD) or an act of use falling outside the realm of copyright protection altogether (Article 4(3) CDSMD). Against this background, the prohibition of formalities following from Article 5(2) BC does not pose obstacles when notifications under Article 17(4)(b) and rights reservations under Article 4(3) are combined with an obligation to create and submit metadata in a standardised format not only to OCSSPs

[76] Senftleben, supra note 55, 1502.
[77] As to the development of the provision in the Berne Convention, see S. Ricketson/J.C. Ginsburg, *International Copyright and Neighbouring Rights – The Berne Convention and Beyond*, 3rd ed., Oxford: Oxford University Press 2022, 796-800; L. Guibault, 'The Press Exception in the Dutch Copyright Act', in: P.B. Hugenholtz/A.A. Quaedvlieg/D.J.G. Visser (eds.), *A Century of Dutch copyright law – Auteurswet 1912–2012*, Amstelveen: deLex 2012, 443 (447-450).
[78] See L. Guibault, *Copyright Limitations and Contracts – An Analysis of the Contractual Overridability of Limitations on Copyright*, The Hague/London/New York: Kluwer Law International 2002, 58, as to the rationales underlying the newspaper exemption.
[79] Guibault, supra note 77, 444-445: "Hardly any newspaper in those days could survive without citing or borrowing articles from prestigious foreign publications."
[80] For a more detailed discussion of these opt-out mechanisms, see Senftleben, 'Generative AI and Author Remuneration', supra note 1, 1544-1546.

and AI trainers but also to a central EU data collection point that could be established at the EUIPO.

6. Conclusion

To improve the visibility and accessibility of the diverse European repertoire of literary and artistic works and to enable the creative industries to benefit from new licensing opportunities, in particular in the area of AI training, it is important to establish a comprehensive copyright data infrastructure focusing on European content – a copyright data repository that would devote sufficient attention to smaller and lesser-known (country) repertoires and reflect the full spectrum of cultural diversity in the EU. As the foregoing analysis has shown, obligations to create and share metadata providing information on creators, rightholders and the nature and characteristics of the work itself could be attached to several provisions in the EU copyright acquis. The notification of "relevant and necessary" information for the purpose of content moderation under Article 17(4)(b) CDSMD provides an important starting point. If Article 17(4)(b) notifications sent to OCSSPs are in parallel collected and pooled in a central EU copyright data repository, the resulting accumulation of EU copyright data could lead to a promising data reservoir. For this purpose, notifications under Article 17(4)(b) would have to be enriched. They would have to be detailed and descriptive enough to allow AI trainers to identify content they find desirable for machine training. It would also have to go beyond the right of communication/making available to the public that is central to the Article 17(4)(b) platform context and include information on holders of reproduction rights. To allow an EU data repository to enhance the visibility of the European repertoire in a meaningful way and expand licensing opportunities for copyright holders, "relevant and necessary" information in the sense of Article 17(4)(b) would thus have to be understood broadly: it should cover a wide range of exclusive rights, including the right of reproduction that occupies centre stage in TDM contexts.[81] It should also include descriptive metadata that provide information on the nature and content of the works notified.

Luckily, Article 17(4)(b) CDSMD is not the only copyright norm that offers impulses for the creation and maintenance of accurate copyright metadata. The opt-out mechanism in Article 4(3) CDSMD – directly addressing AI training because it concerns a potential ban on the use of works for TDM purposes – is a further example of a copyright rule that could generate a continuous flow of metadata providing creator, ownership and work-related information. In this case, sufficiently rich metadata can be expected when rightholders are willing to use the opt-out mechanism as a vehicle to offer a license in exchange for the payment of remuneration. For rightholders seeking to attract the attention of AI trainers and enhance the chances of licensing agreements, it makes sense to go beyond the mere opt-out statement and provide additional information, such as descriptive metadata that allow AI developers to determine whether the repertoire – genre, content etc. – is desirable for machine training purposes. Furthermore, metadata are necessary that provide contact details, licensing conditions and potential options for automated rights clearance.

However, the transformation of copyright instruments, such as Article 17(4)(b) notifications and Article 4(3) opt outs, into data improvement tools is only one central building block of a data improvement strategy for Europe. In addition, copyright law and practice[82] should devote

[81] Articles 3 and 4 CDSMD. Cf. Senftleben, supra note 3, 36-37.
[82] Article 17(10) CDSMD.

attention to data harmonisation and interoperability. To provide data that are not only accurate but also harmonised and interoperable, the obligation to transmit metadata to a central EU data collection point could be supplemented with an obligation to provide the data in a specific, standardised format. Hence, the law could be used as a tool to address not only issues of data accuracy and timeliness, but also the problem of data interoperability and data harmonisation. The obligation to submit enriched Article 17(4)(b) notifications and Article 4(3) opt-out statements in a generally-agreed format to a central data repository could enable rightholders to set a data submission standard. This generally accepted standard could pave the way for universal applicability of submitted metadata. If other addressees of metadata information – OCSSPs and AI trainers in the case of data transmissions under Articles 17(4)(b) and 4(3) CDSMD – would legally be bound to accept the information in this standardised format, copyright holders would no longer have to deal with potentially fragmented data transmission standards, which may differ from one data user to the other and require the submission of the same information in various formats.