

# Kluwer Copyright Blog

## Saving research: Lawful access to unlawful sources under Art. 3 CDSM Directive?

Thomas Margoni (Centre for IT and IP Law (CiTiP), Faculty of Law, KU Leuven) · Friday, December 22nd, 2023

Arts. 3 and 4 of the Copyright in the Digital Single Market Directive (CDSMD) introduced two exceptions for Text and Data Mining (TDM) in EU copyright Law. These two exceptions, despite having different objectives, share several similarities, as [scholar analysis](#) has shown. One of these common aspects is the requirement of *lawful access*. Only if the training material can be lawfully accessed do the TDM exceptions apply. However, the exact delineation of lawful access is not clear in the CDSMD, nor in the EU copyright *acquis*. This blog, based on an upcoming paper, advances the view – perhaps unconventional, certainly purpose-driven – that the requirement of lawful access should only cover the behavior of the beneficiary of the exception and not extend to the status of the accessed source. If, as it is here argued, this view is acceptable, then the lawful access requirements of Arts. 3 and 4 may lead to different outcomes than until now assumed. At least in the case of Art. 3 – TDM by research organizations for research purposes – it could be possible to have lawful access to unlawful sources.



Photo by [5477687](#) via Pixabay

### The problem

This interpretation, which departs from most ([not all](#)) of the previous assessment developed in the literature, is not merely intended to broaden the body of knowledge available for scientific research, a [deserving goal in its own right](#). This interpretation is an existential condition for the

effective enjoyment of Art. 3. Should lawful access be interpreted as requiring a lawful source, lawfulness will be assessed in light of cases from the Court of Justice of the EU (CJEU) including *ACI Adam* and *Copydan* which examine the right of reproduction, the same right at the core of Arts. 3 and 4. However, it seems plausible, if not even likely, that this assessment may also look at the more elaborated case law on the right of communication to the public in relation to digital technologies (e.g., *Svensson*, *Stichting Brein*, *GS Media*, etc.). Cases like *Renckhoff* might even unveil a predisposition by the Court to address reproductions in online environments as cases of communication to the public (although the role of referring courts and national law should be duly considered).

The reported line of cases essentially establishes either *i*) a total ban on unlawful sources (without clarifying how this unlawfulness should be assessed, thus arguably strictly, e.g., *ACI Adam*), or *ii*) a presumption of knowledge of the unlawfulness of the source when operating for financial gain purposes (*GS Media*). Accordingly, research organization may have to comply with two alternative frameworks. In the first, they are asked to perform a strict assessment of lawfulness for every single piece of content used for TDM, something that may be quite difficult for most content available on the Internet. This framework will almost unavoidably reduce Art. 3 to an ineffective provision.

This leads to the second option. By adopting the more lenient case law on linking, it is arguable that research organizations operating for research purposes within Art. 3 do not pursue financial gain and thus are not liable if they have no direct knowledge of the unlawfulness of the source. However, even this second option is not as suitable as it might look. Research organizations' (presumed) lack of knowledge will cease once they become aware of the unlawfulness of the sources (e.g., via a notification) and they will be under an obligation to remove it (see *GS Media* at 53). Here lies the problem with this option. Whereas it is possible to remove a link to an unlawful source (*GS Media*), a photograph from a website (*Renckhoff*), and arguably also a work from an Art. 3(2) storage dataset, it is simply impossible or (considering statements of theoretical feasibility) extremely costly to remove the information derived from an unlawful source in a trained model.

Whereas an argument about the enduring lawful effects of acts made during the temporal validity of the presumption could be made, this would introduce yet another severe layer of legal uncertainty for research organizations. The only realistic way to avoid the risk of costly copyright infringement litigation would be to retrain from scratch the whole model on the same dataset cleaned up from the infringing content. Every single time that such knowledge/notification is obtained. For models trained on sometimes millions of works this would lead to an untenable situation. Such an eventuality would be particularly concerning for research organizations acting for research purposes which, differently from for-profit initiatives, often lack the financial structure to clear copyright *ex-ante* or to purchase "copyright shields".

As previously argued, uncertainties like this one constitute a powerful push away from copyright exceptions and an attractive argument in favor licensing or of remuneration schemes based on statutory licensing or levies. While plausible, even perhaps desirable, for commercial uses under Art. 4, a similar result would nullify the effect of Art. 3 and sacrifice the public policy dimension of scientific research at the altar of legal formalism. This is not only regrettable from a policy point of view given the hard-fought legislative compromise on TDM. This course of action would be incompatible with the CJEU case law stating that EU law, particularly copyright exceptions, must be interpreted in a way that safeguards the effectiveness of the provision (see *VOB*).

### The solution (or a possible one)

In order to offer such a judicially required effective interpretation that safeguards the purpose of Art. 3, it is here suggested to simply detach the concept of lawful access from that of lawful source – something that seems plausible given the taxonomical fragmentation in the area – and develop an interpretative framework that supports the fundamental balancing of rights between private and public policy goals established in the CDSMD (Recs. 2-6).

This alternative interpretation starts from a **literal reading of the preamble of the CDSMD**. Rec. 14 identifies four main types of lawful access in relation to Art. 3. They are access to content:

*i)* based on open access policies;

*ii)* through contractual arrangements between right holders and research organisations, such as subscriptions;

*iii)* through other lawful means; or finally

*iv)* that is *freely available online*.

Options *i)*, *ii)*, and *iii)* of Rec. 14 seem unproblematic. Option *iv)* however could offer an interesting opening towards much wider access to training material for research organizations acting for research purposes. Option *iv)*, introduces the concept of content *freely available online* which seems to be new wording in the copyright *acquis* and thus in need of proper demarcation. Cases like *CV-Online Latvia* or *VG Bild-Kunst* do not seem to develop this concept towards a prescriptive category.

Logically, content *freely available online* must have a broader scope than the options *i)*, *ii)* and *iii)* of Rec 14, since these are quite well-defined cases in relation to which the choice of language (“lawful access *should also cover ...*”), appears to create an additional category. Again logically, *freely available online* cannot mean (although it could include) content *made available to the public* online (Rec. 18), for very similar reasons: the choice of wording by the legislator is sufficiently different to create the space to argue that these are – intentionally? – forms of lawful access to different types of content.

### Freely available online v. lawful sources

The CJEU has long insisted on the lawful nature of the *source* needed for specific exceptions and limitations but has otherwise employed an inconsistent taxonomy (lawful use, lawful user, lawful source, legal access, etc) in a way that has contributed to create a degree of uncertainty around the exact perimeter of the elements needed to be lawful – a sort of *lawful-uncertain* approach. [Scholars](#) on the other hand have attempted a much-needed conceptual reconstruction intended to make sense of this terminological uncertainty in a way that coordinates the various concepts around a set of [common criteria](#).

The key hermeneutic issue is whether the CDSMD legislator has embraced this *lawful-uncertain*

approach or, on the other hand, whether it has deviated from it by creating a new and different category. Here we focus on this second possibility, which seems *inter alia* supported by a choice of language that not only introduces the new category of *freely available online*, but which also carefully avoids any reference to the expression *lawful source* in the preamble or in the Articles.

Once again, we start from a literal reading, in this case of Arts. 3 and 4. **The focus of the language is on the lawfulness of the access, not the accessed copy**, or in other words of the source. This makes sense since it is often difficult to establish whether something that is available online is there with the authorization of the right holder. Illustratively, circumventing a paywall, a technological protection measure, or a valid contractual limitation are arguably types of unlawful access that would disqualify the potential beneficiary. But a researcher who behaves lawfully should be shielded by the eventual unlawfulness of the uploaded content. The examples listed in Rec. 14 can be read as confirming this approach. They illustrate types of conducts that the miners should follow to stay within the boundaries of the exception. *A contrario*, the choice of using statutory language in Rec. 18 (*made available to the public*) could be read as the intention, or at least as the plausibility, to subsume this act within the statutorily regulated right of communication to the public in relation to Art. 4.

Cases like *ACI Adam* and *Copydan* do not seem to stand in the way of this interpretation. Both cases start from the observation that the InfoSoc Directive “... *does not address expressly the lawful or unlawful nature of the source from which a reproduction of the work may be made*” (*ACI Adam* at 29). However, the CDSMD does address the lawful or unlawful nature of the source, and it does that through the concept of lawful access which, in turn, refers to the types present in the Recitals.

If we accept that *freely available online* does not require *lawful sources*, the next step is to assess what does it require or refer to. Arguably it refers to **all content that can be freely accessed on the public internet**. This excludes content behind a paywall or protected by valid technological measures. But it includes content that is freely available online without the consent of the right holder or absent any other legal basis. In other words, **unlawful sources**.

**Why should this be allowed and why for Art. 3?** The reason seems to pervade the whole CDSMD and to a good extent most of the existing *acquis*. Scientific research deserves a special treatment due to its extraordinary educational, scientific, social, economic, and cultural importance. Research organizations such as Universities are essential institutions in the social fabric of our societies and as such find explicit recognition in EU and international fundamental right instruments (e.g., Arts. 19, 26, 27 UDHR or Arts. 11, 13 and 14 CFR).

This recognition is regularly translated into specific provisions of EU secondary legislation. There are plenty of examples of research’s special treatment in the copyright *acquis* or, for example, in recent data and digital legislation. Academic freedom is emerging as a key enabler of a thriving participatory democracy and for the development of models of scientific and technological innovation not exclusively driven by profit, but instead by EU core values. The beneficiary-, purpose- and use-limitations explicitly set forth by Art. 3 reflect a balance of interests between the public mission of these organizations and the protection of right holders’ interest. Lawful access, not lawful sources – a category that has existed in the [case law of the CJEU](#) since at least 2014 and which could have been chosen by the legislature had they wanted to – is one of the deliberate embodiments of this balance.

The case of Art. 4 and Rec. 18 may very well depict a different scenario since the scope of the provision is not limited to research organizations but includes also commercial players. In this case, by an explicit use of different statutory language, the choice of the legislature could have been (“including”) to link lawful access to content lawfully made available online and thus to the judicial category of lawful sources, in line with the [CJEU case law](#) and [scholarly analysis](#). After all, that Art. 3 and 4 are the expression of very different stakeholder interests and policy considerations is apparent throughout the whole [legislative history of the CDSMD](#) .

In conclusion, the possibility to TDM unlawful content (as long as the access, i.e., the behavior of the agent, is lawful), has several advantages from the point of view of research organizations. It substantially enlarges the body of knowledge that can be mined and removes the enormous monetary, transaction and legal uncertainty costs that would disproportionately affect research organizations and cultural heritage institutions to the point of relegating Art. 3 to irrelevance. The proposed interpretation, which admittedly favors scientific research in an original, if not provocative, way seems nonetheless well justified. Literal, systematic, constitutional and teleological arguments as well as the *travaux préparatoire* offer a sound basis to draw a distinction that may not have been apparent from an initial reading of Arts. 3 and 4, but which is clearly present in the *acquis* and which has fundamental public policy implications.

---

*The author would like to thank Alain Strowel, Rossana Ducato, Luca Schirru, Roberto Caso, Alina Trapova and João Pedro Quintais for their valuable feedback and comments on previous drafts or presentations of this argument. All errors are of course only mine.*

---

*To make sure you do not miss out on regular updates from the Kluwer Copyright Blog, please [subscribe here](#).*

## Kluwer IP Law

The **2022 Future Ready Lawyer survey** showed that 79% of lawyers think that the importance of legal technology will increase for next year. With Kluwer IP Law you can navigate the increasingly global practice of IP law with specialized, local and cross-border information and tools from every preferred location. Are you, as an IP professional, ready for the future?

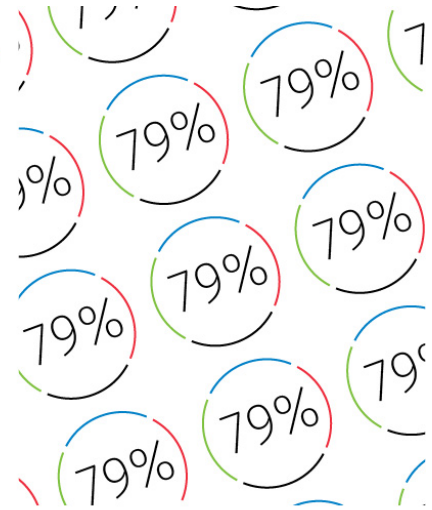
Learn how **Kluwer IP Law** can support you.

---

79% of the lawyers think that the importance of legal technology will increase for next year.

**Drive change with Kluwer IP Law.**

The master resource for Intellectual Property rights and registration.



2022 SURVEY REPORT  
The Wolters Kluwer Future Ready Lawyer  
Leading change

This entry was posted on Friday, December 22nd, 2023 at 8:03 am and is filed under [CDSM Directive](#), [Digital Single Market](#), [European Union](#), [Exceptions and Limitations](#), [Text and Data Mining \(TDM\)](#)

You can follow any responses to this entry through the [Comments \(RSS\)](#) feed. You can leave a response, or [trackback](#) from your own site.