Data mining and Copyright

Kyungsuk Kim

Associate Professor of the Department of Intellectual Property, Sangmyung University, Seoul
miju0704@smu.ac.kr

Abstract

Data mining has broad applications that reach beyond scholarly and scientific research and provide internet search engine services that are commonly used forms of Text and Data Mining ('TDM') of websites. The exceptions and limitations for data mining provide a competitive advantage in the global race for policy innovation because it permits researchers to conduct computational analysis - TDM on any materials to which they have access. For this purpose, Japan and the EU added limitations on copyright to legalize some TDM research through amendments to copyright law, and the U.S. copyright law has allowed data mining by the fair use provision. On the other hand, there are no explicit exceptions and limitations for data mining under the Korean Copyright Act, and there are no cases considering data mining fair use. We review comparatively exceptions and limitations on copyright which will help to encourage AI-related business by using more data smoothly through the mining process and extracting more valuable information.

Keywords: data mining, AI, fair use, exceptions and limitations, Directive on Copyright in the Digital Single Market (CDSM), sui generis, database, opt-out, contract-out, extraction

1. Introduction

It takes a lot of time and effort to find even a small piece of gold by panning. It is estimated that 26 tons of rocks and other stuff have to be sorted through to extract enough to make a single gold ring. That's a lot to sift through. When it comes to data mining, the concept is the same if the gold is replaced with insights and the panning is replaced with algorithms. In a similar fashion to gold mining, the process of extracting valuable nuggets from large rocks, data mining is a method of extracting valuable information from large datasets.

Basically, data mining is about processing data, identifying patterns and trends in that information, and turning it into useful knowledge. Data mining technology has accelerated rapidly over the past two decades, enabling companies to make more informed decisions, and is being used in a wide range of industries, from marketing to education [1]. One of the main advantages of data mining is that it can help businesses make predictions about future trends by analyzing past data. Data mining can also help businesses identify relationships between different pieces of data that businesses might not previously have been able to see, which will help invent new information. The more data is used in the mining process, the more valuable information is extracted.
However, data mining processes can involve copyright issues if the dataset contains photos, videos, text, or other copyright-protected original work. A database showing the originality of the creator can be protected in the same way as the original. In addition, database producers who have compiled databases are protected under copyright law if the data is creatively selected or arranged to achieve the original database. The data mining process involves accessing, collecting, storing (copying), transforming, and transmitting original works that may infringe copyrights, such as the right to reproduce, the right to derive, and the right to communicate with the public. Data mining is essential to the AI industry, which will be the mainstay of the future industry. To promote the AI-related sector, many national copyright laws define data mining as one of the exceptions and limitations of copyright or treat it as fair use.

On the other hand, there are no explicit exceptions and limitations for data mining under the Korean Copyright Act, and there are no cases considering data mining fair use. Therefore, Korea's AI sector could be at a considerable disadvantage if the much higher costs of AI development in Korea are incurred due to the need for licensing negotiations on the vast amount of copyrighted works required for input data. To this end, the copyright revision bill of 2021 contains explicit exceptions and limitations for data mining, but it has not yet passed the National Assembly and is still pending. This article intends to review whether the copyright revision bill will properly respond to future AI-related industries with the exceptions and limitations or fair use for data mining, compared to other national copyright laws.

2. Overseas Copyright Laws for Data Mining

2.1. The Need for Copyright Limitations on Data Mining

When using technologies such as data mining, it is often impossible to obtain permission from the copyright holder for each individual work, and it is difficult to expect permission from the copyright holder because the author is unknown or the original source is unclear. Even if individual permission from the copyright holder is available, it is meaningless to obtain permission if the effort and cost required to obtain such permission exceed the value of information extracted from data mining.

Therefore, there are concerns that AI-related fields might give up their efforts to extract valuable information through data mining. For this reason, the EU and Japan stipulate exceptions and limitations on copyright in data mining activities through copyright revision, and the United States considers data mining to be fair use.

2.2. The EU CDSM

The CDSM defines TDM as ‘any automated analytical technique aiming to analyze text and data in digital form to generate information such as patterns, trends and correlations’ (Art. 2(2)) as well as ‘the automated computational analysis of information in digital forms, such as text, sounds, images or data’ enabled by new technologies (Recital 8). Art. 2(2) is a comprehensive definition that appropriately identifies the potential of tools able to analyze vast amounts of data autonomously or semi-autonomously [2].

The European Union also has a separate sui generis database right in the Database Directive (96/9/EC), which applies to the content of databases where significant investments have been made to acquire, verify or present data [3]. This sui generis database right is always granted to investors—for example, an employer company, or so-called database producer. Therefore, when datasets and personal works are used in the TDM process, both the rights of the database producers and the rights of the authors may be infringed.

For the smooth exploitation of copyrighted works and datasets in the data mining process, the EU introduced
two exceptions to the TDM (Arts. 3 and 4) in the CDSM, both of which are mandatory for the Member States. Art. 3 is imperative and seeks application in the case of TDM for scientific research in research and cultural institutions [4].

The goal of Art. 3 is to introduce a mandatory exception under EU copyright law which exempts acts of reproduction and extraction made by research organizations and cultural heritage institutions in order to conduct TDM for the purposes of scientific research. In addition, contracts cannot prevent the opt-out of TDM conducted by the research and cultural institutions for scientific research, and the circumvention of technical protection measures for TDM is also allowed (Art. 7(2)) [5].

Art. 4 mirrors Art. 3 with significant differences. Art. 4 allows anyone to use the copyrighted works for TDM, however, it can be expressly reserved by rightsholders with ‘opt-out’ or ‘contract-out’. In other words, it may be overridden by ‘opt-out’ or ‘contract-out’ [6].

As a result, if the use of works and other subject matter is expressly reserved by rightsholders with ‘opt-out’ or ‘contract-out’, firms, governments, citizens, journalists, and anyone else who is not a research and cultural organization acting for research purposes have to obtain a specific authorization from rightsholders to develop AI. If there is no ‘opt-out’ or ‘contract-out’, the reproductions and extractions may be retained for as long as is necessary for the purposes of TDM (Art. 4(2)).

In sum, Art. 3 for scientific research by research and cultural institutions does not allow contract-out or opt-out with Technological Protection Measures (‘TPM’) to prevent a research and cultural organization from accessing, copying, and extracting the copyrighted works. On the other hand, for uses other than scientific research, TPM and contract may override Art. 4(2) by allowing copyright holders to reserve the use of works and other subject matter. European countries have already introduced TDM copyright exceptions and limitations into their copyright laws including the UK (Art. 29A), Germany (Art. 60d) and Switzerland (Art. 24d), by implementing the CDSM [7]. The CDSM characterizes that the provisions are stipulated in the different articles depending on the purpose of use.

2.3. The U.S. Copyright Law: Fair Use

There is no explicit specific provision in the U.S. Copyright Law regarding exceptions and limitations only for data mining. Instead, in a series of cases involving digital technologies, the federal courts have held that fair use permits conducting computational analysis and creating digital archives to enable search services. If the use of the copyrighted works is fair, then the user does not need to seek or receive permission from the copyright holders to use the works [8]. Whether such ‘fair use’ exists involves a case-by-case determination using four statutorily provided non-exclusive factors in light of copyright purposes [9]. The following examines factors to fairly use the works for data mining by analyzing four factors through the court’s findings that conducting computational analysis is fair use.

The first factor is the purpose and nature of the use of works, including whether such use is of a commercial nature or is for nonprofit educational purposes. The courts have ruled that the extraction from searchable databases or search engines is very transformative and is likely to be fair use. In Authors Guild v. HathiTrust, transformative work is one that serves a new and different function from the original work and is not a substitute for it [10]. In Perfect 10, Inc. v. Amazon.com, Inc., the court held that the use of copyrighted thumbnail images in internet search results was transformative because the thumbnail copies served a different function from the original copyrighted images [11]. In Kelly v. Arriba Soft Corp, the Ninth Circuit held that systematic and institutional copying of images for the transformative purpose of providing a commercial image
search service is fair use [12]. And also in Authors Guild v. Google, the court ruled that Google's systematic and institutional copying of books to provide a full-text search that yields snippets of text containing the search term(s) is fair use [13]. Transformative use means conveying a different meaning, transformative in that it adds something new, with a further purpose or different character, altering the first with a new meaning or message, rather than merely superseding the original work [14]. This factor arises when a secondary user makes unauthorized use of copyrighted material to gain a profit through copying the original work without paying the customary price [15]. The more transformative the new work, the less important the commercial purpose [16].

As a result, if data mining adds value to the original without paying any fee to the right holders —if copyrightable expression in the original work is used as raw material, transformed in the creation of new information, new aesthetics, new insights, and understandings, the data mining satisfies this first factor [17]. The above-mentioned cases satisfied this condition so the court held that the use of works regarding computational analysis is transformative.

The second factor is the nature of the copyrighted work. This factor calls for the 'recognition' that some works are closer to the core of intended copyright protection than others [18]. In case secondary use fairly includes copyright-protected works, the court has generally judged that it is more advantageous to the right holder. However, if original works are not recognizable in secondary use despite the use of copyrighted works, the secondary use may fall within fair use [19]. In Authors Guild, Inc. v. Hathitrust, the use of works for a full-text search of books does not allow users to view any portion of the books they are searching [20]. When it comes to data mining, this factor is more advantageous to users because it is difficult to recognize the original work from information abstracted by data mining.

The third factor considers the amount and substantiality of the proportion used in relation to the copyrighted work as a whole. The third factor asks whether the secondary use employs more of the copyrighted work than is necessary [21]. Copying an entire work militates against a finding of fair use, but the courts have judged that the third-factor element is neutral as far as computational analysis is concerned [22]. If TDM does not involve copying the full text or entire datasets, researchers will not be able to extract valuable information by analyzing the contents necessary for research or the AI-related sector. In these terms, this factor is correlated to the second factor. If original works are not recognizable in data mining abstraction despite copying the entire copyrighted works, the use for data mining may fall within fair use [23]. In Authors Guild, Inc. v. Hathitrust, this service doesn’t provide users with any new, human-readable copies. The court concluded that this is fair use [24].

The fourth factor considers the effect of the use upon the potential market for or value of the copyrighted work. Potential harm to be considered encompasses not only that which usurps the demand for the original market but also harm to markets for derivative works [25]. The fourth factor focuses solely on "the harm that results because the secondary use serves as a substitute for the original work" [26]. However, if TDM as the secondary use is transformative and it is not recognizable of original works, this use is not the substitute for the original works. A transformative use diminishes the role of the fourth factor in that the more the copying is done to achieve a purpose that differs from the purpose of the original, the less likely it is that the copy will serve as a satisfactory substitute for the original [27].

As a result of the close linkage between the first and fourth factors, any economic 'harm' caused by transformative uses does not count unless such uses serve as substitutes for the original work" [28].

In conclusion, if data mining adds value to the original despite being free of charge and serves a different function from the original, the uses of works for data mining are transformative [29]. Also, if the information resulting from data mining is transformative and unrecognizable despite copying of entire copyrighted works,
the information is unlikely to be a substitute for the original works and also unlikely to supersede the potential market for original works [30]. When all four statutory factors are explored together, data mining amounts to the fair use. In addition, the circumvention of TPM for data mining is not disadvantageous to fair use(17U.S.C. § 1201(c)).

2.4. The Japanese Copyright Law Art. 30-4

The Japanese TDM copyright exception was first introduced in 2009 (Art. 47-7 before the enactment of the 2018 Amendment) and revised by the 2018 Amendment (Art. 30-4(ii)), through which the requirement of ‘by using a computer’ was deleted and the expression ‘to exploit by any means’ was added [31]. The Japanese copyright law of 2018 explicitly stipulates the exception clause for TDM in Art. 30-4.

Art. 30-4 allows users to exploit works smoothly without permission from right holders in cases where such exploitations are not for ‘enjoying’ the ideas or emotions expressed in the copyrighted works. Art. 30-4 enumerates the kinds of uses not to enjoy the ideas or emotions expressed in the works as the following; a) Experiments for technological development (Art. 30-4(i)), b) TDM(Art. 30-4(ii)), and c) other exploitations without perceiving by human senses [32].

Article 30-4 (i) permits the use of work necessary for experiments for the development and utilization of technology. For example, when a company researches or develops high technology used for a movie player, it would be helpful for the company to experimentally copy a cinematographic work in order to evaluate the quality of that technology.

Article 30-4(ii) permits the use of all works for TDM. Article 30-4 (ii) allows users to use all copyrighted works in any way to the extent deemed necessary unless such exploitation unreasonably prejudices the copyright holder’s interests regardless of commercial or non-commercial purposes for TDM [33]. Article 30-4(iii) permits the exploitation of works that do not involve perceiving expressions in the work through human senses, such as exploitation in the process of computer data processing.

The key point of Art. 30-4 is the definition of ‘enjoyment’. Here the ‘enjoyment’ is similar to ‘recognition’, the requirement of the second factor of fair use in the US copyright law. The copyright works are permissible where the original works are not recognizable from secondary uses in the ways enumerated in Art. 30-4.

Like the requirement of the second factor under the fair use principle, TDM does not allow users to view or enjoy the original works from the information resulting from TDM. Accordingly, Art. 30-4 does not apply, for example, to an experimental screening of a cinematographic work to the public in order to evaluate the effect of the projection technology, even if the IT company researches and develops technology for an impressive projection to be used in cinemas. This is because the public enjoys or recognizes the original works through watching the movie, even if the screening is conducted primarily for experimental purposes [34].

The Japanese copyright Act Art. 30-4 permits comprehensive TDM by anyone – research organizations or business companies regardless of commercial or non-commercial purpose unless it prejudices to the copyright holders’ interest.
2.5. The Comparative Summary

The CDSM Directive stipulates two articles to distinguish the use of works by research organizations (Art.3) from anyone’s use (Art.4). Art.3 is mandatory, and the use of works is not reserved by copyright holders. Therefore, ‘opt-out’ with TPM and ‘contract-out’ are not allowed. On the other hand, Art.4 allows copyright holders to reserve the use of the copyright works with TPM and contract-out. Next, the U.S. copyright law determines whether data mining falls under fair use by exploring all four statutory factors together. If data mining serves a different function or purpose from the original by adding value to the original works, the uses of works for data mining are transformative even if the users do not pay for the use of the works. Also, if original works are unrecognizable from abstracted information through data mining despite copying of entire copyrighted works, the information is unlikely to be a substitute for the original works and unlikely to supersede the potential market for original works. The circumvention of TPM for data mining is allowed for fair use. The Japanese copyright Act Art. 30-4 allows the use of any kind of copyright works for anyone’s data mining regardless of commercial or non-commercial purposes unless it prejudices the copyright holders’ interest by allowing users to enjoy the original works.

3. Amendment Bill to the Korean Copyright Act for Data mining

3.1. Cases

In the series of cases over the last few years, courts have ruled that the defendant’s crawling act unreasonably prejudiced the plaintiff’s interests by copying and posting information from the plaintiff’s website [35]. The courts have ruled that data mining by crawling infringes on the rights of database producers. The Korean Copyright Act provides that database producers shall own the rights to reproduce, distribute, broadcast, or interactively transmit the whole or considerable parts of the relevant database (Art. 93(1)). Provided the copying of their considerable parts of the database conflicts with the normal exploitation of the relevant database, or considerable parts of the database are copied repeatedly or systematically for specific purposes, the copying of the database infringes unduly on the interests of database producers(Art. 93(2)).

However, in 2022, the criminal case Yanolja v. Yogi-eottae, the Supreme Court ruled that the defendant’s crawling act does not only amount to the copying of considerable parts of the database which conflicts with the normal exploitation of the relevant database, but also considerable parts of the database were copied repeatedly or systematically for specific purposes [36]. As a result, the defendant’s copying of plaintiff’s database does not prejudice to the interests of database producers(Art. 93(2)) and it does not fall on the infringement of copyrights. In addition, the court ruled that there was no intent that the defendant tried to infringe copyrights.

Despite the similarities in the above cases, the Yanolja case was judged differently from the former cases. However, the reasoning is not specific. When it comes to the continuous occurrence of these similar cases from now on, more specific requirements are necessary.

3.2. Review on the Copyright Bill for data mining

Hereinafter the requirements for copyright restrictions on data mining based on the amendment bill to the Copyright Act, which is currently pending in the National Assembly, will be reviewed. The amendment bill provides the limitations on data mining in its Art. 43.
Article 43 (Reproduction and Transmission for Analysis of Information)
(1) Reproduction or transmission of works is allowed to the necessary extent for the creation of additional information or additional value (extraction of information such as rules, structure, tendency, and correlation, etc.) from a large volume of information including a number of works by applying automated analysis technology of computers if such creation is possible without enjoying ideas or feelings expressed in such works. Provided, that this shall only be allowed if lawful access to the works is available.
(2) Reproductions made in accordance with Paragraph (1) may be kept to the necessary extent required for analysis of information.

According to the amendment bill, data mining is permitted if the work is not used for enjoyment and only to the necessary extent in case that lawful access is available. So, the requirements to use the works for data mining are ‘no enjoyment’, ‘the necessary extent’, and ‘lawful access’.

The first requirement ‘enjoyment’ is the same wording in the Japanese copyright law Art. 30-4. Also, the ‘enjoyment’ is a very similar concept to the second factor of fair use which requires ‘recognition’. The second requirement ‘the necessary extent’ is also the same wording in the Japanese copyright law Art. 30-4. Accordingly, if original works are used beyond the necessary extent, the use will be an infringement of copyright. The third requirement ‘lawful access’ is the same wording as the CDSM Art.3. and Art.4. The ‘lawful access’ means that circumvention of TPM can be unlawful. To encourage data mining for researching or archiving by libraries, educational establishments or museums, or by archives, the CDSM allows the users to exceptions or limitations for TDM by circumventing technical protection measures (Art.7). However, the amendment bill does not have any provisions to allow users to circumvent TPM for data mining. It will interrupt the encouragement to use the works for data mining if the bill is taken into force. The fair use clause Art.35-5 would be more useful than the bill Art. 43.

4. Conclusion

Data mining has broad applications that reach beyond scholarly and scientific research and provide internet search engine services that are commonly used forms of TDM of websites. The exceptions and limitations for data mining provide a competitive advantage in the global race for policy innovation because it permits researchers to conduct computational analysis - TDM on any materials to which they have access. For this purpose, Japan and the EU added limitations on copyright to legalize some TDM research through amendments to copyright law, and the U.S. copyright law has allowed data mining by the fair use provision. As a result of reviewing copyright laws, it was found that each copyright law has different characteristics.

The amendment bill of Korean copyright law also contains a provision on the exceptions and limitations for data mining. In that data mining is permitted if the work is used not to enjoy original works to the necessary extent in case that lawful access is available, the amendment bill of Korean copyright law is very similar to the Japanese copyright law Art.30-4. However, the requirement of ‘lawful access’ would interrupt the encouragement to use the works for data mining if the bill is taken into force. Rather the fair use clause Art.35-5 would be more useful rather than the bill Art. 43. We reviewed that Korea also needs an appropriate provision of exceptions and limitations on copyright which will help to encourage AI-related business by using more data smoothly through the mining process and extracting more valuable information.
References


[6] Ibid.


[10] Authors Guild v. HathiTrust, 755 F.3d 87 (2d Cir. 2014);


[27] Brief of Digital Humanities and Law Scholars as Amici Curiae in Partial Support of Defendants’ Motion for Summary Judgment, Authors Guild, Inc. v. HathiTrust, 755F.3d 87 (2d Cir. 2014) (No. 11 Civ. 06351); Authors
Guild, Inc. v. Google, Inc. (Google Books), 804 F.3d 202, 222-223 (2d. Cir. 2015).

[28] HathiTrust, 755 F.3d at 99 (citing Bill Graham Archives v. Dorling Kindersley Ltd., 448 F.3d 605, 614 (2d Cir. 2006))


[32] Referring to the translation see Japan Copyright Office (n 30), at 11-12.


[36] Yanolja v. Yogi-eottae [Supreme Court 2021Do1433 (Criminal case); Seoul High Court 2021Na2034740 (Civil Case)]