B · M · G
AVOCATS

Newsletter
Geneva, 11 November 2024
Author : Me Philippe Gilliéron, philippe.gillieron@bmglaw.ch

Creation of a dataset for AI training purposes : towards a non-infringement of copyrights?

I. Introductory remarks: the facts

On September 27, 2024, the Hamburg Landgericht (LG) handed down a long-awaited ruling on the issue to know whether the creation of a database (LAION) reproducing almost six billion images, together with descriptive texts and related links, infringes existing copyright on these images.

As a preliminary point, it should be noted that the question was not whether the use of this database for training an AI model infringed copyright, but whether the creation of such a database, made available free of charge to third parties likely to use it for such purposes, infringed copyright.

It should also be remembered that under German law, images are protected irrespective of whether they have an individual character, in accordance with art. 72 para. 1er of the German Copyright Act (UrhG), which states that "*Lichtbilder und Erzeugnisse, die ähnlich wie Lichtbilder hergestellt werden, werden in entsprechender Anwendung der für Lichtbildwerke geltenden Vorschriften des Teils 1 geschützt*".

It was established that the image which had led to the action against LAION had indeed been reproduced and processed in the disputed data set, and that it had been extracted from the www.bigstockphoto.com image bank under license from the plaintiff. The plaintiff's standing to sue, a recurrent issue in this type of cases, was therefore not open to discussion in this case.

If the reproduction of the image was not contested, the question arose as to whether LAION was in a position to avail itself of an exception provided for under German copyright law. Three provisions are examined in turn by the LG, with interesting considerations: § 44a, 44b and 60d of the German Copyright Act.

II. § 44a UrhG

According to this provision, which transposes art. 5 ch. 1 of Directive 2001/29 into German law and corresponds in Swiss law to art. 24a of the LDA :

"Zulässig sind vorübergehende Vervielfältigungshandlungen, die flüchtig oder begleitend sind und einen integralen und wesentlichen Teil eines technischen Verfahrens darstellen und deren alleiniger Zweck es ist, (1.) eine Übertragung in einem Netz zwischen Dritten durch einen Vermittler oder (2.) eine rechtmäßige Nutzung eines Werkes oder sonstigen Schutzgegenstands zu ermöglichen, und die keine eigenständige wirtschaftliche Bedeutung haben "

Or, in English :

"*Permitted are temporary, transient or accessory acts of reproduction, which are an integral and essential part of a technical process and whose sole purpose is to enable (1) transmission over a network between third parties by an intermediary or (2) lawful use of a work or other protected object, and which have no autonomous economic significance.*"

LG Hamburg rightly holds that this exception is inapplicable to the creation of such a database, insofar as the reproductions generated are neither transitory nor incidental.

For the purposes of this provision, a copy can only be transitory if its lifetime is limited to what is necessary for the proper functioning of the technical process concerned, it being understood that this process must be automated so that it automatically removes this act, without human intervention, as soon as its function of enabling such a process to be carried out is completed (CJEU, C-5/08, July 16, 2009, *Infopaq/Danske Dagblades Forening*, cons. 64).

In this case, apart from the fact that the duration of reproduction remained a matter of debate, only manual intervention could lead to the destruction of reproduction, which ruled out its transitory nature.

Moreover, only a copy whose sole purpose is to facilitate the performance of a technical process, to the exclusion of any other purpose whatsoever, can be incidental within the meaning of this provision (CJEU, C-360/13, June 5, 2014, *Public Relations Consultants Associations Ltd v. Newspaper Licensing Agency Ltd and Others*, cons. 43).

However, this was not the case here, where the images were reproduced for analysis by software, which was the primary and sole purpose of the download, and not an incidental one.

III. § 44b UrhG

According to this provision, which transposes art. 4 of Directive 2019/790 into German law:

"*(1) Text und Data Mining ist die automatisierte Analyse von einzelnen oder mehreren digitalen oder digitalisierten Werken, um daraus Informationen insbesondere über Muster, Trends und Korrelationen zu gewinnen.*

(2) Zulässig sind Vervielfältigungen von rechtmäßig zugänglichen Werken für das Text und Data Mining. Die Vervielfältigungen sind zu löschen, wenn sie für das Text und Data Mining nicht mehr erforderlich sind.

(3) Nutzungen nach Absatz 2 Satz 1 sind nur zulässig, wenn der Rechtsinhaber sich diese nicht vorbehalten hat. Ein Nutzungsvorbehalt bei online zugänglichen Werken ist nur dann wirksam, wenn er in maschinenlesbarer Form erfolgt."

That is, in French:

(1) Text and data mining is the automated analysis of one or more digital or digitized works to extract information, including patterns, trends and correlations.

(2) Reproductions of legally accessible works for text and data mining are permitted. Reproductions must be removed when they are no longer required for text and data mining.

(3) The uses referred to in paragraph 2, first sentence, are only permitted if the rights holder has not reserved them. A reservation of use concerning works accessible online is valid only if it is made in machine-readable form."

A. § 44b para. 1er UrhG

With regard to paragraph 1er , the LG notes the following points:

In the first place, the fact that the images were reproduced in order to examine them and compare them with a pre-existing description corresponds to the establishment of correlations required by the text and data mining exception. The fact that the defendant inserted a "*disclaimer*" to the effect that its database had not been cleaned was irrelevant, insofar as this lack of cleaning related only to the absence of filtering of potentially sensitive content, without

calling into question the actual analysis of the images for the purposes of establishing the aforementioned correlations.

Secondly, the argument that the text and data mining exception can only concern the analysis of data underlying reproduced works, to the exclusion of the exploitation of their content, appears to have little relevance. In a digitized world, the distinction between the underlying data and the content itself is delicate, to say the least, as the two tend to merge.

Thirdly, the mere creation of an image bank does not yet allow us to draw any conclusions as to its future use. The creation of such an image bank must be distinguished from the subsequent training of the algorithm within a neural network, and then from the use of the trained system to generate new images. At the time of setting up the image bank, it is not possible to know how successful the training of the algorithm will be, or whether it will be possible to generate new content. As the possibilities of application are not foreseeable at the time the image database is set up, the resulting legal uncertainty means that the lawfulness of setting up such a database cannot be judged on the basis of the supposed future intention to generate artificial content at a later date.

Fourthly, the application of § 44b UrhG cannot be ruled out simply because the legislator could not have had artificial intelligence systems in mind when adopting it. This is demonstrated by art. 53 para. 1er lit. c of the AI Regulation, which stipulates that providers of general purpose AI models must have a strategy in place to ensure compliance with any reservations made in accordance with art. 4 para. 3 of Directive 790/2019.

Finally, the application of the three-step test does not lead to a different result. The exception of data analysis and mining, as in the present case, is a special case, and it is difficult to see how it would be detrimental to the normal use of images. Although the subsequent generation of images on the basis of the system trained for use could be considered as potentially prejudicial to the normal exploitation of the works, this is not the case for the mere creation of an image bank and data sets for training purposes. These two stages must be clearly distinguished.

The LG therefore emphasizes the distinction to be drawn between the simple creation of data sets, which are then used to train algorithms, and the use of the trained system to generate content. The only question for the LG to consider is whether or not this constitution is lawful, independently of its subsequent use, which is another matter. In the LG's view, and in my opinion rightly so, judging the lawfulness of the creation of such an image bank on the basis of its possible (but at this stage uncertain) subsequent use would be tantamount to excluding the application of § 44b UrhG in the majority of cases, which would be contrary to the legislator's intention.

B. § 44b para. 2 UrhG

The LG noted that the images reproduced by the defendant were freely accessible to the public. Only images with a digital watermark were reproduced and made available free of charge as previews, to the exclusion of images without a watermark, which could only be downloaded subject to prior payment.

C. § 44b para. 3 UrhG

Notwithstanding the foregoing, the LG considers that the plaintiff had availed itself of the proviso in § 44b para. 3 UrhG allowing a content provider to prohibit the extraction of content for the purposes of data mining and analysis.

In this respect, the LG notes the following points:

Firstly, the copyright owner is entitled to rely on the reservation made by one of its licensees. In this case, the reservation was made by the www.bigstockphoto.com platform, to which the

photographer had granted a license on the offending image, subject to sub-licenses in favor of users. The plaintiff should have been able to take advantage of this.

Secondly, the reservation must have been made expressly and be easily understood by users. This is the case of an express reservation made in the general terms and conditions available on the website concerned.

Finally, a reservation contained in general conditions must be accepted as machine-readable. As far as the LG is concerned, there are sufficient AI tools available today for the automated reading of texts to expect interested parties to make use of them. This expectation is expressly stated in art. 53 al. 1er lit. c of the AI Regulations, which refers to the expected use of "advanced technologies" to identify such a reservation.

D. Intermediate conclusion

At this stage, therefore, the LG considers that the creation of an image database which can then be made available free of charge to third parties for the purpose of training AI models falls within the scope of § 44b UrhG, but that the insertion of a reservation, as was the case here, prohibits extraction.

The LG does not stop there, however, and continues its analysis by examining the application of § 60d UrhG.

IV. § 60d UrhG

According to this provision, which transposes art. 3 of Directive 790/2019 into German law, and whose first two paragraphs have particularly caught the LG's attention:

"(1) Vervielfältigungen für Text und Data Mining (§ 44b Absatz 1 und 2 Satz 1) sind für Zwecke der wissenschaftlichen Forschung nach Maßgabe der nachfolgenden Bestimmungen zulässig.

(2) Zu Vervielfältigungen berechtigt sind Forschungsorganisationen. Forschungsorganisationen sind Hochschulen, Forschungsinstitute oder sonstige Einrichtungen, die wissenschaftliche Forschung betreiben, sofern sie (1.) nicht kommerzielle Zwecke verfolgen, (2.) sämtliche Gewinne in die wissenschaftliche Forschung reinvestieren oder (3.) im Rahmen eines staatlich anerkannten Auftrags im öffentlichen Interesse tätig sind.

Nicht nach Satz 1 berechtigt sind Forschungsorganisationen, die mit einem privaten Unternehmen zusammenarbeiten, das einen bestimmenden Einfluss auf die Forschungsorganisation und einen bevorzugten Zugang zu den Ergebnissen der wissenschaftlichen Forschung hat".

That is, in English:

"(1) Text and data mining (art. 44b para. 1 and 2, first sentence) is permitted for scientific research purposes, in accordance with the following provisions.

(2) Research organizations are authorized to make reproductions. Research organizations are universities, research institutes or other institutions that carry out scientific research, provided that (1) they pursue non-commercial objectives, (2) all profits are reinvested in scientific research or (3) they act in the public interest as part of a mission recognized by the State.

Research organizations that cooperate with a private company that has a decisive influence on the research organization and privileged access to the results of scientific research are not authorized organizations within the meaning of the aforementioned provision.

The LG retains the following three points:

With regard to the first point, the concept of "scientific research purposes" must be interpreted broadly. Even if the act in question is not directly carried out for "scientific research purposes", it is sufficient that it constitutes a stage in such research, necessary to enable such research to be carried out at a later date. This is the case for LG in the creation of an image bank, as in the present case, which is made available free of charge and which may subsequently be used by third parties to train a neural network. Furthermore, it is irrelevant whether the third parties in question exploit the database for commercial purposes, since such a purpose does not exclude scientific research considerations, which are generally accepted in the context of data analysis to follow the LG.

With regard to the second figure, the LG considers that it has been achieved, given that the image bank is made available to third parties free of charge, regardless of whether the third parties using it to train their models are likely to be engaged in commercial activity.

Finally, the LG concludes that it has not been sufficiently established by the plaintiff that the defendant would have had any connection whatsoever with a particular enterprise having a determining influence on it. In this respect, the LG considers that the plaintiff has not met the burden of proof required to establish such an affiliation.

Finally, the LG concludes that the defendant is entitled to rely on the exception provided for in § 60d UrhG. Accordingly, the plaintiff's claim for copyright infringement is dismissed.

V. Comment

What can we learn from this ruling?

First of all, we need to clearly distinguish between the different stages in the development and operation of AI models, i.e. :

- the establishment of a data set enabling such models to be trained, the only one at issue in this case;
- training the algorithm using this set; and
- the use of the trained model by users.

The legal assessment may vary according to the stage we are at and the players involved.

In the case in point, it was only the first stage, involving the creation of an image bank, that was at issue. Without indulging in a detailed exegesis of this first-instance judgment, the following points are worth noting:

The creation of a database involving correlations to verify the accuracy of data relating to images and their textual descriptions, meant to ensure the quality of the database thus created, constitutes an analysis that falls within the scope of the exception provided for text and data mining.

An author whose work is included in an image bank set up by an agency to which he has granted exploitation rights (license) is entitled to rely on reservations made in the general terms and conditions by his licensee, prohibiting third parties from using robots to extract images. A textual reservation in the general terms and conditions is considered to be machine-readable and therefore acceptable, since third parties can be expected to use advanced technological means, such as AI, to enable them to take cognizance of these reservations, which are thus enforceable against them.

The above analysis, based on art. 44b UrhG transposing art. 4 of Directive 790/2019, thus led to the conclusion that the creation of the image bank infringed the plaintiff's copyright, insofar as the defendant had not taken into account the contractual prohibition on extracting such images for analysis purposes.

But the LG doesn't stop there. For the LG, the training of a neural network constitutes research "for scientific purposes", and the establishment of a database, a necessary step to enable this training, is therefore part of it. It is irrelevant that third parties using the database are pursuing a commercial goal, which does not exclude research "for scientific purposes". The LG therefore concludes that Art. 60d UrhG, which transposes Art. 3 of Directive 719/2019, validates LAION's activities.

While we can follow the LG's reasoning with regard to art. 44b UrhG, its reasoning with regard to art. 60d UrhG is perplexing for two reasons:

- Firstly, while we can accept that scientific research is only possible when an image bank has been set up, and that the entity setting it up can therefore benefit from the exception available to its users, it is surprising that the exploitation of the database to produce a model for commercial use can be considered to be for scientific purposes. If the analysis and processing of data constitutes a scientific purpose by the mere fact of such processing, then it is hard to see what would not fall under the heading of "scientific purposes". What, then, would be the *raison d'être* of art. 4 of Directive 790/2019 if its art. 3, which is more generous since it is not possible to exclude its application by contract, were to apply systematically?

- Secondly, if the legal processing of a dataset is to be distinguished from its training, does this mean that only the entity that creates the dataset reproduces the works contained therein, to the exclusion of the entity that exploits the dataset for the purpose of training its algorithm, and from which only tokens, modified by techniques such as diffusion models, would be exploited? Only the constitution of such a set would then be likely to trigger the application of copyright, to the exclusion of its training. To be continued.