GRUR International, 74(4), 2025, 307–319

https://doi.org/10.1093/grurint/ikaf029
Advance Access Publication Date: 10 March 2025
Article

KACPER SZKALEJ*

The Paradox of Lawful Text and Data Mining? Some Experiences from the Research Sector and Where We (Should) Go from Here

Scientific research can be tricky business. This article critically explores the 'lawful access' requirement in European copyright law which applies to text and data mining (TDM) carried out for the purpose of scientific research. Whereas TDM is essential for data analysis, artificial intelligence (AI) and innovation, the paper argues that the 'lawful access' requirement in Art. 3 CDSM Directive may actually restrict research by complicating the applicability of the TDM provision or even rendering it inoperable. Although the requirement is intended to ensure that researchers act in good faith before deploying TMD tools for purposes such as machine learning, it forces them to ask for permission to access data, for example by taking out a subscription to a service. That provides the opportunity for copyright holders to apply all sorts of commercial strategies to set the legal and technological parameters of access and potentially even circumvent the mandatory character of the provision. The paper concludes by drawing on insights from the recent European Commission study 'Improving access to and reuse of research results, publications and data for scientific purposes' that offer essential perspectives for the future of TDM, and by suggesting a number of paths forward that EU Member States could take now to support a more predictable and reliable legal regime for scientific TDM, and potentially code mining, to foster innovation.

I. Introduction

The balance between copyright protection and public access to knowledge is more relevant than ever, especially in research and technological development. One area where this balance has become particularly tricky is the intricate realm of text and data mining (TDM). In this contribution I would like to consider an issue that rarely gets attention in copyright scholarship, namely the interaction between permissible use through a copyright exception and the ostensible need to ask for permission from the rightholder *as a prerequisite* for the applicability of the exception.[1] It is a rare type of situation but it does exist in a few cases, and the new TDM provisions introduced by Arts. 3 and 4 CDSM Directive[2] are no exception. Both provisions are conditioned on lawful access to the copy. My argument is that in the digital environment, especially with online uses, lawful access requirements lead to a paradox – they are meant to ensure that the user's access is bona fide, but in doing so they risk rendering the exception inoperable.

It is useful to start with an exploration of how the discourse around copyright exceptions has evolved in European copyright law. I will then delve into these requirements, focusing primarily on the exception permitting TDM for scientific research, and end with some insights from a recent study and how we should move forward.

II. What copyright exceptions do and how they are supposed to be handled

Together with the rules for subsistence of copyright protection, such as originality,[3] the time-limited duration of copyright,[4] or the principle of exhaustion,[5] copyright exceptions constitute tools that mediate between different

* Dr (LL.D.), Researcher in Intellectual Property Law, Institute for Information Law (IViR), University of Amsterdam, The Netherlands. E-mail: k.szkalej@uva.nl. ORCID: 0000-0002-7834-5514. This article is based on a talk prepared for the conference 'Copyright, TDM and AI', 14 June 2024 in Stockholm. I am grateful to the organisers for the invitation.

1 This issue receives increasingly more attention. For example Tatiana Eleni Synodinou, 'Lawfulness for Users in European Copyright Law: Acquis and Perspectives' (2019) 10(1) JIPITEC 20 and, more evocatively, Thomas Margoni, 'Saving Research: Lawful Access to Unlawful Sources' (*Kluwer Copyright Blog*, 22 December 2023), <https://copyrightblog.kluweriplaw.com/2023/12/22/saving-research-lawful-access-to-unlawful-sources-under-art-3-cdsm-directive/> accessed 10 February 2025.

2 Directive 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.

3 Harmonised by the Court of Justice of the European Union (CJEU). See further Eleonora Rosati, *Originality in EU Copyright Law: Full Harmonization through Case Law* (Edward Elgar 2013).

4 Harmonised initially by Council Directive 93/98/EEC of 29 October 1993, then by the current Directive 2006/116/EC of the European Parliament and of the Council of 12 December 2006 on the term of protection of copyright and certain related rights, which was amended in 2011 by Directive 2011/77/EU. At international level the term continues to be harmonised by art 7 Berne Convention, art 14 Rome Convention, art 17 WIPO Performances and Phonograms Treaty. The WIPO Copyright Treaty incorporates the Berne provision through art 1(4).

5 Codified in art 4(2) 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 (InfoSoc Directive), art 4(2) Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs (Software Directive).

interests. To realise this objective, they come in different shapes and sizes. For instance, some are conditioned on a more or less complex set of cumulative requirements, such as respectively the temporary copying exception,[6] or the news reporting exception.[7] Some are conspicuously concise, like the parody exception,[8] while others pay for themselves, like the public lending exception that permits public libraries to lend on condition that rightholders receive 'fair remuneration',[9] or the private copying exception that permits such copying on condition that rightholders receive 'fair compensation'.[10] Some exceptions also extend over such esoteric uses as the making of *ephemeral* copies by broadcasting organisations or use during religious or official celebrations.[11]

It is self-evident from the plethora of uses that they permit and the beneficiaries that they apply to that copyright exceptions are intended to realise the cultural or innovation policy of a state, regulate industry practice, alleviate market failure, or safeguard the protection of constitutional rights.[12] Indeed, one does not have to be a copyright or constitutional lawyer to realise that copyright exceptions evidently give expression to constitutional rights, particularly freedom of expression and access to information, the right to privacy, freedom of the arts and sciences, freedom to conduct a business, or even the right to fair trial.[13] Despite this, European copyright law, especially the InfoSoc Directive which contains the main catalogue of exceptions, does not make these values or objectives particularly explicit, nor has it been reflected in the way exceptions have been treated.

From recital 31 of the Directive we learn that harmonisation of copyright *must* safeguard a fair balance of rights and interests of both rightholders and users. But what rights and interests are at stake? With the exception of an ambiguous reference to freedom of expression in recital 3 and to education and teaching in recital 14, the preamble of the Directive is virtually silent on the normative aspects of exceptions or interests of users compared with the attention it devotes copyright protection.[14] The question that largely remains unanswered in the Directive is against what, in particular, the need to ensure a high level of protection (as follows from recital 9) should be balanced so that a 'fair balance' can actually be safeguarded. Instead, we learn in recital 32 that the list of 22 copyright exceptions in the Directive is exhaustive and that it takes account of different traditions of the Member States. In the absence of uniform values, or at least an indication of what exceptions are supposed to do, it seems Member States have not only been able to cherry-pick the exceptions (as follows from Art. 5 of the Directive) but also their justifications and rationales.[15]

Beyond the foibles of the text of the Directive, the CJEU's case law on exceptions did not make recital 31 any clearer. In cases such as *Infopaq* from 2009 and the later order in *Infopaq II* not only did the CJEU not mention it but also established and entrenched the idea that copyright exceptions constitute a derogation from a general principle established by the directive – that principle being copyright protection – and had therefore to be interpreted strictly;[16] a formula understood to imply the selection of that interpretation which, in relation to all other possible meanings of an interpreted term, has the strictest scope of application.[17] Such an approach has been reflected at national level.[18] Coupled with the oft-repeated objective of providing a high level of protection and the formula of broad interpretation of exclusive rights,[19] such a rule of interpretation invited to a normative categorisation of copyright exceptions as mere

6 InfoSoc Directive, art 5(1).

7 InfoSoc Directive, art 5(3)(d).

8 InfoSoc Directive, art 5(3)(k).

9 Directive 2006/115/EC of the European Parliament and of the Council of 12 December 2006 on rental right and lending right and on certain rights related to copyright in the field of intellectual property (RLD), art 6(1).

10 InfoSoc Directive, art 5(2)(b). Other examples include arts 5(2)(a) on reprography or 5(2)(e) InfoSoc Directive on uses by social institutions pursuing non-commercial goals. Despite concrete itemisations, at least in the case of the InfoSoc Directive Member States may require compensation for any of the exceptions as clarified by recital 36. Hence, if it aligns with the copyright policy and balance sought in a Member State, all exceptions can pay for themselves.

11 InfoSoc Directive, arts 5(2)(d) and 5(3)(g) respectively.

12 See generally Lucie Guibault, *Copyright Limitations and Contracts. An Analysis of the Contractual Overridablity of Limitations on Copyright* (Kluwer Law International 2001) 28-87. A lot has been written about exceptions in recent years. See especially Tito Rendas, *Exceptions in EU Copyright Law. In Search of a Balance Between Flexibility and Legal Certainty* (Kluwer Law International 2021) and Kacper Szkalej, *Copyright in the Age of Access to Legal Digital Content. A Study of EU Copyright Law in the Context of Consumptive Use of Protected Content* (Uppsala University 2021).

13 See in case of fair trial art 5(3)(e) InfoSoc Directive. In such cases the definition of exclusive rights has also relevance, such as where holders of copyright decide to initiate infringement proceedings against someone submitting copyright protected material as evidence in another case against the rightholder; see C-637/19 *BY v CX* ECLI:EU:C:2020:863 and Kacper Szkalej, 'Looking for the edge of Article 3 InfoSoc Directive and finding it twice – in a car and in the court' (*Kluwer Copyright Blog*, 25 November 2020) <https://copyrightblog.kluweriplaw.com/2020/11/25/looking-for-the-edge-of-article-3-infosoc-directive-and-finding-it-twice-in-a-car-and-in-the-court/> accessed 10 February 2025.

14 cf InfoSoc Directive, recitals 4 and 9 relating to high level of protection and 21 and 23 relating to a broad interpretation of rights. According to Thomas Dreier, 'Limitations: The Centerpiece of Copyright in Distress: An Introduction' (2010) 1(2) JIPITEC 50, 51, 'nowhere has [the aim to provide for as much protection as possible] been formulated more clearly as in the recitals of the InfoSoc Directive'. See further Szkalej, *Copyright in the Age of Access to Legal Digital Content. A Study of EU Copyright Law in the Context of Consumptive Use of Protected Content* (n 12) 203-05.

15 For more on cherry-picking, see Lucie Guibault, 'Why Cherry-Picking Never Leads to Harmonisation: The Case of the Limitations on Copyright under Directive 2001/29/EC' (2010) 1(2) JIPITEC 55.

16 Case C-5/08 *Infopaq International v Danske Dagblades Forening* (*Infopaq*) ECLI:EU:C:2009:465, paras 56-57; Case C-302/10 *Infopaq International v Danske Dagblades Forening* (*Infopaq II*) ECLI:EU:C:2012:16, para 27. The rule of strict interpretation is sometimes denoted by alternative terms such as *narrow* or *restrictive* interpretation.

17 Tito Rendas, 'Copyright, Technology and the CJEU: An Empirical Study' (2018) 42(2) IIC 153, 157.

18 Such as in Sweden; Regeringens proposition 2004/05:110, 83.

19 For example Case C-607/11 *ITV Broadcasting and Others v TVCatchup* ECLI:EU:C:2013:147, para 20; Case C-466/12 *Nils Svensson and Others v Retriever Sverige* ECLI:EU:C:2014:76, para 17; Case C-351/12 *OSA v Léčebné lázne Mariánské Lázne* ECLI:EU:C:2014:110, para 23; Case C-325/14 *SBS Belgium* v *Belgische Vereniging van Auteurs, Componisten en Uitgevers* ECLI:EU:C:2015:764, para 14; Case C-160/15 *GS Media BV v Sanoma Media Netherlands and Others* ECLI:EU:C:2016:644, para 30; Case C-527/15 *Stichting Brein* v *Filmspeler* ECLI:EU:C:2017:300, para 27; Case C-610/15 *Stichting Brein v Ziggo BV and XS4ALL Internet* ECLI:EU:C:2017:456, para 22; Case C-161/17 *Land Nordrhein-Westfalen v Dirk Renckhoff* ECLI:EU:C:2018:634, para 18; Case C-484/18 *Spediddam and Others v Institut national de l'audiovisuel* ECLI:EU:C:2019:970, para 36 with 39; and Case C-263/18 *Nederlands Uitgeversverbond and Groep Algemene Uitgevers v Tom Kabinet Internet and Others* ECLI:EU:C:2019:1111, paras 48-49. In Case C-145/10 *Eva-Maria Painer v Standard Verlags and Others* ECLI:EU:C:2013:138 the two components were mentioned

subsidiary provisions regulating an abnormality or, more evocatively, as islands in a sea of exclusivity.[20]

Nevertheless, what we now see in copyright discourse (and legislation) is a clear shift of perspective. When the CJEU was again confronted with the complicated Art. 5(1) InfoSoc Directive after *Infopaq* in *FAPL*, involving the showing of football matches from a legal source via satellite on television screens, the Court realised that strict adherence to the rule of strict interpretation of exceptions is simply not a workable approach. According to the Court, 'the interpretation of [an exception] (…) must enable the effectiveness of the exception (…) to be safeguarded and permit observance of the exception's purpose as resulting in particular from recital 31 [InfoSoc Directive]'.[21] Such a rule requires the interpreter to ask the question why the exception is there in the first place. That's not only a broadening of the approach in the direction of discovering the interest that lies behind an exception, but also a clear invitation to a purposive interpretation of exceptions. In *FAPL* the CJEU confirmed this by stating that Art. 5(1) 'must [in accordance with its objective] allow and ensure the development and operation of new technologies and safeguard a fair balance between the rights and interests of right holders, on the one hand, and of users of protected works who wish to avail themselves of those new technologies, on the other'.[22]

Libraries are another great example where change is taking place. Just like operators of TV screens, they are considered users in the copyright ecosystem and therefore need to organise their activities in a way that is compatible with the conditions contained in various copyright exceptions. That's not only the obvious act of lending but also archiving, cultural preservation, facilitation of research and even carrying it out.[23] But why does copyright law permit public libraries to lend books that people can borrow for free? Is it because lending constitutes a means for others to impart and receive information and that promotes the development of a democratic society and free formulation of opinions? This is actually not that different from shadow libraries like SciHub and LibGen or The Pirate Bay platform.[24] Is it to promote literature and society's interest for self-development, enlightenment, research, and culture? Is it purely mechanical so that libraries can fulfil their statutory tasks? Or is it all of these at the same time?[25] Against this backdrop, when libraries are confronted with 21st century consumption patterns, under a regime of purposive interpretation of copyright exceptions, there may be every reason to extend the scope of the lending exception to cover e-lending and that is exactly what the CJEU has done.[26]

The same goes for research. There are four different exceptions permitting scientific research – in Arts. 5(3)(a) InfoSoc Directive, 6(2)(b) and 9(b) Database Directive and 10(1)(d) Rental and Lending Directive. Why are they there? To promote the progress of science and the useful arts?[27] To make Europe a leader in innovation? To realise the principle of academic freedom? If we take science seriously, we need to recognise that researchers must be able to choose not only the topic of research and questions to be answered, but also the methods and the materials to find those answers and present and disseminate the results.[28] However, according to the CJEU, academic freedom has not only this individual dimension but an institutional and organisational dimension reflected in the autonomy of institutions.[29] In this context, universities, libraries and other research institutions emerge as critical facilitators of scientific research. They provide the infrastructure, access to resources, and the organisational freedom necessary for researchers to pursue their inquiries. Acting as custodians of science, these institutions preserve and make available the vast body of knowledge needed for the pursuit of innovation. By safeguarding academic freedom both for individuals and as part of their institutional mission, universities and libraries play an essential role in fostering a research environment that nurtures curiosity and creativity, encourages diversity in methods, and ensures that scientific progress remains unhindered by restrictive policies or commercial interests. With few

separately (paras 96 and 107); similarly Case C-469/17 *Funke Medien NRW v Bundesrepublik Deutschland* ECLI:EU:C:2019:623, paras 50 and 70 and Case C-516/17 *Spiegel Online v Volker Beck* ECLI:EU:C:2019:625, paras 35 and 54. The canon of broad interpretation has also been invoked without direct reference to the objective of high level of protection; for example in Case C-301/15 *Soulier and Doke v Premier ministre and Ministre de la Culture et de la Communication* ECLI:EU:C:2016:878, para 30 (right of reproduction), Case C-117/15 *Reha Training Gesellschaft für Sport-und Unfallrehabilitation v GEMA* ECLI:EU:C:2016:379, para 36 (communication to the public) or Case C-265/16 *VCAST v RTI* ECLI:EU:C:2017:913, para 40 (communication to the public). See further Rendas, *Exceptions in EU Copyright Law. In Search of a Balance Between Flexibility and Legal Certainty* (n 12) 55-58.

20 Christophe Geiger, 'Promoting Creativity through Copyright Limitations: Reflections on the Concept of Exclusivity in Copyright Law' (2010) 12(3) Vand.L.Rev. 515, 521, for whom such a hierarchy is already embedded in the term 'exception'; cf Rendas, *Exceptions in EU Copyright Law. In Search of a Balance Between Flexibility and Legal Certainty* (n 12) 70.

21 Joined Cases C-403/08 and C-429/08 *FAPL v QC Leisure and Others* ECLI:EU:C:2011:631, para 163.

22 ibid para 164.

23 See Ángel Borrego, Jordi Arrdanuy and Cristobal Urbano, 'Librarians as Research Partners: Their Contribution to the Scholarly Endeavour Beyond Library and Information Science' (2018) 44(5) JAL 663, and Rebecca Watson-Boone, 'Academic Librarians as Practitioner-Researcher' (2000) 26(2) JAL 8.

24 cf *Neij and Kolmissoppi (TPB) v Sweden* ECtHR App No 40397/12, according to which TPB constituted the means for others to impart and receive information and those actions were considered to be afforded protection under Art. 10 European Convention on Human Rights (ECHR) (freedom of expression) by default. One can have different opinions about TPB but an often missed circumstance in the debate is that in the national case preceding the application to the ECtHR, and which concerned contributory infringement of copyright, the Svea Court of Appeal stated that it had not been shown what proportion of torrent files on the platform led to material made available *with* rightholders' consent, and that the court therefore had to assume that that amount was not negligible. Moreover, expert witnesses called in the case provided drastically opposing opinions on the proportion; Svea HovR deldom 26 november 2010 mål B 4041-09 (*TPB-målet*), p 25.

25 See for example 2 and 12 §§ bibliotekslagen (2013:801) (Swedish Libraries Act).

26 Case C-174/15 *VOB v Stichting Leenrecht* ECLI:EU:C:2016:856. The 'EU rationale' of the lending exception seems to be contribution to cultural promotion, as follows from para 51 of the judgment.

27 I borrow this from the US copyright clause in art I, s 8, clause 8 US Constitution.

28 Vilius Stančiauskas and others, 'Improving access to and reuse of research results, publications and data for scientific purposes. Study to evaluate the effects of the EU copyright framework on research and the effects of potential interventions and to identify and present relevant provisions for research in EU data and digital legislation, with a focus on rights and obligations' (Study for European Commission, DG RTD, 2024) 153 <https://data.europa.eu/doi/10.2777/633395> accessed 10 February 2025 (Commission Study).

29 Case C-66/18 *Commission v Hungary* ECLI:EU:C:2020:792, paras 225 and 227.

strings attached – use for the purpose of research is permitted to the extent justified by the non-commercial purpose to be achieved and as long as the source and name of the author is indicated unless this turns out to be impossible[30] – the research exceptions have probably always been intended to be interpreted broadly.

The new formula of purposive interpretation has been repeated many times in subsequent case law,[31] until finally in 2019 the CJEU declared that copyright exceptions constitute *user rights*.[32] Whilst not necessarily evident from the rising amount of copyright cases at the time, this decade-long development was promulgated by significant constitutional shifts at EU level, and in particular the incorporation of the Charter of Fundamental Rights of the European Union (EU Charter) into primary EU legislation.[33] Fundamentally, the status of the EU Charter means that directives and any other secondary legislation must be interpreted in its light.[34] The CJEU's proclamation in the same cases (and one other) that copyright exceptions constitute mechanisms through which fundamental rights find concrete expression may therefore not be surprising to a constitutionalist.[35] It is simply a coherent adherence to the hierarchy of legal sources, and one that should probably have occurred much sooner, already in *Infopaq*. But for a copyright lawyer of our time, a qualification of exceptions as user rights is a very strong statement that signifies a shift has occurred since the adoption of the InfoSoc Directive.[36] Most importantly, the EU Charter patches the InfoSoc Directive by injecting a common frame of reference of familiar constitutional values into the structure of the harmonised copyright framework, and gives sense to recital 31 by providing a much stronger normative foundation for the rights and interests of copyright users.[37] This same shift has also been codified in provisions such as Art. 17(9) CDSM Directive, which requires Member States to ensure that users can enforce copyright exceptions against unjustified copyright claims on online content platforms.[38] In Member States such as Sweden it has even led to the introduction of an entirely new cause of action, enabling users to request an injunction against a content platform, or even damages.[39]

Through these developments copyright exceptions are no longer a derogation from a general rule. They are the rule once again, together with exclusive rights.

III. The act of text and data mining, big data, and AI

Although the list of exceptions was supposed to be exhaustive, the first thing the CDSM Directive does in its opening provisions is to introduce new ones. Amongst these are Arts. 3 and 4, which enable reproductions of works and extractions of databases[40] for the purpose of text and data mining.[41] What clearly lands in the foreground is the remarkable definition of TDM, denoting the use of 'any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations'.[42] Clearly the Directive is indifferent as to the type of information that is generated. It *can* be some form of patterns, trends, or correlations. It *can* also be information that eventually makes it possible for a computer to evoke the image of some unknown distant planet and distinguish it from a car exhaust pipe.

But the evident breadth of the definition should not be surprising at all. In the mid-2010s the central buzzword on everybody's mind was *big data* – extremely large datasets that may be analysed computationally to reveal patterns, trends and associations.[43] This was the time when Europe was half-way through the Digital Agenda,[44] the GDPR[45] was in the making, and policy discussions focused on the *digital economy* and datafication of society.[46] We still live in

30 The provision in the Rental and Lending Directive has no such requirements. Although the subject matter covered by this directive is also covered by the InfoSoc Directive and therefore the InfoSoc exception, I can envisage the argument that the former constitutes *lex specialis* relative the latter in view of art 1(2)(b)-(c) and because the current Rental and Lending Directive, Directive 2006/115/EC, is a recast of Council Directive 92/100 that was adopted five years after the InfoSoc Directive.

31 *Painer* (n 19) para 133; Case C-201/13 *Deckmyn and Vrijheidsfonds VZW v Vandersteen and Others* ECLI:EU:C:2014:2132, paras 23 and 27; Case C-117/13 *Technische Universität Darmstadt v Ulmer* ECLI:EU:C:2014:2196, paras 32 and 43; *Spiegel Online* (n 19) para 36; *Funke Medien* (n 19) para 51.

32 *Spiegel Online* (n 19) para 54; *Funke Medien* (n 19) para 70.

33 Treaty on European Union (TEU), art 6(1) which states that the EU Charter shall have the same legal value as the Treaties.

34 See also EU Charter, art 51(1).

35 *Spiegel Online* (n 19) para 43; *Funke Medien* (n 19) para 58; as well as Case C-476/17 *Pelham* ECLI:EU:C:2019:624, para 60.

36 Not since the adoption of the first copyright directive; see Council Directive 91/250/EEC of 14 May 1991 on the legal protection of computer programs, arts 5 and 6 and preamble (recital 26).

37 Szkalej, *Copyright in the Age of Access to Legal Digital Content. A Study of EU Copyright Law in the Context of Consumptive Use of Protected Content* (n 12) 205.

38 See further Tito Rendas, 'Are copyright-permitted uses 'exceptions', 'limitations' or 'user rights'? The special case of Article 17 CDSM Directive' (2022) 17(1) JIPLP 54.

39 6 b kap. 52 r – 52 s §§ Lag (1960:729) om upphovsrätt till litterära och konstnärliga verk (Swedish Copyright Act).

40 Protected through the Database Directive.

41 Others have also been added in the meantime in a piecemeal fashion through Directive 2012/28/EU of the European Parliament and of the Council of 25 October 2012 on certain permitted uses of orphan works (Orphan Works Directive), and Directive (EU) 2017/1564 of the European Parliament and of the Council of 13 September 2017 on certain permitted uses of certain works and other subject matter protected by copyright and related rights for the benefit of persons who are blind, visually impaired or otherwise print-disabled and amending Directive 2001/29/EC (Marrakesh Treaty Directive) implementing the Marrakesh Treaty of 2013 to Facilitate Access to Published Works for Persons Who Are Blind, Visually Impaired or Otherwise Print Disabled.

42 CDSM Directive, art 2(2).

43 See generally Daniel Gervais, 'Exploring the Interfaces Between Big Data and Intellectual Property Law' (2019) 10(1) JIPITEC 3, 3-5.

44 European Commission, 'A Digital Agenda for Europe' COM(2010)245 final, 19 May 2010.

45 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC.

46 See for example contributions in Sebastian Lohsee, Reiner Schulze and Dirk Staudenmayer (eds), *Trading Data in the Digital Economy: Legal Concepts and Tools. Münster Colloquia on EU Law and the Digital Economy III* (Nomos 2017); Article 29 Working Party, 'Statement of the WP29 on the impact of the development of big data on the protection of individuals with regard to the processing of their personal data in the EU' 14/EN WP 221 (16 September 2014) <https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp221_en.pdf> accessed 10 February 2025; Council of Europe, 'Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data' (Directorate General of Human Rights and Rule of Law, 23 January 2017) <https://rm.coe.int/16806ebe7a> accessed 10 February 2025; Cassidy R Sugimoto, Hamid R Ekbia and Michael Mattioli (eds), *Big Data is Not a Monolith* (MIT Press 2016); Bart van der Sloot and Sascha van Schendel, 'Ten Questions for Future Regulation of Big Data: A comparative and Empirical Legal Study' (2016) 7(2) JIPITEC 110.

this time and witness how the European legislator adopts all sorts of heavy-duty data-relevant legislation – the DGA,[47] DMA,[48] DSA,[49] and the Data Act[50] (which even has its own mini-TDM exception in Art. 43).[51] Not all of these take big data as their central theme, but they all relate to the use and management of data in digital ecosystems, and most realise what is now called the European Strategy for Data.[52] We have also grown enough in the meantime to not be impressed by the size of data anymore, which saves the need for something like a Big Data Act. That these provisions of the CDSM Directive are intended to spearhead big data analytics and realise the new technological frontier follows not only from the clear forward-oriented itemisations in recital 8 and 18, which recognise the capacity to process large amounts of data to gain new knowledge, discover new trends, or the wide use of TDM techniques by private and public entities in different areas of life, including for developing new applications and technologies. In a fashion that is completely unnecessary for copyright legislation, recital 9 clarifies that TDM can be carried out in relation to mere facts or data not protected by copyright, or that it may even not involve the reproduction right at all.

The sort of computational analysis involved in TDM is of course very relevant for machine learning processes and raises the question of whether the new exceptions apply to the development of AI systems and large language models. There is a recent argument that the legislature did not intend the TDM exceptions to cover it.[53] Of course, the CDSM Directive neither makes reference to machine learning nor to AI. But why should it? Both are terms of the 1950s.[54] Similarly, it does not make reference to big data (but instead to 'large amounts of data') or another buzzword of the mid-2010s – deep learning.[55] Although one popular interpretation of data mining is that it simply is machine learning,[56] to require legislators to positively itemise every conceivable situation they have mind is absurd. Especially when exceptions are user rights and mechanisms through which fundamental rights find concrete expression. To handle that stupendous task, exceptions must be given the space to do it and cannot be drafted like 'technical repair manuals'.[57] Good policy avoids codifying the flavour of the day and focuses on the general principle. The TDM exceptions express the principle that if someone works with data analysis using modern techniques, whatever the preponderant purpose, they should be able to use copyright-protected material 'in order to generate information'.

Today we have the AI Act[58] whose recitals 104-106 and Art. 53(1)(c) confirm that the TDM exceptions are relevant for machine learning and development of AI systems. This not only settles the question but seems to confirm that the TDM exceptions are *the* exceptions that are relevant for developing AI systems. But there are also others, like the temporary copying exception, which applies concurrently with the TDM provisions as clarified by recital 9 CDSM Directive, or the research exceptions mentioned earlier.

Since none of the research exceptions circumscribe scientific research as such, taken together one could therefore say that there has always been an opportunity to carry out TDM, and certainly when exceptions are to be given a purposive interpretation. What then do the new provisions add? Recitals 8, 10, 11 CDSM Directive make the case that the provision is intended to rectify the legal uncertainty that universities and other research institutions are confronted with. What exactly that uncertainty consists of is not self-evident from the recitals,[59] but certainly the discretionary character of all the general scientific research exceptions has made it possible for Member States to opt out of science or narrow the national provision down to teaching.[60] However, while the provision is said to rectify legal uncertainty, it also adds new complexity.

47 Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act).

48 Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act).

49 Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).

50 Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act).

51 'The sui generis right provided for in Article 7 of Directive 96/9/EC shall not apply when data is obtained from *or generated by* a connected product or related service falling within the scope of this Regulation, in particular in relation to Articles 4 and 5 thereof' (emphasis added).

52 European Commission, 'A European strategy for data' COM(2020) 66 final, 19 February 2020.

53 Tim W Dornis and Sebastian Stober, *Urheberrecht und Training generativer KI-Modelle* (Nomos 2024) report published 13 September 2024, at 94; cf Kacper Szkalej and Martin Senftleben, 'Mapping the Impact of Share Alike/CopyLeft Licensing on Machine Learning and Generative AI' (*Open Future*, 12 June 2024) <https://openfuture.eu/publication/the-impact-of-share-alike-copyleft-licensing-on-generative-ai/> accessed 10 February 2025.

54 John McCarthy and others, 'A Proposal For The Dartmouth Summer Research Project On Artificial Intelligence' (31 August 1955) <http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf> accessed 10 February 2025; Arthur L Samuel, 'Some Studies in Machine Learning Using the Game of Checkers' (1959) 3(3) IBM JR&D 210.

55 Cade Metz, '2016: The Year That Deep Learning Took Over the Internet' (*WIRED*, 25 December 2016) <https://www.wired.com/2016/12/2016-year-deep-learning-took-internet/> accessed 10 February 2025.

56 Rajkumar Buyya, Rodrigo N Calheiros and Amir Vahid Dastjerdi, *Big Data: Principles and Paradigms* (Morgan Kaufmann Publishers 2016) 14-15. Also Daniel Gervais, 'Exploring the Interfaces Between Big Data and Intellectual Property Law' (2019) 10(1) JIPITEC 3.

57 cf Bernt Hugenholtz, 'Exploitation and Liability' in Jan Rosén and Per Jonas Nordell (eds), *Copyright, Related Rights and Media Convergence in the Digital Context* (ALAI Nordic Study Days 18-20 June 2000) (Swedish Copyright Society 2000) 124.

58 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828.

59 In case of all other users, whose TDM is covered by art 4 CDSM Directive, the case is made in recital 18 that such users *could* be faced with legal uncertainty as to whether TDM can be carried out on lawfully accessed material when the reproductions or extractions made for the purposes of the technical process do not fulfil all the conditions of the temporary copying exception in art 5(1) InfoSoc Directive. The recital also explicitly states that the objective is to encourage innovation in the private sector. See however Jean-Paul Triaille, Jérôme de Meeus d'Argenteuil and Amélie de Francquen, 'Study on the legal framework of text and data mining (TDM)' (Study for the European Commission, DG MARKT, 2014).

60 This is partly recognized in recital 10. See however Commission Study (n 28) 148 and Annex I, 'State of harmonisation and its impact on Open Science' 189-90.

IV. Lawful text and data mining: Between a rock and a hard dataset

Article 3 CDSM Directive covers research organisations and cultural heritage institutions, including 'persons attached thereto' as explained in recital 14.[61] The provision allows TDM, for the purpose of scientific research, of copyright-protected material to which the research organisation has 'lawful access'. In other words, *before* a researcher mines data, they need to have lawful access to the material they want to mine. That is subject to one peculiar exception addressed at the end of this section.

1. Access granted? A lawful quandary in the mines of copyright

Lawful access requirements are rather unusual – the adopted, binding, wording of none of the exceptions in Art. 5 InfoSoc Directive refers to such a condition. In the Directive a similar requirement is instead imposed in the English version of Art. 6 ('legal access') to justify Member State intervention when DRM systems prevent reliance on certain exceptions. The whole idea behind such requirements is probably to ensure that users obtain copies from bona fide sources. But the more one ponders what that could mean, the more complicated it gets, given the few incarnations of this type of requirement in the copyright *acquis*. Apart from 'lawful access' in Art. 3 CDSM Directive and 'legal access' in Art. 6(4) InfoSoc Directive, there is the concept of 'lawful acquirer',[62] 'lawful user',[63] 'person having a right to use *the computer program*',[64] 'licensee',[65] 'person having the right to use *a copy* of the computer program',[66] 'person authorised to use a copy of the computer program on their behalf',[67] and in the case of Art. 4 CDSM Directive, the provision that allows anyone to TDM for any purpose, there is 'lawfully accessible'. All these different terms cause quite the headache – they all seem related, yet they are differently formulated and some can be understood to refer to completely different things, such as the bona fide character of the user, the status of the copy, or the status of the source from which the copy is obtained. Moreover, and in relation to the latter type, in two cases the CJEU seems to have introduced an actual lawful source requirement. Depending on how one understands it, the interpretation of the provision in question can be different each time and therefore so too can the scope of the exception.[68] Within the CDSM Directive alone there is the question of whether 'lawful access' to works and 'lawfully accessible' works convey the same requirement. In law, grammar and the selection of terms matters, so if we treat legislative prerogative seriously the answer can only be no.

Neither the provisions nor recitals of the CDSM Directive shed light on what lawful access is supposed to mean. Such requirements are often understood as use on the basis of a copyright exception, or that is not otherwise unlawful. This does not bring us any closer to understanding what the exception *requires*. But of course, cultural heritage institutions that can digitise their collections pursuant to Art. 6 CDSM Directive thus have lawful access (as required by Art. 3) on the basis of another copyright exception and can therefore rely on Art. 3, even when the initial exception imposes limits on use (in this case the purpose of preservation of collections). The initial exception, Art. 6, merely operates as a 'portal provision' that confirms lawfulness of access to the material for the purpose of Art. 3 which, in this case, happens to be in such an institution's collection. Stated differently, the material that is already in the user's collection was digitised (created in digital form) on the basis of a copyright exception and the actual copyright-relevant use (TDM) of that already existing digital material is governed by Art. 3. Although in this particular case there does not seem to be anything to indicate that digitisation of collections for the purpose of being able to carry out TDM should not be governed by Art. 3. Digitisation is, after all, an act of reproduction defined broadly in Art. 2 InfoSoc Directive and that act is captured by the provision as long as it serves TDM.

When material needs to be obtained directly from rightholders there is more nuance to the equation. Recital 14 of the Directive refers to subscriptions and open access as examples of lawful access. So, that seems to be straightforward – a research organisation gets a subscription to some service that hosts data. They have lawful access, so they can carry out TDM. But does using someone else's login credentials to a bona fide service not mean that access is unlawful? Viewing the content will not constitute copyright infringement because it will likely be exempted by the temporary copying exception – the source is lawful – yet the user in question is not the intended recipient. Do a bona fide service's errors in rights clearance render a user's access unlawful? Those errors inevitably cause the service to infringe copyright (the rights of reproduction, communication to the public, extraction or reutilisation),[69] so the source needs to be considered unlawful even though the service offering the subscription may have acted in good faith.

Moreover, just as *access* is not an exclusive right, *subscription* is not a surgical designation of the form of access. It merely denotes a continuing relationship without indicating the exact conditions. The very impulse of a diligent user of having to reach out to obtain a subscription or some other authorisation to access content inevitably activates a market opportunity. This opportunity places rightholders in a position to determine the specific parameters of access, on a contractual or a technical level, or both. When limits to access are imposed, and a researcher steps outside of those limits to carry out TDM, or even to create a mineable dataset first, do they still have lawful access in the sense that is understood by the

61 These terms warrant an analysis of their own, for which I do not have the space in this contribution. Suffice it to say that they raise complicated questions and challenges when viewed particularly from the perspective of collaboration, private-public partnerships and expectations of research financing institutions. See also João Pedro Quintais, 'What is a 'research organisation' and why it matters' [2025] GRUR International (forthcoming).

62 Software Directive, art 5(1).

63 Database Directive, arts 6(1) and 9(1).

64 Software Directive, art 5(2).

65 Software Directive, art 6(1)(a).

66 Software Directive, art 6(1)(a).

67 Software Directive, art 6(1)(a).

68 Thomas Margoni, 'TDM and generative AI: Lawful access and opt-out' (30 May 2024) 12-16 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5036164> accessed 10 February 2025.

69 InfoSoc Directive, arts 2-3 and Database Directive, arts 5 and 7(1).

provision? If copyright respects freedom of contract, then I hardly think so. The permitted (excepted) act of reproduction that may very well precede the act of TDM does not of itself justify lawfulness of access to the material. It is justified by an act external to the provision. If rightholders are the source of the material, they determine the conditions under which it is released, and can therefore affect the applicability of the provision.

The Directive seems to recognise some sort of market opportunity by exemplifying in recital 10 that the terms of a licence (including an open access licence) could exclude TDM. But the unfortunate solution is a clarification in Art. 7(1) that contractual provisions preventing TDM are unenforceable. It is unfortunate because the prerogative of the drafter of a contract is that it is they who decide what's in it. Refraining from mentioning TDM effectively renders this provision rather useless because there is no contractual term to 'unenforce'. What the European legislature appears not to have accounted for is that the market opportunity does not concern the potential ability to prevent the possibility of exploiting a copyright exception by way of contract, but to define what lawful access is actually supposed to entail. This is the step preceding TDM and on which the whole exception is conditioned.

Besides, rendering a contractual clause unenforceable means just that – it will not be enforced by a court. It is a private law-type of remedy, not a prohibition whose breach carries with it a penalty or fine. Even a risk-averse rightholder might be tempted to use such a clause in a contract. But a risk-averse user will not engage in TDM if the contract explicitly prevents it, especially when they lack copyright expertise. Indeed, the study that was recently carried out for the European Commission on copyright and the research sector emphasises that researchers often lack a clear understanding of their prerogatives under copyright law.[70] This is not at all surprising. They are, after all, not copyright lawyers.[71] For most purposes that concern law and bargaining, researchers are more like consumers, even though they are formally institutional users. Therefore, regular market dynamics apply and they may be discouraged from pursuing a legitimate activity already when it transpires that the rightholder is prepared to contest it.[72]

Moreover, letting a contract be governed by the law of any other state but the law of an EEA Member State where the Directive applies, or a jurisdiction with a similar contractual limitation, might actually preserve the TDM-exclusion as a matter of contract law. Hence, even a national court might actually have to enforce such a contractual term against a user (or consider it a breach of contract), even though copyright will not have been implicated under domestic law.

The opportunities of what lawful access can entail are endless. Here is one simple example. Rightholders can make use of territoriality of copyright – a principle that is a limitation and a burden when enforcing copyright but a phenomenal bargaining tool in contracting.[73] A territorial restriction may, for example, limit a user's access to Sweden.[74] Because of territoriality, this is a fully legitimate restriction. It also relies on the proprietary character of copyright. So a project involving collaboration with researchers in the Netherlands through a consortium set up for the needs of the research project is likely to require those researchers to travel to Sweden or alternatively obtain their own subscription that applies for the territory of the Netherlands. Connecting from the Netherlands, even if technologically possible, will simply not satisfy a contractual restriction designed to permit access only from Sweden. But even if a Dutch university obtains its own subscription, presuming that the service is even available for Dutch users, will Dutch researchers have access to exactly the same data and on the same terms?

A rational rightholder may be tempted to fragment markets, price-discriminate and – something that is not unusual in the entertainment sector in Europe – offer a different repertoire in each Member State, or even an entirely different access experience and functionality. This is the very core of the open-endedness and mercurial character of a *subscription* – conditions for access are not determined by the subscriber but by the service provider. The Directive seems to completely ignore technological parameters of use that shape the specific form of lawful access. While geofencing of IP-addresses is an obvious measure that technologically enforces a territoriality condition, other examples shaping lawful access include:

- Imposing individualised credentials for accessing material by requiring individual researchers to register on the service despite an overarching institutional subscription or limiting access to one user at a time
- Requiring access from specific devices (device fingerprinting)
- Presenting material for viewing through a browser only, without providing access to raw files or offering download functionality
- Delivering material with strong DRM-protection combining different solutions, such as encryption and watermarking that might require circumvention to accomplish copying or extraction.

2. Mining the obstructions: tools of inconvenience and technological tangles

But the complexity of the lawful character of access does not end with the mere opportunity to apply various technological solutions at the supply end. Other provisions in Art. 3 CDSM Directive support the conclusion that the rightholder *is* intended to be placed in a position to determine the exact parameters of access. In particular Art. 3(4) recognises that rightholders can ensure the

70 Commission Study (n 28) 147.

71 This is not a definitive indication of knowledge. As rightly observed by the Swedish government, it may even be difficult for a lawyer specialising in the area to assess whether a copyright exception is applicable; Regeringens proposition 2021/22:278 130.

72 cf Robert Bradgate, 'Consumer Rights in Digital Products. A research report prepared for the UK Department for Business, Innovation and Skills' (Department for Business, Innovation & Skills, September 2010) 11 <https://assets.publishing.service.gov.uk/media/5a797b29ed915d07d35b5ded/10-1125-consumer-rights-in-digital-products.pdf> accessed 10 February 2025.

73 Szkalej, *Copyright in the Age of Access to Legal Digital Content. A Study of EU Copyright Law in the Context of Consumptive Use of Protected Content* (n 12) 84-87.

74 And because of copyright's proprietary character, use can legitimately be limited to a specific user in Sweden.

security and integrity of networks and databases where the material is hosted. That is an invitation to use a variety of technological solutions and 'tools of inconvenience' to ensure compliance with the intended access conditions. For example:

- Fragmenting data (pagination), such as where it is split across multiple pages requiring user interaction to access additional portions of the data (for example by clicking 'next page' or scrolling)
- Hiding or obfuscating the structure of the data to make it harder to parse (HTML/CSS obfuscation). This can involve rendering text in ways that look normal to a human but is unreadable by bots.
- Imposing rate limits on how many requests a user can make in a given period of time and disconnecting or slowing down the connection when the limit is crossed.
- Blocking IP-addresses for multiple rapid requests so that data is accessed from a single IP-address
- Using so-called CAPTCHA[75] challenges that require users to manually solve a challenge to ensure that interactions with a service are from a human user and not a bot (e.g. clicking images or re-typing text from a distorted image).
- Requiring users to log in again after a period of inactivity as a means of preventing continuous access over long periods of time (session timeout). This can be enhanced by using scripts to monitor browser behaviour, mouse movement or keyboard use to detect inactivity (or scraping activity).

None of these potentially preclude TDM in a way that a contractual prohibition does. But all of them define and affect access to the material in various ways and make TDM activities and preparatory acts, including early automation of the process, much harder or perhaps even impossible.

Moreover, Art. 7(2) clarifies that the TDM exception in Art. 3 CDSM Directive is subject to the so-called three-step test in Art. 5(5) InfoSoc Directive which permits the application of particular copyright exceptions only 'in certain special cases that do not conflict with a normal exploitation of protected subject matter and do not cause unreasonable prejudice to the legitimate interests of the rightholder'. The intuitive approach is to apply this provision to evaluate the scope of the excepted copyright-relevant act – in the case at hand to determine what the act of TDM should and should not cover. But could it also not be applied to determine what 'lawful access' is actually supposed to mean? That is exactly what the CJEU did in *Filmspeler* to determine the meaning of another complicated term – 'lawful use' in Art. 5(1)(b) InfoSoc Directive. Although, in case of this exception, 'lawful use' is what the excepted use (the temporary copying under assessment) is intended to enable.[76]

With the three-step test, territoriality of copyright, and *Filmspeler* in mind, would it not be an unreasonable prejudice to the legitimate interests of the rightholder if they were to be unable to fragment markets and selectively apply each of the approaches above in accordance with their business model and commercial opportunities that copyright exclusivity provides? The fact that a user might not be able to carry out TDM because of technological impediments is not an issue that a rightholder must resolve – or care about – because European copyright law does not require rightholders to pre-emptively design the technical parameters of access in a manner that makes it possible for users to benefit from copyright exceptions. It only requires that contractual terms restricting use under Art. 3 CDSM Directive be unenforceable. Even if one were to be able to argue that technological restrictions constitute an implied contractual term that should be captured by the unenforceability rule in Art. 7(1) of the Directive (the recitals do not imply that), the provision does not preclude the use of technology (nor even the use of unenforceable terms).

The combined effect of the lawful access requirement, the three-step test, and use of technology to define access, raises the question of whether researchers can rely on national measures introduced under Art. 6(4)(1) InfoSoc Directive designed to ensure that beneficiaries of certain privileged exceptions can make use of them when so-called effective technological protection measures (TPMs) prevent the envisaged use.

Not all of the measures itemised above might be considered effective TPMs within the meaning of Art. 6 InfoSoc Directive.[77] But determining their status might not even be necessary. All of the technologies above serve to control the act of accessing and viewing copies (admittedly this could make it a TPM[78]). Such an act, by virtue of being a temporary reproduction, is governed by Art. 5(1) InfoSoc Directive. Could this provision be 'enforced' so that it can, in accordance with its literal wording, enable the 'lawful use' of TDM under Art. 3 CDSM Directive?

75 Completely Automated Public Turing tests to tell Computers and Humans Apart.

76 'Temporary acts of reproduction (…) which are transient or incidental and an integral and essential part of a technological process and *whose sole purpose is to enable* (a) a transmission in a network between third parties by an intermediary, or (b) *a lawful use* of a work or other subject-matter *to be made*, and which have no independent economic significance, shall be exempted from the reproduction right … '.

77 In the example above I refer to DRM – digital rights management (systems). They are often used interchangeably with TPMs, but TPMs is a legal term originating from art 11 WIPO Copyright Treaty. DRM is a much broader, supposedly technological concept encompassing a variety of technological solutions, including technologies that copyright law classifies as TPMs. See further Szkalej, *Copyright in the Age of Access to Legal Digital Content. A Study of EU Copyright Law in the Context of Consumptive Use of Protected Content* (n 12) 135-37; 'When DRM is detached from the copyright realm, a more nuanced view of technology surfaces. While supporting rightholders in their commercial endeavours, any technological solutions that are used do not have to, but may, reflect the exercise of prerogatives that copyright law confers. Under this view copyright exclusivity merely serves as a template for imaginable uses in respect of which technological solutions can ensure compliance. This broader view of technology does not lock attention to copyright-relevant acts at the [user] end but embraces the use of technology on the entire distribution chain. Moreover, it more transparently opens up for the flexible nature of technological solutions and the business operations those solutions might be intended to give effect to.' (136).

78 Pursuant to InfoSoc Directive, art 6(3): '"technological measures" means any technology, device or component that, in the normal course of its operation, is designed to prevent or restrict acts, in respect of works or other subject matter, *which are not authorised by the rightholder* of any copyright or any right related to copyright as provided for by law or the sui generis right provided for in Chapter III of Directive 96/9/EC. Technological measures shall be deemed "effective" where the use of a protected work or other subject matter is controlled by the rightholders through application of an access control or protection process, such as encryption, scrambling or other transformation of the work or other subject-matter or a copy control mechanism, which achieves the protection objective' (emphasis added).

Well, this provision is not included in the list of privileged uses in Art. 6(4)(1) InfoSoc Directive, so the answer is an unambiguous no. Moreover, although the scientific research exception in Art. 5(3)(a) InfoSoc Directive is itemised, whenever researchers access material available on an online service on the basis of 'agreed contractual terms' – perhaps a subscription – the national measures cannot cover such material pursuant to Art. 6(4)(4) InfoSoc Directive. In such a case the answer is the same. In other words, as far as these two exceptions are concerned, researchers stay empty-handed.[79] Moreover, the measures, to the extent they are to be viewed as TPMs, are legally protected from circumvention pursuant to Art. 6(1) InfoSoc Directive. But if they are not to be viewed as such, and are circumvented, the question arises whether the lawful access requirement in Art. 3 CDSM Directive is still met.

Article 3 CDSM Directive is, however, included in the list and is not covered by the peculiar limitation in Art. 6(4)(4) InfoSoc Directive.[80] Can this exception be 'enforced'? There are at least four issues to consider. First, in the event that the measures are not TPMs, this scheme does not apply. Second, the scheme is contingent on one of the headache terms – the user must have 'legal access' to the copy, as follows from the English language version of the provision. Somehow this term feels narrower than 'lawful access', as if access can only be based on permission obtained from the rightholder. Either way, it is unlikely that users can be considered to have legal access to the material if they want to connect from a different territory despite geoblocking, if they use someone else's credentials or a different device from the one designated, or if the claim concerns copies that the user managed to create even though the service only intended to make them available for viewing through a browser.[81] Third, all measures above, except perhaps the example with downloadable DRM-protected files, can in fact be argued to ensure the security and integrity of networks and databases where the material is hosted and be allowed under Art. 3(3) CDSM Directive, thus raising the complex question of whether this provision indeed curbs the scope of the exception and whether they are really TPMs. If their aim, in their normal course of operation, is to ensure the security and integrity of databases, they might actually not be.[82] And fourth, the scheme under Art. 6(4)(1) InfoSoc Directive requires the making available of means to benefit from a privileged exception 'to the extent necessary' to benefit from the exception. Do these technologies really prevent reproductions for the purpose of TDM? They are tools of inconvenience. I can envisage the argument that researchers are free, or are even explicitly permitted on the basis of a contract, to exploit the so-called 'analog gap' to reproduce the material; for example, by printing all the materials they can lawfully access, then scanning them to create DRM-free copies and then carrying out TDM of those scans. This would be extremely costly and utterly pointless, but, whilst Art. 6 InfoSoc Directive has been rightly criticised during the last quarter of a century, the provision is not designed to make reliance on privileged copyright exceptions more convenient but only possible. Except in the narrow context of the Cross-border Portability Regulation,[83] user expectations do not as such matter in copyright law analysis. This is certainly one thing to consider for future copyright policy, but for now we might have to accept that lawful TDM might be something like a Schrödinger's cat – simultaneously allowed and denied.

3. The curious case of software and the field of mining software repositories (MSR)

Last but not least, for reasons that are not very clear, and despite that recital 84 CDSM Directive confirms that the Directive observes the principle of equal treatment in Art. 20 EU Charter,[84] Art. 3 CDSM Directive does not cover one particular literary work – computer programs. Yet, it is covered by the general TDM exception in Art. 4. Since the Software Directive also does not contain a scientific research exception, code mining for scientific purposes is likely to implicate copyright protection, unless the Software Directive's exceptions can cover it.[85] In that case, while being conditioned on 'the person having a right to use a copy of the computer program' it would have to have the *sole purpose* of determining the ideas and principles which underlie any element of the program.[86] This might not be enough for the field of mining software repositories (MSR). MSR focuses on analysing and cross-linking the rich data available in software repositories (software data such as source control systems, defect tracking systems, code review repositories, archived communications between project personnel recording information about the evolution and progress of a software development project, question-and-answer sites, continuous integration servers, and run-time telemetry). The aims are to uncover interesting and actionable information to understand software development and evolution, software users, and runtime behaviour; support maintenance of software systems; improve software design and reuse; empirically validate novel ideas and techniques; support predictions about software development;

79 See also Commission Study (n 28) 155.

80 CDSMD, art 7(2).

81 In case of the last example the argument could be made that those copies were created already on the basis of art 3 CDSM Directive as they were made at a time that the user had lawful access and any subsequent reproductions, in whatever form, would have been authorized by the exception, and at the time of the claim the user has not managed to circumvent a TPM embedded in the copy (which would contravene the general prohibition on circumvention of TPMs under art 6(1) InfoSoc Directive and possibly render access unlawful/illegal).

82 Compare the definition of TPMs; above.

83 Regulation 2017/1128 of the European Parliament and of the Council of 14 June 2017 on cross-border portability of online content services in the internal market.

84 On which see Case C-260/22 *Seven.One Entertainment Group v Corint Media* ECLI:EU:C:2023:900 and Kacper Szkalej, 'Private copying levies, broadcasters and the principle of equal treatment' (*Kluwer Copyright Blog*, 29 July 2024) <https://copyrightblog.kluweriplaw.com/2024/07/29/private-copying-levies-broadcasters-and-the-principle-of-equal-treatment-c-260-22-seven-one-entertainment-group-v-corint-media/> accessed 10 February 2025.

85 See however Bohdan Widła, 'Though Shalt Not Conduct Research on Software? Text and Data Mining of Computer Programs in the Current EU Copyright Framework' [2025] GRUR International 11, itemising diverging implementations of the TDM exceptions.

86 Software Directive, art 5(3). The exception does not seem to include adaptation, translation or alteration of the program as noted by Rossana Ducato and Alain Strowel, 'Ensuring text and data mining: remaining issues with the EU copyright exceptions and possible ways out' (2021) 43(5) EIPR 322, 328. Arguably, however, neither translation, adaptation nor alteration of a computer program, to the extent this terminology is intended to express transformative use of a program, is necessary for 'determining the ideas and principles' behind a program in accordance with the limited scope of the exception; cf Widła (n 85) 7 and 10.

and exploit this knowledge in planning future software development.[87] Although copyright protection of a computer program includes both source and object code, as well as preparatory design material,[88] some of the data available in software repositories may, to the extent they are original in the sense of being the author's own intellectual creation, be governed by the InfoSoc-regime instead of the Software Directive. While that may enable TDM of such data on the basis of Art. 3 CDSM Directive, or even the scientific use exception in Art. 5(3)(a) InfoSoc Directive, the deployment of an MSR project inevitably falls between two stools. So much for leading the new technological frontier.

V. Is there anywhere we can go from here, right now?

The present situation is lamentable. Many research institutions are public authorities financed by public means and will by default organise their activities by the book. Yet, the legal landscape of TDM for research under the CDSM Directive presents a multitude of challenges and is tangled in a web of abundant technological solutions determining parameters of access, contractual strategies and loopholes, all encapsulated by a lawful access requirement.

In 2024 the Directorate-General for Research and Innovation (DG RTD) of the European Commission published the study *Improving access to and reuse of research results, publications and data for scientific purposes*.[89] Its overarching objective was to assist the European Commission in delivering the primary outcomes of priority action 2 of the European Research Area (ERA) Policy Agenda 2022-2024. The study sought to identify barriers and challenges to access and reuse of publicly funded R&I results, publications and data for scientific purposes; to identify the potential impact on research; and to propose legislative and non-legislative measures to improve the EU copyright and data law frameworks to make it fit for scientific research, open research data and ERA.[90] Despite its broader regulatory context and primary addressee, some important takeaways are worth recalling for the present problem in order to conceptualise possible initiatives at national level.

1. The future of lawful access requirements – recognising users' copyright-relevant expectations

Legislative change at European level that centres on the TDM exceptions might take some time as the CDSM Directive cannot be reviewed sooner than 7 June 2026.[91] Given the pace of technological advances and needs of the research sector, this might feel like decades away. The Commission Study recommends in policy option CRR-03 the issuing of guidance on the TDM provisions that address aspects that may lead to legal uncertainty and divergent approaches, and practices across the Member States, including on what constitutes 'lawful access'.[92] It is obvious that this is one of the most immediate needs. Of course the option of adopting soft law instruments has limits following institutional and constitutional considerations. Pursuant to Art. 288 TFEU such instruments are not binding.[93] They offer flexibility for addressing particular issues and may invite addressees to adopt or follow a certain manner of conduct and, in that way, contribute to better coordination of national policies. They operate as an instrument to exhort and persuade without generating rights or obligations.[94] Most importantly, they are within reasonable reach for the European regulator. But while the European Commission has the functioning of the internal market in mind, Member States need not sit and wait for developments to unfold. Member States such as Sweden that are in the process of reviewing their system of exceptions are particularly well-placed.[95]

The overarching issue produced by the lawful access requirement concerns power dynamics and contractual arrangements rather than copyright law. The unenforceability of contracts preventing TDM under Art. 7(1) CDSM Directive is a step in the right direction (mandatory copyright exceptions are probably even more rare than headache terms!) but it fails to fully address the evidently asymmetrical bargaining position between researchers and rightholders. Copyright industries are today well-placed to use all sorts of contractual arrangements and technological aids that align with their commercial vision and rights exploitation strategy. In this context it is also important to understand that contracts need not itemise every use – it may sometimes bring a better bargaining edge to say as little as possible in a contract and use a badly drafted or vague copyright exception to one's advantage. As I stated in section IV above, the Commission Study affirms the obvious, that researchers often lack a clear understanding of their prerogatives under copyright law. But it also confirms that rights exploitation strategies may

87 Definition used by the Mining Software Repositories Conference <https://2024.msrconf.org/>, <https://2025.msrconf.org/> both accessed 10 February 2025. See generally on MSR Ahmed E Hassan, 'The road ahead for Mining Software Repositories' in Hausi A Muller, Scott Tilley and Kenny Wong (eds), *Proceedings of the 2008 Frontiers of Software Maintenance* (30 September– 2 October 2008) (IEEE 2008) 48-57. See also Nicolas E Gold and Jens Krinke, 'Ethics in the mining of software repositories' (2022) 27(17) Empirical Software Engineering 1, discussing various challenges for mining software repositories, including copyright, licensing and terms and conditions of services at pp 17 and 27.

88 Software Directive, recital 7; C-393/09 BSA, para 37.

89 Commission Study (n 28). The author was part of the study team.

90 Commission Study (n 28) 43.

91 CDSM Directive, art 30(1).

92 Commission Study (n 28) 201-03; cf pp 157-59. The study also recommends a broader approach to the term given the whole spectrum of exceptions. See policy option CRR-01.2.

93 The Commission's autonomous power to issue recommendations follows from art 292 TFEU.

94 Case C-16/16 *Belgium v Commission* ECLI:EU:C:2018:79, para 26. See also generally Case C-16/16 *Belgium v Commission* ECLI:EU:C:2017:959, Opinion of AG Michal Bobek, paras 87-108 and 166-71. Soft law or administrative practice may however produce legal effects against the Commission; see for example Case T-472/12 *Novartis v Commission* ECLI:EU:T:2015:637, para 67; Case T-376/12 *Greece v Commission* ECLI:EU:T:2014:623, para 108; Joined Cases T-61/00 and T-62/00 *APOL* ECLI:EU:T:2003:60, para 72, and Case C-527/07 *Generics (UK)* ECLI:EU:C:2009:197, Opinion of AG Jan Mazak, para 37. See also European Commission, Better Regulation Toolbox, July 2023 which sets out various legal devices ('tools') which the Commission can rely on when preparing new initiatives and proposals or when managing and evaluating existing regulation. Whilst a significant body of academic texts analyse the functions and effects of soft law, a summarising overview has been prepared by the European Parliament; see generally Denis Batta, 'Better Regulation and the Improvement of EU Regulatory Environment: Institutional and Legal Implications of the Use of "Soft Law" Instruments' (DG Internal Policies of the Union, Legal Affairs, PE.378.290, March 2007).

95 See Official Governmental Inquiry *Inskränkningarna i upphovsrätten* SOU 2024:4.

be in full swing. According to one research performing organisation, their researchers refrain from using research tools that make it possible to mine a large number of protected knowledge resources 'not because we do not want to risk copyright infringement, but because the standard agreement with the publisher specifically does not allow such use'.[96] When combined with technology that may simply be designed to impede TDM, that produces misalignments between outcomes envisaged by the legislator when drafting an exception and preferences of market actors. In other words, it might not be enough to give 'lawful access' the broadest possible meaning and permit TDM even when a researcher has had to step outside of the permitted confines of access to ultimately be able to carry out TDM. The Commission Study points in this direction and recommends, as policy option CRR-01.3, the removal of barriers posed by TPMs. This could be done by, *inter alia*, adding all research-related copyright exceptions to the list of privileged exceptions in Art. 6(4) InfoSoc Directive, including Art. 5(1) (temporary copying) and Art. 5(3) (quotation) of the Directive, broadening the intervention options established in Art. 6(4) of the Directive, or excluding the applicability of Art. 6(4)(4) of the Directive to the general research exceptions and other exceptions relevant for research.[97] But even if a revision of the InfoSoc Directive may be for the European legislature to implement, Member States can also act.

If we perceive the EU legislature's intention as desiring to make TDM for scientific purposes possible, and using on this occasion the known mechanism of unenforceability of contractual clauses, Member States may consider bolstering this mechanism by closing the gap that the lawful access requirement inevitably creates by giving more prominence to the expectation embedded in the exception: being able to rely on the exception in order to carry out TDM.

One way of accomplishing this is by enhancing contract and e-commerce legislation that affects electronic service providers and is subject to supervisory control. Instruments like this usually do not 'cause prejudice to Union law on copyright or related rights', such as consumer law.[98] But adopting provisions that are designed to protect user expectations based on exceptions does not, or even cannot, cause prejudice to copyright law because copyright law does not normally regulate user expectations in their dealings with rightholders. However, this is also changing, as the adoption of the Cross-Border-Portability Regulation demonstrates.[99] In any event, the option of protecting copyright users' expectations seems to lie within the purview of Member State prerogative because it seems to fall neither within the copyright *acquis* nor – at least as far as researchers and institutional users are concerned – within the consumer protection *acquis*. If anything, a provision like this would in fact give effect to Union law on copyright by ensuring that a mandatory copyright exception cannot be rendered inoperative through the exploitation of stronger bargaining power. In that way it also ensures that fundamental rights to which the exception gives expression can find very concrete application.

Another approach could be to enhance national rules transposing Art. 6(4) InfoSoc Directive by extending it to technological mechanisms or processes that are designed, in the normal course of operation, to ensure the security and integrity of networks and databases as permitted by Art. 3(3) CDSM Directive, if those solutions necessitate a TDM user to infract the conditions for lawful access in order to carry out TDM, or if they otherwise obstruct that possibility. This could prevent researchers from shooting themselves in the foot by making previously lawful access unlawful. Such a solution could even extend over contracts defining parameters of access if they appear to convolute the lawful access requirement and prevent a determination of whether the beneficiary actually has such access. Member States ought to be free to design such a mechanism since it also falls outside of the copyright *acquis* by virtue of not, as such, concerning the use of TPMs harmonised by Art. 6 InfoSoc Directive. On the other hand, a broad interpretation of what TPMs are (but do we really want that?)[100] might just extend the national Art. 6(4)-mechanism to cover technologies that shape the exact parameters of access. In this respect the InfoSoc Directive might actually serve as a legal basis since Art. 6(4) requires Member States to make available to a beneficiary 'the means' of benefiting from a privileged exception.

Considering this alternative, it is important to note that rightholders are not as such precluded from using different technological mechanisms, as permitted by Art. 3(3) CDSM Directive. The scheme under Art. 6 InfoSoc Directive is based on voluntariness. A rightholder may very well be happy to offer unconstrained access if a beneficiary of a privileged exception asks. The scheme kicks in first when that does not happen. In this way it arguably preserves a rightholder's freedom to conduct a business protected by Art. 16 EU Charter and generally use such technologies.[101] The same goes for the first option if rightholders can choose how to comply.

There is a third option too. Under Art. 3(4) CDSM Directive, Member States shall encourage the parties involved to define commonly agreed best practices concerning the application of measures that ensure the

96 Commission Study (n 28) 158.

97 Commission Study (n 28) 195-96; cf pp 154-55.

98 For example Directive (EU) 2019/770 of the European Parliament and of the Council of 20 May 2019 on certain aspects concerning contracts for the supply of digital content and digital services, art 3(9); Regulation (EU) 2018/302 of European Parliament and of the Council of 28 February 2018 on addressing unjustified geo-blocking and other forms of discrimination based on customers' nationality, place of residence or place of establishment within the internal market and amending Regulations (EC) No 2006/2004 and (EU) 2017/2394 and Directive 2009/22/EC, art 1(5).

99 The Regulation obliges rightholders and service providers to provide access to their streaming services to their own subscribers when they temporarily visit another Member State and prohibits them from entering into contractual arrangements between each other or with subscribers that are contrary to the Regulation One can ask what the 'exception-based expectation' is in this case. The Regulation was not drafted using copyright terminology so references to copyright law are few, although it is perceived of as belonging to copyright legislation; see Irini Stamatoudi and Paul Torremans (eds), *EU Copyright Law* (2nd edn, Edward Elgar 2021). But isn't it obvious that it's the expectation to be able to rely on art 5(1) InfoSoc Directive in every Member State?

100 The temptation may be high when afficionados of copyright exceptions analyse interventions that Member States can take pursuant to art 6(4), but very low when discussing the anti-circumvention provisions. The broader the concept, the broader the prohibition on circumvention.

101 cf Case C-314/12 *UPC Telekabel Wien v Constantin Film* ECLI:EU:C:2014:192 (open injunction against an ISP allowing it to select the means does not interfere with freedom to conduct a business).

security and integrity of networks. This obligation may not be straightforward to imagine. However, one way of resolving the issue would be to facilitate stakeholder dialogues to establish use of application programming interfaces (APIs) for researchers to access databases in ways that respect security and copyright but also allow for efficient data mining without excessive technical hurdles. If difficulties arise, Member States could 'encourage' stakeholders by adopting a mandatory provision that requires such use. But this is perhaps not that different from the first option, except that it may affect rightholders' capacity to 'freely use, within the limits of its liability for its own acts, the economic, technical and financial resources available to it' as protected by Art. 16 EU Charter.[102]

2. Facilitating mining software repositories (MSR)

The exclusion of computer programs from Art. 3 CDSM Directive, leaving them covered only by the broader TDM exception in Art. 4, is difficult not to see as a glaring inconsistency.[103] As a mode of strengthening the general research exceptions, the Commission Study recommends in policy option CRR-01.1 the 'introduction of a fully harmonised, mandatory and open-ended exemption of scientific research that applies horizontally across the ISD, RLD, DBD, as well as the Software Directive'.[104] It also considers issuing guidance, but highlights that such an approach will not overcome conceptual differences in the *acquis* nor that it would be an efficient tool to address the absence of a scientific research provision in the Software Directive.

Considering this recommendation, it is important to recognise that a fragmented legal landscape consisting of several scientific use exceptions scattered around several directives is an EU *acquis* problem, not a national one. Where Member States have introduced such exceptions it usually takes the form of just one provision rather than three or four (two for databases) separate ones. But can Member States introduce a scientific use exception proper covering software when no such exception exists in the Software Directive? The issue seems particularly vital to the field of mining software repositories (MSR) as the use of software repositories involves copyright information protected both by the Software Directive and by the InfoSoc Directive.[105]

In the course of the last fifteen years we have learnt a lot about EU copyright law through the astonishing number of judgments from the CJEU (over a hundred cases and counting). But only two are necessary to offer an answer. In *Pelham* the CJEU was essentially asked by the German Supreme Court if it was possible to introduce new copyright exceptions on the basis of fundamental rights under the InfoSoc Directive.[106] The CJEU's negative answer can be boiled down to these two arguments. Having in mind that the list of exceptions in the InfoSoc Directive is exhaustive pursuant to recital 32, it would endanger the effectiveness of harmonisation of copyright effected by the Directive and the objective of legal certainty pursued by it; and the requirement of consistency in the implementation of exceptions would not be able to be ensured if Member States were free to provide for exceptions beyond those expressly set out in the Directive.[107] This is a strong argument that may very well be considered in the context of the Software Directive, even if the list of exceptions in that Directive does not appear to be exhaustive.[108] However, if we are prepared to accept that the Software Directive actually is *lex specialis* as the court made clear in *UsedSoft*,[109] then we might also have to accept that the CJEU's decision in *Pelham* does not apply to that Directive. In other words – and having regard in particular to the need to ensure that the rights and freedoms in the EU Charter are given concrete expression – it appears to me that it lies in Member States' prerogative power to introduce a research exception that covers software.

To the extent that Art. 25 CDSM Directive can be read as *constraining* Member States' discretion to adopt or maintain broader exceptions,[110] it does not appear to pose a barrier to such an initiative because the provision only applies to uses or fields 'covered by exceptions and limitations provided for in this Directive', which arguably excludes scientific research of software. In any case, a software research exception would hardly be incompatible with the equivalent exceptions in the InfoSoc Directive or the Database Directive.[111] Additionally, the CJEU's decision in *Top System* also seems to confirm that there is a little bit more flexibility to construe software exceptions.[112] More broadly, the need for a scientific use exception encompassing software arises for simple reasons of coherency of the copyright system. The principle of equal treatment enshrined in Art. 20 of the EU Charter requires that comparable situations must not be treated differently and that different situations must not be treated in the same way unless such treatment is objectively justified.[113] A copyright system that is incomplete in the sense of providing a copyright exception for the benefit of researchers working with particular categories of protected material – such as text, images, sound, film, or databases – but not software, risks preventing the rights and freedoms following from the Charter to be given practical meaning for those researchers working with such latter material. It moreover insufficiently accounts for expectations and practices within a discipline or diversity of research data. Conversely, an incomplete framework treats authors and holders of neighbouring rights and producers of databases differently from developers of computer programs. In this respect, it is essential to observe that the reuse and development of computer programs or systems is not necessarily limited to historically relevant fields typically associated

102 ibid para 49.

103 Ducato and Strowel (n 86) 328 calling it a systematic inconsistency.

104 Commission Study (n 28) 185-87; cf 147-50.

105 Above section IV.3.

106 *Pelham* (n 35) para 25 (Question 3 read together with Question 6).

107 ibid paras 63-64.

108 In addition, recital 19 Software Directive confirms derogations from the directive are possible on points not covered by it as long as conformity with the Berne Convention is ensured.

109 Case C-128/11 *UsedSoft v Oracle International* ECLI:EU:C:2012:407, para 56.

110 The exceptions must be 'compatible with the exceptions and limitations provided for' in the Database Directive and the InfoSoc Directive.

111 cf Widła (n 85) 10 who offers an argument based on recital 19 Software Directive and art 9(2) Berne Convention.

112 Case C-13/20 *Top System v Belgian State* ECLI:EU:C:2021:811 (operative part).

113 *Seven.One Entertainment Group* (n 84) para 45.

with these activities, such as engineering or computer science. It may also extend to natural sciences, especially in the case of once esoteric and now established fields such as bioinformatics or neuroinformatics,[114] or any other field requiring an interdisciplinary approach involving informatics. Considering that computer programs qualify as literary works for copyright purposes under international, EU and, inevitably, national law,[115] just like traditional literature, maintaining a distinction that carves out computer programs from scientific research may be difficult in an increasingly computerised research sector without an objective justification.

3. Concluding remark

A desire to lead the charge in innovation requires streamlined rules that balance copyright exclusivity with access to data for scientific ends. If simple solutions are not found, research institutions might find themselves spending more time mining legal texts than actual data.[116]

114 See generally Jeff Gauthier and others, 'A brief history of bioinformatics' (2019) 20(6) Briefings in Bionformatics 1981; Losiana Nayak and others, 'Computational neuroscience and neuroinformatics: Recent progress and resources' (2018) 43(5) J Biosci 1037; B Nolan Nichols and Kilian M Pohl, 'Neuroinformatics Software Applications Supporting Electronic Data Capture, Management, and Sharing for the Neuroimaging Community' (2015) 25(3) Neuropsychol Rev 356; Marc-Oliver Gewaltig and Robert Cannon, 'Current Practice in Software Development for Computational Neuroscience and How to Improve It' (2014) 10(1) PLoS Comput Biol 1; Susan M Baxter and others, 'Scientific Software Development Is Not an Oxymoron' (2006) 2(9) PLoS Comput Biol 975; Jacques Cohen, 'Computer Science and Bionformatics' (2005) 48(3) Commun ACM 72.

115 Agreement on Trade-Related Aspects of Intellectual Property Rights, art 10; WIPO Copyright Treaty, art 4; Software Directive, art 1.

116 Lawyers, on the other hand, might need some convincing before mining Big Law. See Johan Lindholm, 'Textual Insights: What Can Computers Teach Legal Scholars About Law?' (4 December 2024) <https://ssrn.com/abstract=4942744> accessed 10 February 2025.