Martin Senftleben,[a] Thomas Margoni,[b] Daniel Antal,[c] Balázs Bodó,[d] Stef van Gompel,[e] Christian Handke,[f] Martin Kretschmer,[g] Joost Poort,[h] João Quintais,[i] Sebastian Schwemer[j]

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

In the European Strategy for Data*, the European Commission highlighted the EU's ambition "to acquire a leading role in the data economy."*[1] *At the same time, the Commission conceded that the EU would have to "increase its pools of quality data available for use and re-use."*[2] *In the creative industries,*[3] *this need for enhanced data quality and interoperability is particularly*

[a] Professor of Intellectual Property Law and Director, Institute for Information Law (IViR), University of Amsterdam, The Netherlands.
[b] Research Professor of Intellectual Property Law, Centre for IT & IP Law (CiTiP), Faculty of Law, KU Leuven, Belgium.
[c] Independent Researcher, The Hague, The Netherlands.
[d] Associate Professor, Institute for Information Law (IViR), University of Amsterdam, The Netherlands.
[e] Associate Professor, Institute for Information Law (IViR), University of Amsterdam, The Netherlands.
[f] Associate Professor of Cultural Economics, Erasmus University Rotterdam, The Netherlands.
[g] Professor of Intellectual Property Law and Director, CREATe, University of Glasgow, United Kingdom.
[h] Associate Professor and Co-Director, Institute for Information Law (IViR), University of Amsterdam, The Netherlands.
[i] Postdoctoral Researcher, Institute for Information Law (IViR), University of Amsterdam, The Netherlands.
[j] Associate Professor, Centre for Information and Innovation Law (CIIR), University of Copenhagen, Denmark; Adjunct Associate Professor, Norwegian Research Center for Computers and Law (NRCCL), University of Oslo, Norway.
[1] European Commission, 19 February 2020, "A European Strategy for Data", Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Document COM(2020) 66 final, 1, available at: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy.
[2] European Commission, id., 1.
[3] As to the contours of the cultural and creative sectors, see the definition provided in Article 2(1) of the Regulation 1295/2013 of 11 December 2013 establishing the Creative Europe Programme (2014 to 2020), available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32013R1295: "all sectors whose activities are based on cultural values and/or artistic and other creative expressions, whether those activities are market- or non-market-oriented, whatever the type of structure that carries them out, and irrespective of how that structure is financed. Those activities include the development, the creation, the production, the dissemination and the preservation of goods and services which embody cultural, artistic or other creative expressions, as well as related functions such as education or management. The cultural and creative sectors include inter alia architecture, archives, libraries and museums, artistic crafts, audiovisual (including film, television, video games and multimedia), tangible

strong (section 1). Without data improvement, unprecedented opportunities for monetising the wide variety of creative content in EU Member States and making this content available for new technologies, such as artificial intelligence ("AI") systems, will most probably be lost (section 2). The problem has a worldwide dimension. While the US have already taken steps to provide an integrated data space for music as of 1 January 2021,[4] the EU is facing major obstacles not only in the field of music but also in other creative industry sectors (section 3).[5] Weighing costs and benefits (section 4), there can be little doubt that new data improvement initiatives and sufficient investment in a better copyright data infrastructure should play a central role in EU copyright policy. A trade-off between data harmonisation and interoperability on the one hand, and transparency and accountability of content recommender systems on the other, could pave the way for successful new initiatives (section 5).

1. Introduction

Since the early days of the digital revolution, the dream of the free flow of information across cultures and continents has been accompanied by the hope that digital rights management in the area of copyright ("DRM") would maximise the spectrum of available literary and artistic productions (including content for niche audiences), minimise transaction costs, pave the way for ubiquitous and differentiated licensing solutions and allow the creative industries to thrive. In reaction to the challenges arising from the digital environment, the 1996 WIPO "Internet" Treaties[6] introduced new international standards against the circumvention of technological measures that are employed to protect copyrighted works, and the removal or alteration of copyright management information.[7] The 2001 Directive on the Harmonisation of Copyright and Related Rights in the Information Society ("Information Society Directive" or "ISD")[8] transposed these international standards into EU law.

Besides applications by individual companies, the issue of copyright data management – in the sense of attaching and standardising metadata to works stemming from various authors and producers – has traditionally played a crucial role in the area of collective licensing of creative content. Nowadays, content distribution platforms that operate internationally, such as Spotify, iTunes, YouTube, Netflix and Getty Images, play a central role as well. With Article 17 of the Directive on Copyright in the Digital Single Market ("Digital Single Market Directive" or "DSMD"),[9] the topic receives an important additional dimension. Article 17 addresses specifically online platforms that allow users to upload and share user-generated content

---

and intangible cultural heritage, design, festivals, music, literature, performing arts, publishing, radio and visual arts."

[4] See the information on the launch of the Music Licensing Collective (MLC) at https://www.themlc.com/press/mechanical-licensing-collective-begins-full-operations-envisioned-music-modernization-act.

[5] As to data space priorities in the EU, see European Commission, supra note 1, 22-23 and Appendix.

[6] WIPO Copyright Treaty and WIPO Performances and Phonograms Treaty, adopted in Geneva on December 20, 1996. See https://www.wipo.int/treaties/en/ip/wct/ and https://www.wipo.int/treaties/en/ip/wppt/.

[7] Articles 11 and 12 of the WIPO Copyright Treaty; Articles 18 and 19 of the WIPO Performances and Phonograms Treaty.

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

[9] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9/EC and 2001/29/EC, OJ 2019 L 130, 92.

("UGC").[10] The collaboration between the creative industry and these platforms – Online Content Sharing Service Providers ("OCSSPs")[11] – has already led to the creation of content identification systems (and corresponding databases) in the past, and can be expected to foster the establishment of more extensive content libraries and corresponding metadata for the purposes of online content identification and moderation in the future.

While the digital environment, in theory, offers unprecedented opportunities for commercialising literary and artistic productions and serving consumers, several practical problems prevent the creative industries from realising the full potential of copyright data management and digital modes of exploitation to this day. The lack or inaccuracy of metadata prevents or delays the disbursement of royalties. Moreover, inaccurate and incomplete metadata make content hard to find, or license, and, as a result, may contribute to digital piracy. From an economic perspective, it may be said that even if certain content is technically available via legal channels, inaccurate and incomplete metadata may increase search costs for users to such an extent that data problems de facto create incentives to make unauthorised

[10] As to the underlying debate on new licensing and content moderation obligations, see Geiger, Christophe and Jütte, Bernd Justin, "Platform liability under Article 17 of the Copyright in the Digital Single Market Directive, Automated Filtering and Fundamental Rights: An Impossible Match", available at https://ssrn.com/abstract=3776267; Sebastian Felix Schwemer, "Article 17 at the Intersection of EU Copyright Law and Platform Regulation", *Nordic Intellectual Property Law Review* 2020, 400-435; Martin R.F. Senftleben, "Institutionalized Algorithmic Enforcement – The Pros and Cons of the EU Approach to Online Platform Liability", Florida International University Law Review 14 (2020), 299-328; Martin Husovec and João Pedro Quintais, "How to License Article 17? Exploring the Implementation Options for the New EU Rules on Content-Sharing Platforms", *Gewerblicher Rechtsschutz und Urheberrecht International*, forthcoming 2021, doi: 10.1093/grurint/ikaa200, available at: https://ssrn.com/abstract=3463011, 10-20; Matthias 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), 123-214; João Pedro Quintais/Giancarlo Frosio/Stef van Gompel et al. "Safeguarding User Freedoms in Implementing Article 17 of the Copyright in the Digital Single Market Directive: Recommendations from European Academics", *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* 10 (2020), 277-282; Giancarlo Frosio, "Reforming the C-DSM Reform: A User-Based Copyright Theory for Commonplace Creativity", *International Review of Intellectual Property and Competition Law* 51 (2020), 709 (724-726); Sebastian Felix Schwemer and Jens Schovsbo, "What is Left of User Rights? – Algorithmic Copyright Enforcement and Free Speech in the Light of the Article 17 Regime", *Intellectual Property Law and Human Rights,* 4th ed., Alphen aan den Rijn: Wolters Kluwer 2020, 569-589; Martin Senftleben, "Bermuda Triangle: Licensing, Filtering and Privileging User-Generated Content Under the New Directive on Copyright in the Digital Single Market" 41(8) (2019) *European Intellectual Property Review*, 480 (483-484); M Senftleben, et al, "The Recommendation on Measures to Safeguard Fundamental Rights and the Open Internet in the Framework of the EU Copyright Reform", European Intellectual Property Review 40 (2018), 149; C Angelopoulos, "On Online Platforms and the Commission's New Proposal for a Directive on Copyright in the Digital Single Market" (2017), available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2947800; G Frosio, "From Horizontal to Vertical: An Intermediary Liability Earthquake in Europe", Oxford Journal of Intellectual Property and Practice 12 (2017), 565-575; G Frosio, "Reforming Intermediary Liability in the Platform Economy: A European Digital Single Market Strategy", Northwestern University Law Review 112 (2017), 19; RM Hilty & V Moscon V. (eds.), "Modernisation of the EU Copyright Rules – Position Statement of the Max Planck Institute for Innovation and Competition", Max Planck Institute for Innovation and Competition Research Paper No. 17-12, Max Planck Institute for Innovation and Competition: Munich 2017.

[11] See the definition in Article 2(6) DSMD. For a more detailed discussion, see A. Metzger/M. Senftleben, "Comment of the European Copyright Society: Selected Aspects of Implementing Article 17 of the Directive on Copyright in the Digital Single Market into National Law", Journal of Intellectual Property, Information Technology and Electronic Commerce Law 10 (2020), 115-131.

use where copyright enforcement is weak. Alternatively, potential uses of works may simply be forgone due to such transaction costs. In addition to these problems at the level of individual data sets, the lack of interoperability between data management systems and related data libraries forces stakeholders to deal with a highly inefficient, and often inaccurate, piecemeal network of data providers, systems, datasets, and standards. It increases all types of transaction costs because it obliges stakeholders to learn about, identify, and deal with various types of metadata, as well as individual terms and modalities of use. The high costs of dealing with inaccurate and incomplete metadata may moreover favour big providers of copyright-intensive products and services who can afford to invest in database building, data cleansing, and are capable of bearing the costs of lawsuits arising from data-related conflicts. This enhances the risk of economic concentration in the digital content distribution market and a corresponding power imbalance between copyright holders and content distributors, such as online platforms.

2. Need for Improved Copyright Data Management

Emerging new technologies that require the use of large repertoires of creative content shed light on the dimension of transaction cost problems in the creative industries – and the risk of losing substantial revenue. The situation in the field of AI systems can serve as an example. For a long time, mankind assumed that only humans were capable of creating literary and artistic works. With developments in the field of AI giving birth to a new kind of algorithmic work creation in the realm of cultural creativity, this assumption no longer seems valid. Today, AI systems increasingly assist in the creation of works of art and literature ("AI-assisted works"). Sometimes, on the basis of appropriate training material, they may also be capable of mimicking human literary and artistic productions, such as poems, music and paintings ("AI-generated works").[12] The technology enabling their creative functions is becoming more and more advanced and instead of fully relying on human instructions, contemporary AI systems are becoming increasingly autonomous. Certain types of deep-learning systems may give users the impression of being capable of cultural creation, potentially almost independently,

[12] See A. Elgammal, B. Liu, M. Elhoseiny, M. Mazzone, "*CAN:* Creative Adversarial Networks Generating "Art" by Learning About Styles and Deviating from Style Norms", June 2017, available at: https://www.researchgate.net/publication/317823071_CAN_Creative_Adversarial_Networks_Generating_Art_by_Learning_About_Styles_and_Deviating_from_Style_Norms, 17 (Elgammal and his fellow researchers carried out an experiment to determine whether humans were capable of distinguishing computer-generated art from human art by its appearance. 75% of the research subjects assumed that the computer-generated paintings were created by a human artist). Cf. P.B. Hugenholtz, D. Gervais, J.P. Quintais, C. Hartmann & J. Allan, "Trends and Developments in Artificial Intelligence: Challenges to the Intellectual Property Rights Framework: Final Report, Report written for the European Commission, Amsterdam: Institute for Information Law 2020, available at https://op.europa.eu/en/publication-detail/-/publication/394345a1-2ecf-11eb-b27b-01aa75ed71a1/language-en; M.R.F. Senftleben/L.D. Buijtelaar, "Robot Creativity: An Incentive-Based Neighbouring Rights Approach", European Intellectual Property Review 42, No. 12 (2020), 797-812; D. Gervais, "The Machine as Author", Iowa Law Review 105 (2020), 2053; J.C. Ginsburg & L.A. Budiardjo, "Authors and Machines", Berkeley Technology Law Journal 34 (2019), 343 (395-396); M.-C. Janssens & F. Gotzen, "Kunstmatige Kunst. Bedenkingen bij de toepassing van het auteursrecht op Artificiële Intelligentie", Auteurs en Media 2018-2019, 323 (325-327); W.T. Ralston, "Copyright in Computer-Composed Music: HAL Meets Handel", J. Copyright Society U.S.A. 52 (2005), 281; S. Yanisky & S. Moorhead, "Generating Rembrandt: Artificial Intelligence, Copyright and Accountability in the 3A Era", Michigan State Law Review (2017), 659 (662); A. Bridy, The Evolution of Authorship: Work Made by Code, Columbia Law Journal & Arts 39 (2016), 395 (397); R.C. Denicola, "Ex Machina: Copyright Protection for Computer-Generated Works", Rutgers University Law Review 69 (2016), 251.

allowing for broad-scale production of cultural objects which eye and ear often fail to distinguish from human creations.[13]

In this context, however, it must not be overlooked that "artificial creativity" is impossible without source material in a harmonised and interoperable format that can be used for feeding and instructing AI systems. Without machine-readable literary and artistic input stemming from authors of flesh and blood, an AI system has no template for its own processes of mimicking human creativity. Modern data-driven statistical AI often uses Text-and-Data Mining ("TDM")[14] techniques to extract the data needed for machine learning. TDM has emerged as one of the most powerful digital tools in the AI environment which enables the discovery and extraction of patterns, correlations and more generally of (often hidden) knowledge from existing content and data.[15] Both high-tech and creative industries are currently being revolutionised by the advancements in this data-driven type of AI. Techniques that are currently discussed under the headings of Machine Learning ("ML"), Natural Language Processing ("NLP") and Deep Neural Networks ("DNN"), require the "training" on vast amounts of content and data in order to achieve reliable results that may finally lead to new scientific and technological advancements, products and services. This information is often deduced, through automated machine-reading processes, from books, magazine articles, music works or their fixations, or films enjoying copyright protection. Not surprisingly, the insatiable appetite of "creative" AI systems for literary and artistic data input is often regarded as a promising new source of revenue for the creative industries.[16]

The use of copyrighted works as training material for this type of AI applications, however, raises complex questions. When humans learn a new task or skill (e.g. a new language), they usually store the training information (e.g. the textbook rules and examples used to learn the language) as an electrochemical trace in the area of the brain dedicated to language. Humans do not need a copyright exception in order to store that copy. However, it is far from clear that when a computer makes the corresponding digital copy of training material in order to learn a language – or any other task for that matter – this activity is likewise excluded from the copyright domain. On the contrary, the use of any digital copy, temporary or permanent, in

---

[13] The impact that AI is having in the field of IP, and copyright in particular, has been recognised by the European Commission, which has specifically identified a number of ambitious interventions in this area in its recent "IP Action Plan", see European Commission, 15 November 2020, "Making the most of the EU's innovative potential – An intellectual property action plan to support the EU's recovery and resilience", Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Document COM(2020) 760 final, p. 12.

[14] The abbreviation "TDM" is used here for text-and-data mining in accordance with the use that has become customary in the domain of copyright. It is not to be confused with "term document matrix" – an important standard organizational form of data describing natural language texts for NLP algorithms.

[15] Margoni T., Text and Data Mining in Intellectual Property Law: Towards an Autonomous Classification of Computational Legal Methods, CREATe working paper 01/2020 Forthcoming in Calboli I. & Montagnani L., Handbook on Intellectual Property Research, available at: https://www.create.ac.uk/blog/2020/05/01/new-working-paper-text-and-data-mining-in-intellectual-property-law-towards-an-autonomous-classification-of-computational-legal-methods/.

[16] Cf. Covington, Paul, Jay Adams, and Emre Sargin. 2016. "Deep Neural Networks for Youtube Recommendations." In Proceedings of the 10th Acm Conference on Recommender Systems, 191–98. RecSys '16. New York, NY, USA: Association for Computing Machinery, https://doi.org/10.1145/2959100.2959190; Jacobson, Kurt, Vidhya Murali, Edward Newett, Brian Whitman, and Romain Yon. 2016. "Music Personalization at Spotify." In Proceedings of the 10th Acm Conference on Recommender Systems, 373. RecSys '16. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2959100.2959120.

whole or in part, direct or indirect, may amount to the infringement of the right of reproduction laid down in Article 2 ISD.

The right of reproduction thus constitutes a pivotal element in AI training processes. ML-based systems may require numerous and different types of reproductions: certain copies may be just temporary (the conversion of .pdf into .xml for annotation and enrichment purposes), others may be permanent (the initial creation of *corpora* or databases of training material, or the final storage of said material for replicability, accountability and verifiability of the training process). Some copies may be in whole (such as the initial reproduction of the *corpora*), others may be in part (such as the information stored in the "trained models" which will be used by the AI algorithm to perform the intended task). Finally, some reproductions may be direct and others may be only indirect (again the final "trained models" may contain only partial and modified copies of the original material). Further steps in the AI training process and the distribution and use of the final outcome may involve additional rights that are exclusively reserved to copyright holders, such as the right of distribution and the right of communication to the public. If no exceptions or limitations permit the use of copyrighted material without authorisation, all these individual acts of use require licenses.

Against this background, appropriate copyright data management and licensing infrastructures are not only desirable to offer the creative industries the opportunity of exploiting the promising new market for AI training data. Improved copyright data management is also indispensable to enable EU high-tech industries to compete with AI system developers in other regions. In Article 3(1) DSMD, EU legislation has granted a statutory permission to reproduce literary and artistic works for AI training purposes. This limitation of copyright protection, however, only covers TDM in the context of scientific research carried out by non-profit research organisations and cultural heritage institutions.[17] Article 4(1) DSMD supplements this research privilege with a general TDM exemption that can also be invoked by commercial AI system developers. This broader copyright limitation, however, is only applicable as long as copyright holders refrain from reserving their exclusive rights under Article 4(3) DSMD. The need to obtain licenses for commercial applications is thus the rule in EU copyright law; a use permission without prior rightholder authorisation is the exception. With regard to commercial AI training, Article 5(1) ISD only provides a loophole for TDM processes that keep within the confines of transient, temporary copying.[18] This restrictive

[17] Cf. the definition in Article 2(1) and (3) DSMD.

[18] CJEU, 16 July 2009, case C-5/08, Infopaq/Danske Dagblades Forening, para. 56-58; CJEU 17 januari 2012, case C-302/10, Infopaq II, para. 36, 44 and 51-56. Cf. Geiger et al (2018) The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Legal Aspects, Policy Department for Citizens' Rights and Constitutional Affairs, Directorate General for Internal Policies of the Union PE 604.941- February 2018, available at: https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/604941/IPOL_IDA(2018)604941_EN.pdf; Christophe Geiger et al., "Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?", *International Review of Intellectual Property and Competition Law* 49 (2018), 814 (814-844); Thomas Margoni, "AI, Machine Learning and EU Copyright Law: Who owns AI?", *Annali Italiani del Diritto d'Autore, della Cultura e dello Spettacolo* XXVII (2018); 281 (281-304); Rossana Ducato and Alain Strowel, "Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to 'Machine Legibility'", *International Review of Intellectual Property and Competition Law* 50 (2019), 649; Eleonora Rosati, "An EU Text and Data Mining Exception for the Few: Would it Make Sense?", *Journal of Intellectual Property Law and Practice* 13 (2018), 429 (429-430); Andres Guadamuz and Diane Cabell, "Data Mining in UK Higher Education Institutions: Law and Policy", *Queen Mary Journal of Intellectual Property* 4 (2014), 3 (3-29).

approach may be insufficient for the needs of high-tech firms focusing on AI development. Considering current industry practices, it seems safe to assume that more than temporary takings from copyrighted source material will be necessary in many cases.

Main international competitors of the EU have chosen an approach that markedly departs from the focus on copyright licensing adopted in Europe. Countries such as the US, Canada, Singapore, South Korea, Japan, Israel or Taiwan have adopted regulatory measures which, in the natural tension between the protection of investments and the promotion of innovation, have opted for broader copyright limitations arguably favouring the latter over the former. The specific measures that have been adopted in order to gauge the proper balance have evolved from, and thus mirror, the domestic legal culture and characteristics. In the US, for instance, TDM and ML analyses are routinely considered to be transformative uses and as such to constitute fair use which is permissible without the prior authorisation of the right holder and which does not generate claims for fair compensation. This means that using protected works not as works but as input data to extract information that will be used to create new knowledge – so called non-consumptive or non-expressive uses[19] – is considered a free activity that does not require licensing efforts. Japan is another interesting example as its copyright law can be considered closer to continental-European models. Instead of a broad standard (i.e. fair use), Japanese copyright legislation provides for a list of exceptions and limitations that resembles to a certain degree the approach taken in Article 5 ISD. Japan has implemented in its copyright legislation a broad TDM exception back in 2009. This provision refrains from precluding commercial users from invoking the TDM exception.[20] The US and Japan are interesting examples because, while belonging to different copyright traditions, they both have thriving creative and cultural industries as well as a highly competitive high-tech sector in the field of AI.

Considering this global scenario, it is of particular importance to establish efficient copyright data management and licensing infrastructures. In the current policy debate, creative industry representatives in European countries often express a preference for a restrictive approach that only leaves room for narrow copyright exceptions. They fear that a more flexible solution would allow the high-tech industry to exploit copyrighted source material for AI training purposes without sharing the benefits that accrue from the development of AI products and services on this basis. This approach may disadvantage EU-based high-tech industries in comparison with their peers in other legal systems that are willing to favour the high-tech sector. The need to obtain an authorisation to train AI algorithms on vast amounts of data – including copyrighted works – constitutes an additional cost factor in the form of transaction costs and licensing fees. When the costs involved are too high, it will negatively impact the

[19] Sag M., Copyright and Copy-Reliant Technology, Northwestern University Law Review, Vol. 103, 2009; Hargreaves I., Digital Opportunities - A Review of Intellectual Property and Growth; UK Department for Business, Innovation and Skills, 18 May 2011.

[20] The Japanese Copyright Act envisages an exception for TDM that is not limited to non-commercial or to research only purposes, see Art. 47-septies Japanese Copyright Act reported and discussed in Guibault & Margoni (2015) Legal Aspects of Open Access to Publicly Funded Research, in OECD (Eds) Enquiries into intellectual property's economic impact, Chapter: 7, OECD, 373 – 414, 396 available at: https://www.oecd.org/sti/ieconomy/intellectual-property-economic-impact.htm. See also Future TDM (2016), Baseline report of policies and barriers of TDM in Europe, 75-76.

ability of the EU's AI sector to compete on the world market and consequently reduce the potential economic value of licensing content for training purposes.[21]

Against this background, the concern must be taken seriously that, in terms of regulatory competition, foreign countries opting for less strict regulatory solutions may appear more attractive to high-tech businesses. Appropriate solutions for copyright data management in the EU, however, may change the equation. Enhanced cooperation between high-tech companies and the creative industries on the basis of licensing agreements, mutually-agreed use protocols and safeguards against algorithms that disregard competition and media regulations may increase the quality and customisation of AI input also. Benefits flowing from enhanced cooperation and better input quality may compensate the costs arising from an obligation to obtain licenses while, at the same time, ensuring that the benefits of copyright-based AI training are fairly shared.

3. Herculean Task of Copyright Data Improvement

A scenario with mutual benefits for creative and high-tech industries, however, will only arise if the considerable problems and obstacles in the field of copyright data management can be overcome. To better illustrate the problems and obstacles arising from (meta-)data obstacles to efficient licensing in European creative industries, the situation in the music sector can serve as a starting point.

3.1 Experiences in the Music Industry

The music segment of the creative industry offers several well-known examples of data infrastructures, such as the Common Information System ("CIS") of the International Confederation of Societies of Authors and Composers ("CISAC"). With its different nodes in several regions of the world, the CIS-Net system and accompanying standards constitute a global tool seeking to facilitate music licensing and the distribution of revenues.[22] In terms of data standardisation, the International Standard Work Code ("ISWC") of the music publishing industry,[23] the International Standard Recording Code ("ISRC") of the recording industry, the Interested Party Information ("IPI") number, and the International Standard Name Identifier ("ISNI") offer prime examples of existing initiatives to enable the exchange of accurate data related to the identification of repertoire or related to the mitigation of ex post transaction costs that arise in relation to the operation of licensing agreements.

At the same time, these examples reveal data deficiencies and interoperability problems arising from different sets of metadata and different approaches to data identification and verification. To this day, initiatives to harmonize ISWC and ISRC metadata and incorporate them into a single, comprehensive database have failed. In the EU, former Commissioner

[21] Handke, C., Guibault, L., & Vallbé, J. J. (2015). Is Europe falling behind in data mining? Copyright's impact on data mining in academic research. New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science: Scale, Openness and Trust - Proceedings of the 19th International Conference on Electronic Publishing, Elpub 2015.

[22] See https://www.cisac.org/What-We-Do/Information-Services/CIS-Net.

[23] ISWC has been developed by CISAC, in collaboration with ISO, as "a unique, permanent, and internally recognized reference number for the identification of musical works". As an example of a further unique identifier system, see also GRiD (Global Release Identifier) which has been developed by IFPI. Cf. Katz, "The potential demise of another natural monopoly: New technologies and the administration of performing rights", 276.

Neelie Kroes launched a working group to stimulate the establishment of a Global Repertoire Database ("GRD") in 2008. While the working group participants, including producers, collective management organisations ("CMOs") and distribution platforms, arrived at recommendations on the way forward,[24] the project was abandoned in 2014.[25] Other unsuccessful attempts include the International Music Joint Venture in 2000, which was formed by several CMOs in Europe and North America, and a project initiated by the World Intellectual Property Organization ("WIPO") aiming at the establishment of a common rights database in 2011.[26]

In the US, by contrast, a new initiative to form a comprehensive database follows from the 2018 Music Modernization Act ("MMA").[27] In Title I, the MMA establishes the Mechanical Licensing Collective ("MLC") as a one-stop shop for obtaining music licenses. For this new licensing body to function properly, it is necessary to have an authoritative and comprehensive database of music rights in place.[28] The MLC seeks to achieve this goal by working closely together with major providers of music streaming services, in particular Apple and Spotify.[29] The new licensing hub offers a US-wide platform for licence administration, enforcement, and royalty processing as of 1 January 2021.[30]

This recent US initiative shows that – despite general metadata infrastructures, such as the CIS-Net system and the ISWC/ISRC standards – a strong need is felt in the music industry to combine, streamline and improve rights databases and establish overarching licensing platforms. New initiatives in Europe point in the same direction. The Technical Online Working Group Europe ("TOWGE") brings together a large group of European CMOs, music publishers and rights agencies developing a digital royalty processing system. TOWGE is based on a small group of direct licensors reporting back to local societies.[31] An initiative with similar objectives has been taken by the Finnish CMO Teosto. A collaboration between Teosto and the start-up company Mind Your Rights has led to the "Concertify" platform seeking to provide

---

[24] 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.

[25] Cf. 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/; Schwemer, S.F. (2019). 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), pp. 68–73.

[26] Schwemer, S.F. (2019). 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), pp. 69–70.

[27] H.R. 1551, Pub. L. 115–264.

[28] Cf. F. Lyons/H. Sun/D. Collopy et al., "Music 2025 – The Music Data Dilemma: issues facing the music industry in improving data management", Newport: UK Intellectual Property Office 2019, available at: https://www.gov.uk/government/publications/music-2025-the-music-data-dilemma, p. 34.

[29] See https://www.appleworld.today/blog/2019/11/18/apple-spotify-to-fund-new-music-royalties-collective.

[30] See https://www.themlc.com/press/mechanical-licensing-collective-begins-full-operations-envisioned-music-modernization-act. As to the underlying planning and preparations, see 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, p. 21, available at: https://bw-98d8a23fd60826a2a474c5b4f5811707-bwcore.s3.amazonaws.com/photos/Proposed_MLC_-_Reply_Comments.pdf.

[31] See https://www.digitalmusicnews.com/2019/07/26/towge-digital-royalty-group/.

– on top of existing industry structures – an efficient and transparent cross-border copyright licensing system. Concertify allows artists, copyright holders, including CMOs, music publishers and event organisers to interact directly by using modules, such as a module for setlist reporting.[32] With the support of the Slovak Art Council, a collaboration between the collecting society SOZA and various stakeholders has led to the creation of a prototype for a comprehensive data and metadata database of the Slovak music repertoire. The consortium also created the prototype of a "Listen Local" recommender system that meets the requirements of the trustworthy AI recommendations of the High-Level Working Group on AI.[33] The accompanying feasibility study highlighted and quantified the problems that arise from incomplete copyright data in existing databases and commercial AI-solutions. For example, it demonstrated that at least 15% of Slovak, Estonian, Hungarian and Dutch works are unlikely to be ever exploited due to data problems.[34] In the area of standardisation, the work of Digital Data Exchange ("DDEX") is of particular interest. The DDEX system has continuously been expanded to all aspects of the digital music value chain. At the interface between ISWC and ISRC, it provides linkages between work and recording data.[35]

3.2 Steps Taken in Other Creative Industry Segments

Other sectors of the creative industry are facing similar data problems and have embarked on initiatives for data improvement, harmonisation and combination as well. In the field of book publishing, industry initiatives, such as the establishment of different e-book platforms and catalogues, play an important role. Flickr and Google Images offer a search option for material covered by a creative commons licence.[36] Another example is the Entertainment Identifier Registry (EIDR), which is a universal unique identifier system for movie and television assets based on DOI technology.[37] As to standardisation, the International Standard Book Number ("ISBN"), the International Standard Serial Number ("ISSN") for journals, the International Standard Music Number ("ISMN") for notated music, and the International Standard Audiovisual Number ("ISAN") for audiovisual works can serve as examples. Moreover, the standardisation work of the international EDItEUR group – leading to the "ONIX" family of standards[38] – is important in the field of books, e-books and serials.[39] With regard to the digital environment, the International DOI Foundation provides the aforementioned Digital Object Identifier ("DOI") services and registration: 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.[40]

[32] See https://www.mindyourrights.fi/.
[33] See https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
[34] Daniel Antal, Feasibility Study On Promoting Slovak Music In Slovakia & Abroad. The Creation of the Comprehensive Slovak Music Database, the Slovak Music Monitor and the Slovak Music Recommendation System, with a supporting Demo Slovak Music Database and the Listen Local Recommendation System, forthcoming 2021. Cf. https://reprex.nl/project/listen-local/.
[35] See https://ddex.net/about-ddex/purpose/.
[36] See https://www.theverge.com/2020/8/31/21408305/google-images-photo-licensing-search-results (Google Images) and https://www.flickr.com/creativecommons/ (Flickr).
[37] See https://www.eidr.org/.
[38] See https://www.editeur.org/8/ONIX/.
[39] See https://www.editeur.org/2/About/#Intro.
[40] See https://www.doi.org/.

In the area of visual arts, CISAC's Visual Arts Council has extended its initial work on the right of resale and established an online licensing hub[41] under the umbrella of the International Council of Creators of Graphic, Plastic and Photographic Arts ("CIAGP").[42] OnLineArt ("OLA") is a one-stop shop for obtaining licenses for worldwide online use of works of visual art currently encompassing works of 60.000 artists.[43] While existing initiatives in the visual arts sector – in particular museums and other cultural heritage institutions digitising works in their holdings – have substantially extended the data coverage of works of fine art, the situation in the field of photography and illustrations is much less transparent. Major visual arts libraries, such as Getty Images, may consistently use data management tools. The costs of properly documenting individual works, however, may be prohibitively high for smaller providers of photography and illustrations in the light of the low average value of individual works.[44] In comparison with the status quo reached in the field of music, the process of harmonising, attaching and bundling (meta-)data still seems in its infancy in the area of visual arts.

3.3 Supportive New Technologies

In the discussion on copyright data improvement, it is important to note that the lack of high quality, publicly accessible metadata for copyrighted material also prompted intense innovation among technology developers. Existing initiatives show that new technologies, in particular AI and blockchain, may serve as catalysts for the streamlining and improvement of copyright data. The aforementioned Concertify platform, for instance, is the result of a collaboration between Teosto and the start-up company Mind Your Rights. The nucleus of the Concertify system for efficient and transparent cross-border copyright licensing was a setlist app which Mind Your Rights had initially developed for Teosto to facilitate setlist reporting on the basis of blockchain technology.[45] Similarly, ASCAP, SACEM and PRS launched a partnership[46] to "prototype a new shared system of managing authoritative music copyright information using blockchain technology."[47] The concept of the project is to develop a blockchain-based solution built on IBM's Hyperledger Fabric that links and manages two standards for copyright-protected content used for music recordings: the International Standard Recording Code (ISRC) and the International Standard Work Code (ISWC). The link between these data would improve royalty matching and licensing. The ultimate goal of the project is to enable a "shared, decentralized database of musical work metadata with real-time update and tracking capabilities."[48]

These examples reflect initiatives to employ distributed ledger (blockchain) technology as a technological architecture for creating and operating shared metadata resources in highly fragmented domains of literary and artistic production. The underlying projects seek to recognise and respond to the metadata issues in the area of copyright. The initiatives, however, may stem from tech companies outside the literary and artistic field – a fact that may

[41] See https://www.cisac.org/What-We-Do/Creators-Relations/CIAGP.
[42] See http://www.ciagp.org/.
[43] See https://onlineart.info/.
[44] Cf. R.A. Posner, "Transaction Costs and Antitrust Concerns in the Licensing of Intellectual Property", *John Marshall Review of Intellectual Property Law* 4 (2005), 325.
[45] See https://www.mindyourrights.fi/.
[46] See https://www.ascap.com/press/2017/04-07-ascap-sacem-prs-blockchain.
[47] See https://societe.sacem.fr/en/press-resources/per-publication/press-releases/ascap-sacem-and-prs-for-music-initiate-joint-blockchain-project-to-improve-data-accuracy-for-rightsholders.
[48] Id. See also https://www.ascap.com/press/2017/04-07-ascap-sacem-prs-blockchain.

indicate structural problems preventing the incumbent creative industries from embracing and fully developing the potential of new technologies. Substantial further innovation in the field was clearly limited by the lack of high quality, comprehensive metadata, which prompted some start-ups to experiment with bottom-up, collaborative metadata pooling, similar to the efforts made for establishing Wikidata.[49]

3.4 Different Settings for Data Improvement

The described experiences with existing data infrastructures and current initiatives to arrive at better results shed light on different settings for the improvement of copyright data management. The initiative to harmonise, combine and enhance the coverage of work-related data may come from different actors and employ different public and private tools:

- *legislation*: the MLC, for instance, is the result of US legislation that explicitly mandates the establishment of a nationwide licensing hub for mechanical music rights. In the EU, Article 17 DSMD, indirectly, may have similar effects if the new obligations to license user-uploaded content and exchange work-related data for content moderation purposes leads to shared data standards and content identification libraries. In addition, the 2014 Directive on Collective Management of Copyright and Related Rights ("Collective Rights Management Directive" or "CRMD")[50] incentivizes CMOs to cooperate in licensing hubs for multi-territorial licensing of online rights in musical works and adopt voluntary industry standards to improve efficiency in the exchange of data. Any legislation at national or EU level for the improvement of copyright data management, however, must observe Article 5(2) of the Berne Convention for the Protection of Literary and Artistic Works ("BC"), which prohibits subjecting the enjoyment and exercise of copyright to mandatory formalities, such as registration requirements;[51]

- *public institutions*: impulses for the further development of the copyright data infrastructure may also arise from non-legislative initiatives taken by national, European or international public bodies. The 2008 GRD working group, for instance, came together under the auspices of former Commissioner Neelie Kroes. WIPO initiated the aforementioned 2011 project for the establishment of a common rights database and has embarked on surveys on voluntary registration systems for copyright and related rights in 2005, 2010 and 2020;[52]

[49] Cf. Bodó, B., Gervais, D., & Quintais, J. P. (2018). Blockchain and smart contracts: the missing link in copyright licensing?. *International Journal of Law and Information Technology*, *26*(4), 311-336.
[50] Directive 2014/26/EU of the European Parliament and of the Council of 26 February 2014 on collective management of copyright and related rights and multi-territorial licensing of rights in musical works for online use in the internal market [2014] OJ L 84/72.
[51] For an in-depth analysis of the impact of this international ban on formalities, see Stef van Gompel, *Formalities in Copyright Law: An Analysis of Their History, Rationales and Possible Future*, Alphen aan den Rijn: Kluwer Law International 2011.
[52] The results of the WIPO surveys carried out in 2005 and 2010 can be found at: https://www.wipo.int/meetings/en/doc_details.jsp?doc_id=52829 and https://www.wipo.int/copyright/en/registration/registration_and_deposit_system_03_10.html. The 2020 survey has been announced, but the results have not yet been published. See: https://www.wipo.int/copyright/en/registration/index.html.

- *private entities*: the initiatives that have led to TOWGE, the Concertify platform and SOZA's Listen Local platform show that private entities, in particular CMOs, may play a decisive role in the further harmonisation and combination of copyright-related data. In addition, individual companies, such as Apple and Spotify, may obtain a market position that allows them to bring together an unprecedented volume of data and establish de facto data standards with a major impact on the sector. External technology start-ups also invest heavily in solutions based on blockchain or related technologies.

For the analysis of copyright data management issues, it is important to bear these different settings in mind. To arrive at a substantial improvement of the copyright data infrastructure, it may be necessary to combine public and private initiatives and seek to offer both legislative and market incentives. The legislation-made MLC initiative in the US, for instance, relies on Apple and Spotify as central sponsors and data providers. A similar, large-scale public/private partnership may be necessary to allow European creative industries to compete at eye level with data and licensing improvement on the other side of the Atlantic.

3.5 Sector-Specific Stumbling Blocks

For the success of European initiatives, however, it is also important to consider potential stumbling blocks and corrosive dynamics which large-scale data improvement projects may unleash in the creative industry sector:

- *rivalry between small and big players*: the establishment of overarching, comprehensive data infrastructures and licensing hubs in the music industry may be perceived as a threat by small players and repertoire holders. For example, small European CMOs may fear to be left behind[53] when major European CMOs take joint initiatives and organise data and licensing processes in a way that enhances the visibility and availability of their content – potentially at the expense of repertoire administered by other CMOs which do not have comparable tools to enhance content visibility and availability.[54] At the global level, individual companies with considerable market power, such as Apple, Spotify, YouTube and Netflix, may establish individual data standards that require European right holders to deal with different data systems for the purposes of distributing content and monitoring the volume of use. European artists and music distributors may also fear to be left behind and lose visibility and market shares on the world market after US legislation, as explained above, established a new US licensing hub in collaboration with US-based streaming services that may become a central data resource in the sector while lending insufficient weight to foreign repertoire;

[53] The risk of a "de facto copyright register in the hands of dominant platforms" was also identified by Germany, in its statement accompanying the Council vote on the DSM Directive. See S.F. Schwemer, "Article 17 at the Intersection of EU Copyright Law and Platform Regulation", Nordic Intellectual Property Law Review 2020, 400-435.

[54] Cf. Lucie Guibault and Stef van Gompel, "Collective Management in the European Union", in: Daniel Gervais (ed.), *Collective Management of Copyright and Related Rights*, 3rd ed., Alphen aan den Rijn: Kluwer Law International 2015, 139 (172).

- *fear of losing traditional gatekeeper position*: in sectors with a less developed data infrastructure, such as the field of visual arts, traditional content gatekeepers – holders of individual work libraries, including CMOs – may feel uneasy about initiatives to systematically attach metadata to copyrighted content and include resulting data sources in a comprehensive database and licensing infrastructure. Once a comprehensive and authoritative platform for rights clearance is in place, traditional "middlemen" in the rights clearance process may fear to become obsolete. The creation of non-harmonised and non-interoperable coding systems and data silos may be part of a survival strategy seeking to preserve a position on the content market, which a more efficient, overarching system for copyright data management may put at risk.

- *path dependence*: stakeholders are likely to have invested substantially in their own proprietary, and often incompatible (meta-)data systems. This investment in individual data infrastructures causes considerable switching costs in case an overarching, harmonised standard is set. This provides a strong disincentive to support initiatives to establish a common, harmonised data standard that requires changes to pre-existing individual data management systems.

This outline of problems arising from data harmonisation and improvement projects sheds light on central obstacles to the establishment of integrated data spaces which the European Commission also highlighted in its *European Strategy for Data*.[55] In this Communication, the Commission referred not only to insufficient data quality and interoperability as problem drivers but also to imbalances in market power, a lack of trust and insufficient economic incentives as obstacles to initiatives seeking to ameliorate and finally overcome the problematic status quo.[56]

4. Costs and Benefits

Considering difficulties and obstacles, it becomes apparent that the improvement of the copyright data infrastructure in the EU is not an easy task. As a highly complex endeavour, it can hardly be accomplished without substantial investment in metadata creation and improvement, technical data management infrastructure, and harmonisation initiatives. The foregoing analysis already offers first insights into the costs which an initiative to improve copyright data may entail in different creative industry sectors.

4.1 Considerable Investment Necessary

With regard to the overall costs of setting up and maintaining a comprehensive copyright data management system, the aforementioned music industry examples provide some indications. Reportedly, the European GRD initiative that had commenced in 2008, finally collapsed after an investment of £8 million because the CMOs involved could no longer agree on the funding of the project.[57] The MLC project in the US rests on a start-up investment of $33.5 million.[58]

---

[55] European Commission, supra note 1.
[56] European Commission, supra note 1, 7-8.
[57] See https://completemusicupdate.com/article/prs-confirms-global-repertoire-database-cannot-move-forward-pledges-to-find-alternative-ways/.
[58] See https://www.appleworld.today/blog/2019/11/18/apple-spotify-to-fund-new-music-royalties-collective.

After the start-up phase, MLC expenditures are expected to average $30 million annually and amount to $227 million from 2021 to 2028.[59]

According to these figures, there might be a substantive gap between the investment which interested parties in the EU, such as CMOs, are willing to make, and the budget that would be necessary to establish a comprehensive data infrastructure and, if this is desired, run a licensing hub. Before leaning too heavily on cost estimates made in a US context, however, it is important to note that MLC calculations were based on data input from only two central sources: iTunes and Spotify. Given the cultural diversity and wide variety of copyright data sources in the EU, a European data integration project (not relying exclusively on US-based Apple and Spotify data) would probably require an even larger investment in the start-up phase and following years.

Looking at the visual arts sector, an additional cost dilemma comes to the fore: the individual costs to be made in respect of each individual content item. In the field of photography, for instance, databases would have to contain an extremely high number of works. In many cases, these works will have a relatively low average licensing value. This constellation raises the problem that, even if a harmonised data format and a central data recording system become available, the required investment in metadata entry and maintenance may still not come forward because the revenue accruing from visibility and "findability" in the comprehensive database can hardly be expected to outweigh the costs of data entry. The expected market value does not justify the time and money that would have to be spent for each individual content item. Hence, the mere existence of a comprehensive and authoritative data infrastructure in a given sector does not automatically ensure that all right holders provide the data necessary to maintain data accuracy and completeness. Revisiting the potential discrepancy between the interests of small and big players, it can be added that, in the light of economies of scale, continuous data entry and maintenance may be less burdensome for holders of big work libraries. For instance, it is conceivable that holders of big repertoires are able to switch from manual data entry to the use of automated or machine-learning systems which substantially reduce the cost per unit.

Finally, it is to be noted that "costs" can also be understood in a broader sense. Instead of confining the analysis to monetary aspects, it is important to consider broader cultural repercussions, in particular the impact of standardised data formats and comprehensive copyright data systems on *cultural diversity*, recognition and attribution (in the sense of the moral rights enjoying protection under international copyright law and the national copyright systems of EU Member States) and the visibility and availability of the full spectrum of European creative works. In the case of photography, for instance, the commercial value of a work for right holders, as explained, will often be smaller than the cost of documenting the work – the outlined problem scenario that raises concerns about large economies of scale favouring large repertoire owners who can automate the documentation and indexation process. Considering this problem scenario, it becomes apparent that the burden of documenting and promoting content in large, supranational content repositories should not increase data management burdens to such an extent that it becomes unprofitable for smaller entities to comply with data standards and data entry requirements. Otherwise, the measures

[59] U.S. Congressional Budget Office Cost Estimate, S. 2823 – Music Modernization Act, as reported by the Senate Committee on the Judiciary on 12 September 2018 (revised version of 17 September 2018), p. 3, available at https://www.cbo.gov/system/files/2018-09/s2823.pdf.

taken to improve copyright data management may discriminate against holders of small repertoires – and potentially even against smaller national repertoires in the EU – and reduce the cultural diversity which the improved data system is intended to reflect.

4.2 Benefits Accruing from Improved Copyright Data

Benefits that can be expected to flow from an improved data management infrastructure are enhanced licensing opportunities, more efficient enforcement of rights, the reduction of royalty losses and the enhancement of access of high-tech industries to copyright data. Conversely, missing or inaccurate copyright metadata can lead to various types of welfare losses:

a. a work is not found and therefore not licensed. That is, the licensing transaction does not take place, depriving both right holders and consumers of the potential welfare gains (producer surplus and consumer surplus) which a transaction would generate in the counterfactual of accurate metadata;

b. a work is found or the potential licensee is aware of the work, but information to license is missing. This may result in two outcomes:

    i. the work is not used/consumed, as under (a);

    ii. the work is pirated/used without a license. In this case, all welfare effects of the transaction are generated on the demand side, while right holders do not benefit;

c. The work is found and licensed, but no proper remuneration is provided to right holders as a consequence of the inaccurate metadata, i.e., licensing revenues are collected but do not reach the right holders due to metadata issues.

Missed licencing and remuneration opportunities not only entail so-called static welfare losses; there can be dynamic effects as well. Efficient licensing can enable more creators to draw on existing copyrighted works, reducing the costs of follow-on creativity. Secondly, smaller markets for copyrighted works and greater costs of licensing will entail lower incentives to invest in innovative complementary goods and services (e.g. innovative ways of disseminating copyrighted works online, or innovative recommendation systems). Thirdly, high transaction costs, legal uncertainty, competition from unlawful competitors, market concentration and barriers to entry that result from (the requirement to incur) sunk costs can inhibit innovation. Efficient licensing systems – including metadata – can mitigate these issues. An obvious remedy, thus, would be to correct and complete the metadata.

In addition, the aforementioned cultural dimension must be taken into account – in the sense of benefits accruing from better visibility and availability of European cultural productions on the world market. To the extent to which European creative industries do not have their own comprehensive repertoire database, they depend on the configuration of content recommendation and licensing systems developed elsewhere. This entails the risk of

insufficient influence on the promotion, sales and distribution process.[60] In theory, the repertoire databases of iTunes, Spotify, YouTube or Deezer, for instance, may offer all providers of cultural content similar opportunities to reach out to end consumers. In practice, however, the visibility and success of a work will depend on the way in which these providers organise work-and creator-related (meta-)data and generate recommendations for end consumers. This implies that European content producers depend heavily on metadata and recommendation systems that have been developed by powerful individual companies. In the field of music, the MLC initiative that follows from US legislation may strengthen this trend. As the MLC database has been established with a focus on the US market and in collaboration with Apple and Spotify, European content is unlikely to occupy the centre stage. A further risk arises from the diversity of European content in terms of cultural backgrounds and languages. Descriptive metadata is usually connected with natural languages. However, the costs of documenting in smaller European languages relative to the expected sales value can be significantly higher for language groups with fewer potential buyers. This creates an incentive to replace higher cost-to-market repertoires from smaller language groups with (translations of) lower cost-to-market repertoires from large language groups, such as works for English-speaking audiences.[61]

5. Conclusion

In sum, the foregoing discussion of a potential need to improve copyright data indicates that, with a view to visibility on the world market and promising licensing opportunities resulting from new technologies, it seems desirable to arrive at a comprehensive database with a focus on European content, including smaller and less-known repertoires reflecting the full cultural diversity across EU Member States. The added value of an improved copyright data infrastructure for European creative industries – in the sense of enhanced visibility and accessibility at a global scale – is a core argument in the cost/benefit analysis that can tip the scales in favour of new efforts to create and harmonise metadata. An improved copyright data infrastructure is likely to enhance licensing, enforcement and royalty opportunities for creative industries. At the same time, it will provide developers of new technologies, such as AI system developers, broad access to diverse data resources. As a counterweight to initiatives in other regions, such as the MLC in the US, it can be expected to allow European creative industries to innovate and emancipate themselves from other data infrastructures and related content distribution and recommendation systems. It may also prevent a non-European bias in globally dominant AI systems trained on copyright data.

The foregoing discussion, however, also reflects the considerable obstacles on the way to more comprehensive and accurate European copyright (meta-)data. In addition to substantial financial resources that will be necessary, a key to new and successful initiatives lies in the creation of appropriate incentives for the creative industries, providers of digital content

[60] As to existing legislation seeking to enhance the visibility and prominence of European content, see Article 13(1) of the Audiovisual Media Services Directive 2010/13/EC, as amended by Directive 2018/1808/EU.

[61] Cf. Daniel Antal, Amelia Fletcher and Peter L. Ormosi, Music Streaming: Is It A Level Playing Field?, in: Competition Policy International, forthcoming 2021.

distribution services and high-tech companies in the field of AI to jointly develop solutions. For a trade-off across these industry sectors, the analysis provides several starting points.

Content distribution platforms and AI companies may have a particular interest in rules that make copyright enforceability and remuneration obligations conditional on the provision of metadata in a specific, interoperable format. With regard to TDM, Article 4(3) DSMD already points in this direction when it refers to the reservation of copyright "in an appropriate manner, such as machine-readable means…" The provision reflects the desirability of efficient interfaces between copyright data and TDM systems that allow the automated verification of use permissions. The requirement of providing "relevant and necessary information" for the blocking of infringing UGC in Article 17(4)(b) DSMD also offers room for establishing an obligation to provide work-related data in a standardised and interoperable format.[62] Arguably, information on protected literary and artistic creations is only "relevant" in the sense of Article 17(4)(b) when it is provided in a form that allows content moderation systems to read it. At the core of these considerations lies the more general principle that rights must be clearly drawn to be enforceable. In this vein, it can be posited that right holders must provide interoperable, accessible information to benefit from enhanced enforcement opportunities.

To strike a proper balance, however, it is necessary to consider not only the interoperability interest of platform and high-tech industries but also the interests of the creative industries. Ideally, producers and publishers of literary and artistic works would opt for the development of interoperable metadata collections voluntarily. In the light of the prohibition of formalities in Article 5(2) of the Berne Convention, one may even wonder whether statutory obligations to comply with a specific data format are possible at all.[63] Hence, the question arises which *quid pro quo* could be developed to ensure acceptance of data standardisation and interoperability obligations in the creative industries.

The interest in transparency and accountability of content moderation and recommendation systems could play a central role in this respect. As a countermove to compliance with harmonised and interoperable data formats, content distribution platforms and AI companies could be obliged to open the "black box" of their algorithmic tools. As a result, the creative industries could benefit from transparency with regard to content selection, moderation and recommendation processes in automated systems.

Hence, new approaches in the area of copyright data improvement could potentially evolve from a trade-off addressing interoperability and transparency interests.[64] On the one hand, the

---

[62] As to the use of the requirement of "relevant and necessary information" as a tool to promote specific notification standards, see Martin R.F. Senftleben/Christina Angelopoulos, *The Odyssey of the Prohibition on General Monitoring Obligations on the Way to the Digital Services Act: Between Article 15 of the E-Commerce Directive and Article 17 of the Directive on Copyright in the Digital Single Market*, Amsterdam: Institute for Information Law/Cambridge: Centre for Intellectual Property and Information Law 2020, 31.

[63] See 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, arguing that such obligations would not fall afoul of the Berne prohibition on formalities, as long as they do not function as prerequisites for the coming into being, maintenance or enforcement of copyright.

[64] For current legislative initiatives pointing in this direction, see European Commission, 15 December 2020, "Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and Amending Directive 2000/31/EC", Document COM(2020) 825 final, Articles 13 and 23 (addressing content moderation) and Recital 52 and Article 30 (addressing

interest of online content distributors and AI trainers in standardised and interoperable data formats could be recognised. On the other hand, transparency and accountability in respect of algorithmic content selection, moderation and recommendation systems should be ensured to pave the way for the eradication of systems that may disadvantage small and less-known enterprises and repertoires, or creators with specific racial, ethnic or other minority backgrounds. To make this incentive scheme for collaboration attractive to a broad spectrum of copyright holders, further research is necessary to develop appropriate solutions not only for big companies but also for independent labels and other SMEs in the creative industries. In addition, it remains an open question whether the prospect of enhanced collaboration in the area of interoperability and transparency would also be sufficient to convince central gatekeepers, in particular CMOs, to contribute to fully standardised and interoperable copyright metadata. As pointed out above, the fear of losing their exclusive position in controlling relationships with their members may trigger resistance against injecting data into a fully standardised copyright data system.

Available at SSRN: https://ssrn.com/abstract=3785272

targeted advertising), available at: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en.