

Research on Fair Use of Copyright in Text and Data Mining

Xuanye Sha

Dalian Ocean University, Dalian, China

Abstract: Text and data mining refers to computer processing behavior that extracts relevant information from texts and data. It mainly copies a large number of original texts, extracts relevant information from them, and then recombines information to obtain potential new and unknown information. Nevertheless, from the processing mode of data mining, certain infringement risks exist, which may result in infringements upon the right of reproduction, the right of adaptation and the right of communication through information network. Moreover, there are certain legal obstacles for text and data mining to invoke the fair use of copyright. Therefore, specific rules for the fair use system shall be designed from three dimensions: broadening the scope of subjects, adopting technical measures for data security protection, and restricting usage from extending to subsequent information dissemination. It is appropriate to incorporate text and data mining into the copyright fair use system. On the basis of learning from extraterritorial experience, China shall construct a conditional exception model for fair use. In addition, in view of deficiencies in China's Copyright Law, targeted fair use circumstances shall be added so as to boost mass innovation.

Keywords: Text and Data Mining, Fair Use, Copyright Infringement.

1. Research Background and Significance

1.1. Research Background

Nowadays, digital technologies have been widely applied, and data has become a core competitive advantage. In the big-data era, data contains abundant scientific-research potential and social value. In-depth exploration and effective utilization of data value constitute an important criterion for measuring national competitiveness. Major approaches to utilizing big data include data mining and data analysis, among which data mining occupies a core position. Models combining big data with artificial-intelligence technologies, especially text and data mining, show tremendous application potential and broad development prospects. Nevertheless, while bringing high productivity, text and data mining, as an emerging technology, also faces conflicts with copyright fair use. China's third revision of the *Copyright Law* has not directly responded to text and data mining, which impedes the development of China's digital economy. Therefore, China shall formulate corresponding provisions for text and data mining on the basis of learning from foreign legislative and judicial practices, and lead technological development through sound institutions and laws.

1.2. Research Significance

The theoretical significance of this paper is as follows: research on the copyright fair-use system for text and data mining is of great importance for promoting the further development of copyright theories in the information age. Along with technological advances, the scope of copyright protection keeps expanding and updating. In-depth analysis of potential risks arising from the application of text and data mining and corresponding suggestions will help enrich and improve the existing theoretical framework, and provide solid legal support for the development and application of emerging technologies.

The practical significance of this paper is as follows: under

the current legal framework, there are no clear norms governing text and data mining. Hence, it is critical to reduce infringement risks in text and data mining and offer legal guidance for legal practitioners and users in advance. By exploring fair-use principles for text and data mining, this paper attempts to propose feasible evaluation criteria for judicial practice and prevent unlimited expansion of the scope of fair use. The goal is to strike a balance between industrial innovation and copyright protection and facilitate their coordinated development.

2. Overview of Text and Data Mining

2.1. Definition of Text and Data Mining

Text and data mining refers to computer-technology-enabled behavior of extracting useful information and knowledge from texts or datasets. In the big-data era, "data" is regarded as a critical strategic resource and factor of production, greatly driving China's economic and social development. At present, the growth rate of various data and information far exceeds human prediction and processing capacity. To make effective use of these precious data resources and resolve the contradiction between abundant data and scarce knowledge, intelligent and automatic analytical processing with computer technologies has become an urgent demand. As a unique information-processing and text-analysis tool, text and data mining plays a vital role in this process.

The United Kingdom revised its *Copyright, Designs and Patents Act 1988* in 2014, and the European Union's *Directive on Copyright in the Digital Single Market* entered into force in 2019. Both instruments adopt the term "Text and Data Mining" and define it in detail. Therefore, most scholars in theoretical circles conduct research on the basis of "Text and Data Mining". Nevertheless, understandings of this term differ greatly due to divergent perspectives and linguistic habits. From the perspective of knowledge discovery, text and data mining generally refers to acts of extracting previously-unknown, hidden yet potentially-useful

information and knowledge from massive texts and data [1].

2.2. Significance of Text and Data Mining

The significance of text and data mining lies in that in-depth analysis and mining of massive texts and data enable people to discover potential value within data, thereby offering robust support for decision-making, research or business applications. News reports, academic papers, social-media content and business documents all contain abundant information. Mining such textual data can uncover new viewpoints, trends, patterns or correlations and enrich knowledge systems. Moreover, in a fiercely-competitive market environment, text and data mining helps enterprises identify competitors' weaknesses, understand customers' real demands and grasp market changes. Text-and-data-mining technologies perform important functions in various sectors. For instance, in meteorology, analysis of meteorological data from diverse sources enables weather forecasting. In the healthcare sector, healthcare systems can adopt text-and-data-mining technologies to analyze patient data from hospitals, sensors, monitors and medical-equipment manufacturers so as to obtain valuable medical insights and suggestions. In academic research, text-and-data-mining technologies help researchers identify new research trends and patterns from massive academic-literature databases. In sales and marketing, by analyzing customers' behavioral habits, preferences and consumption patterns, companies can uncover previously-unknown market trends and deliver more personalized and precise products and services. In the data-resource era, effective extraction and utilization of text and data resources have gradually become catalysts for innovative development and important supports for scientific research. Such information is significant for optimizing products, improving service quality and formulating marketing strategies, and helps enterprises strengthen competitiveness.

3. Problems of Incorporating Text and Data Mining into the Copyright Fair-Use System

3.1. Copyright-Infringement Risks of Text and Data Mining

As an emerging technology, text and data mining targets both public data that can be freely used and works protected by copyright law. When conducting market-trend analysis or consumer-behavior research, people may mine textual data from social media, news reports, forum discussions and other sources. Such data may contain large quantities of original content created by individuals or organizations. In academic research, text-and-data mining is also widely applied in literature review, citation analysis and other fields. In these scenarios, researchers need to mine and analyze large numbers of academic papers and other academic materials, all of which are works protected by copyright law. It is necessary to examine each operational link of text-and-data-mining technologies from the perspective of copyright law and analyze corresponding copyright-infringement risks, so as to provide legal guarantees for technological development.

3.1.1. Potential Infringement of the Right of Reproduction at the Stage of Acquiring Mining Objects

First, text and data mining usually involves reproduction

and processing of massive texts and data. If such reproduction is conducted without explicit authorization from copyright owners, especially for commercial purposes, it may infringe upon copyright owners' right of reproduction. The right of reproduction is a fundamental right conferred upon copyright owners by copyright law, which allows copyright owners to control reproductions of their works.

Second, the right of reproduction is not confined to traditional physical reproduction such as printing and photocopying. Along with technological advances, digital reproduction has also become an important dimension of the right of reproduction. In text and data mining, converting textual content into digital form for processing may also be deemed reproduction. Especially when such digital reproduction exceeds the scope prescribed by fair use under law, it may constitute infringement.

Furthermore, text and data mining may also involve in-depth analysis and processing of copyright-protected works, including feature extraction and model construction. Such acts may involve substantive utilization of works and thus constitute infringement upon copyright owners' right of reproduction.

It shall be noted that determination of infringement upon the right of reproduction shall take multiple factors into account, including specific legal provisions, purposes of utilization, authorization status and copyright owners' will. Under certain circumstances, if text and data mining is conducted for non-profit purposes such as research and education and meets statutory fair-use criteria, it may not be deemed infringement. Nevertheless, for-profit text and data mining without explicit authorization will most likely constitute infringement upon the right of reproduction.

3.1.2. Potential Infringement of the Right of Adaptation at the Pre-processing Stage of Mining Objects

China's *Copyright Law* does not adopt the term "right of adaptation". Academically, the "right of adaptation" is a general concept covering the right of cinematography, the right of adaptation, the right of translation, the right of compilation and other rights. In text and data mining, especially automatic retrieval, extraction and analysis of source documents, substantive utilization or adaptation of original works may be involved. For example, at the pre-processing link of text and data mining, unstructured original textual content may need to be converted into structured content or intermediate-data formats for further analysis. Such conversion may involve substantive adaptation or re-creation of original works and may thus infringe upon original authors' right of adaptation. Another viewpoint holds that converting pre-processed content into unified structured data does not alter the expression of sample works, nor does it change sample works' creative thinking, which bears similarities to translation acts [2]. In text-based machine translation and data mining, applications require massive texts as input for translation into machine-recognizable languages. Upon receiving human instructions, algorithmic programs may generate translations. It can be seen that machine translation involves acts such as translation and compilation of works [3]. The "transcoding" act in text and data mining needs to modify the mode of expression of original works so as to form new research samples. The degree of such transcoding varies in research on these samples. If only minor transcoding occurs, the text-and-data-mining program may only infringe upon copyright owners' right of reproduction. In text and data

mining, if transcoded works possess new original expressions, such transcoding acts will fully satisfy constituent elements for infringement upon copyright owners' right of adaptation.

3.1.3. Potential Infringement of the Right of Communication through Information Network at the Stage of Mining-result Presentation

At the presentation stage, text-and-data-mining technologies generate analytical results which may take the form of academic reports or business-activity predictions and ultimately become visible. According to the "access plus substantial similarity" criterion for copyright infringement, if mining outputs constitute brand-new works or only contain minimal information from original works, no infringement is established. On the contrary, if outputs are highly similar to original works, infringement may be constituted. In practical application, users of text and data mining usually need to share their research outcomes via the internet, and such outcomes may involve expressions of copyright-protected original works. For instance, Google Books Project allowed the public to browse book snippets via keyword search. Such online disclosure of book content without copyright-owner authorization may trigger copyright issues and may infringe upon the right of communication through information network.

3.2. Legal Obstacles for Invoking Copyright Fair Use in Text and Data Mining

If works can be fairly used in text and data mining, the most direct legal-application obstacle lies in compliance with the 12 circumstances enumerated in Paragraph 1 of Article 24 of China's *Copyright Law* or the catch-all clause. Nevertheless, the existing fair-use system cannot cover all text-and-data-mining scenarios and restrains its application to a certain extent. Text and data mining still faces infringement risks, which are specifically analyzed as follows.

3.2.1. Text and Data Mining Can Hardly Satisfy the "Personal Use" Clause

China's *Copyright Law* prescribes the fair-use circumstance of "personal use"[4] and clarifies its applicable conditions: first, subjects are limited to natural persons, excluding legal persons and other organizations; second, it must serve non-profit personal learning, research or appreciation purposes; third, personal use does not mean unlimited use of others' works; it shall be small-scale and within a controllable scope. In reality, at least at present, this technology functions more as a tool, and actual subjects are persons conducting mining. From miners' perspective, their obligations and responsibilities determine that text and data mining is not entirely non-profit. On the one hand, university research institutions may obtain breakthrough research outcomes via text-and-data-mining technologies. On the other hand, enterprises or institutions in commercial sectors may adopt such technologies to acquire direct or indirect economic benefits. Therefore, text and data mining may generate profits under specific circumstances. In addition, text and data mining involves massive texts and data and usually requires full-text or large-fragment utilization of works. This is not fully consistent with the "personal-use" clause under the *Copyright Law*, as personal use generally refers to small-scale and non-commercial utilization.

3.2.2. Text and Data Mining Can Hardly Conform to the Circumstance of "Appropriate Quotation"

China's *Copyright Law* stipulates that for the purpose of

introducing or commenting on a work or illustrating a problem, appropriate quotation of others' published works is permitted. The circumstance of appropriate quotation includes appropriately quoting published works in works provided via information networks. The "appropriate-quotation" clause sets restrictions on subjective intent and quotation scope. Its precondition is that quotation is necessary for creating commentaries or academic writings. The core of "appropriate quotation" lies in control of the "degree", namely quoting others' works in small quantities as required in creation to satisfy the requirement of appropriateness. Nevertheless, to guarantee comprehensiveness and accuracy of knowledge-discovery outcomes, text and data mining generally requires all relevant texts and data on given subjects. Such massive quantity exceeds human expectations and handling capacity and completely fails to meet the "appropriateness" standard. Moreover, at the data-mining and result-output stages, part or even most of acquired works may be involved. Hence, text and data mining may fail to satisfy provisions governing "appropriate quotation".

3.2.3. Text and Data Mining Fails to Meet the Circumstance of "Teaching and Scientific Research Use"

China's *Copyright Law* provides that "for the purpose of classroom teaching or scientific research in schools, translation, adaptation, compilation, performance or limited reproduction of published works may be conducted for use by teaching or research personnel, provided that such works shall not be published or distributed". When text and data mining serves classroom-teaching or scientific-research purposes, a small number of published articles may constitute fair use. In current research fields, a considerable share of text-and-data-mining activities are indeed conducted for research purposes. However, fully applying this principle to research-oriented text and data mining faces difficulties. In practice, not only research institutions adopt text-and-data-mining technologies; a growing number of other entities also participate, including corporate product development and governmental public services. Under the current copyright law, when an artificial-intelligence-R&D company conducts research activities with text-and-data-mining technologies, it can hardly invoke this provision to defend against claims from work copyright holders in law. Second, compared with "personal use", "teaching-and-research use" sets clearer limits for the right of reproduction: only limited reproduction is permitted. Nevertheless, text-and-data-mining technologies are built upon massive work data. This feature determines that it can hardly satisfy the requirement for "limited" use of others' works. Finally, restrictions under this clause mainly target staff members of educational and research institutions. In current circumstances, however, this technology does grant the public maximum rights to reproduce works. Therefore, text and data mining can hardly satisfy provisions governing "teaching-and-research use".

4. Appropriateness of Incorporating Text and Data Mining into the Copyright Fair-Use System

4.1. Theoretical Analysis of Incorporating Text and Data Mining into the Copyright Fair-Use System

4.1.1. Compliance with the Legislative Purpose of the Copyright Law

As an important technology for data processing and information extraction, text and data mining complies with the legislative purpose of China's *Copyright Law*. First, as a branch of data mining, text mining mainly focuses on mining valuable and previously-unknown information and knowledge from massive textual information. This process requires in-depth analysis and processing of textual data to uncover hidden patterns, trends and valuable insights. Furthermore, the *Copyright Law* protects copyrights of creators of literary, artistic and scientific works and actively facilitates creation and dissemination of socialist spiritual and material civilization. It can thus be seen that this law aims to protect creators' rights and interests, encourage creation and sharing, and thereby boost vigorous development of culture, science and technology. Combining these two dimensions, practices of text and data mining to some extent promote the development of culture and science. By mining textual data, people can discover new knowledge and new viewpoints and even inspire new creation. Such knowledge extraction and dissemination is exactly what the legislative purpose of China's *Copyright Law* advocates. In addition, text-and-data-mining practices also embody respect for creators' rights and interests. When conducting text mining, relevant parties usually abide by relevant copyright provisions, ensure lawfully-acquired textual data and respect original authors' rights and interests while utilizing such data. Such copyright-compliant acts actually safeguard creators' rights and interests and align with the legislative purpose of the *Copyright Law*.

4.1.2. Meeting the Requirements of the Three-Step Test

The Three-Step Test is a standard for judging whether an act constitutes fair use. Widely adopted in international copyright circles, this standard has been incorporated into multiple international copyright conventions [5] and has been written into Article 24 of China's *Copyright Law*. Courts shall take the Three-Step Test into consideration when rendering relevant judgments.

The first step requires special circumstances subject to 13 enumerated circumstances. Specific provisions on fair use for text and data mining may be added to administrative regulations to bring it within general-protection scope and satisfy requirements of the first step. The second step is to determine whether utilization conflicts with normal exploitation of works. This provision is highly abstract. In practice, judges may hold divergent views on the quantity, mode and scope of work utilization as well as whether quoted works establish a substitution relationship with original works. In text and data mining, operations are predefined by algorithms, and computers execute input, analysis and output according to programs, avoiding ambiguities that easily arise in practice. Since the technology aims to obtain unknown or poorly-understood knowledge or information, final analytical content is based on highly-detailed texts and side data. Content related to original works may only consist of

keywords or other fragmentary information instead of establishing a substitution relationship with original works. The third step is to determine whether legitimate interests of rights-holders suffer unreasonable prejudice. Judges shall judge whether legitimate interests of original copyright holders are impaired by combining market responses and benefits obtained by copyright owners. When developing systems for fair-use of text-and-data-extraction technologies, so long as such systems follow the interest-balance principle, maximize protection of original copyright holders' legitimate interests and promote innovation and development of science and culture, compliance with requirements of the third step may be affirmed.

According to the Three-Step Test standard, three constituent elements shall be satisfied: first, acts shall be conducted under specific circumstances; second, acts shall not conflict with normal exploitation of works; third, acts shall not cause unreasonable prejudice to legitimate rights of rights-holders. Text-and-data-mining acts generally satisfy requirements of the Three-Step Test. On the one hand, text-and-data-mining acts are completely different from traditional modes of work utilization. Their purpose is to process and analyze massive interrelated works so as to discover new knowledge. They are not tools for reading works; on the contrary, they add value to some works and enable creation of new information, new aesthetics and new understandings. Therefore, they basically do not conflict with normal exploitation of works. On the other hand, reproductions of works generated in text and data mining shall not be permitted to be sold or disseminated to the public. Hence, they cannot replace original works in the market and will not exert excessive negative impacts on market value of original works and legitimate interests of rights-holders. They may even boost sales volume and public awareness of original works. To sum up, text-and-data-mining acts generally satisfy standards of the Three-Step Test. Accordingly, incorporating text and data mining into the copyright fair-use system in China will not conflict with the Three-Step Test.

4.2. Extraterritorial Experience of the Copyright Fair-Use System for Text and Data Mining

4.2.1. Practical Model of "Unconditional Exception"

From a pragmatist perspective, the United States and Japan standardize text and data mining as part of industrial policies, thereby emphasizing its economic importance. The United States and Japan support granting miners relatively broad authority to ensure smooth development of text-and-data-mining activities. They classify such activities as fair use to the largest possible extent and impose relatively few restrictions. This practice is termed the "unconditional-exception" model. Two landmark cases incorporating text and data mining into fair use in U.S. judicial practice are the Google Books case and *Authors Guild, Inc. v. HathiTrust* (hereinafter referred to as the HDL case)[6].

Since 2004, Google has expanded its library-cooperation component under the Google Books Project and signed cooperation agreements with multiple libraries. Under such agreements, Google conducts comprehensive digital scanning of books collected by these libraries and provides scanned electronic-book copies to respective libraries for their digital-management and utilization. In addition, Google generates profits by promoting such book content. The public can now search for interested books and relevant information

via Google's search platform using keywords. Nevertheless, this project met fierce opposition from the Authors Guild of America and individual authors, who filed lawsuits. In 2015, U.S. courts ruled that Google's acts fell within the scope of fair use [7]. In this case, courts thoroughly considered four factors for fair use, paid special attention to the first factor, namely purpose and character of use, and applied the transformative-use theory to affirm that relevant text-and-data-mining acts constituted fair use. Keyword-search scenarios were deemed as use that altered original works' purposes. This greatly reduced the weight of the most-controversial market-impact factor in adjudication. However, there was little dispute over the second and third factors. This is mainly because the nature of involved works was necessary for mining purposes, and large-scale full-text reproduction was required to realize mining objectives. Moreover, in this case, the corporate mode of use did not disclose main content of works to the public.

In the HDL case, HathiTrust (HDL), a non-profit organization initiated by multiple American universities, signed a cooperation agreement with Google. Under this agreement, more than ten-million electronic-work versions provided by Google were stored in HDL's database. Based on these electronic works, HDL offered three types of services including text-and-data-mining technologies. With text and data mining, users could retrieve information about specific words in literature via keyword search without reading full texts. The judgment in this case was based on courts' comprehensive consideration of the four-factor principle. Special emphasis was laid on the first factor — purpose and character of use. By applying the transformative-use theory, courts affirmed that relevant data-mining activities constituted fair use. In this case, provision of adjacent fragments of works was deemed alteration of modes of utilization for original works, which greatly reduced the weight of the most-controversial market-impact factor in adjudication. However, there was little dispute over the second and third factors, mainly because the nature of involved works was necessary for mining purposes and large-scale full-text reproduction was required to realize mining objectives. Furthermore, the corporate mode of use did not disclose main content of works to the public. Concerning provision of adjacent fragments, courts pointed out that copyright law did not confer upon copyright owners the right to provide work-related information. Hence, users did not need to obtain authorization even if plaintiffs claimed authorization was required.

Additionally, in the TVEyes case, the U.S. Court of Appeals for the Second Circuit ruled that providing ten-minute video clips of television works for customers exceeded the scope of fair use. When considering reproduction and archiving acts required for search services, courts deemed such acts fair use under copyright law. This judgment highlights the importance of presentation modes for outputs after text and data mining. Even though reproduction, adaptation or dissemination involved in mining processes may be deemed fair use, users conducting text and data mining still need to adopt necessary measures. Such measures shall ensure that outputs do not replace market positions of original works so as to avoid copyright-infringement risks. This case defines concrete boundaries for fair use of text and data mining under U.S. copyright law from detailed dimensions and bears certain flexibility.

4.2.2. Practical Model of “Conditional Exception”

In the digital age, text and data mining has become a new engine driving scientific-technological progress. As a product combining big-data-analysis and artificial-intelligence technologies, it can conduct in-depth analysis of massive digital information and demonstrates tremendous application potential. At present, a country's application-and-development level in science and technology is increasingly regarded as a core indicator measuring its international competitiveness. The European Union expresses concerns over legal risks that may arise from text-and-data-mining technologies. It believes inappropriate policies may restrain progress of new technologies and cause Europe to lag behind other countries in research-and-development of this field. To facilitate future development of the digital economy, the European Union calls for moderate adjustment of existing copyright statutes to adapt to technological advances.

Against this backdrop, the European Commission put forward a draft *Copyright Directive* in September 2016, whose Article 3 stipulated specific exceptions for scientific research and data mining. Subsequently, the European Parliament revised this draft in September 2018. Legislative intent was to broaden the scope of application for this exception clause to cover commercial-sector entities. In April 2019, the European Union formally adopted the *Copyright Directive*, whose Articles 3 and 4 elaborate upon these exception provisions. Specifically, Article 3 is a mandatory clause which prohibits exclusion of its application for research organizations and public-sector bodies via contractual agreements. It clearly defines applicable subjects as research institutions and public-sector bodies and excludes commercial entities. Meanwhile, it requires applicable subjects to obtain to-be-mined content via legal channels and adopt appropriate security measures to protect works from disclosure so as to avoid unnecessary prejudice to copyright owners. Compared with the mandatory nature of Article 3, Article 4 grants greater options to copyright owners. Copyright owners have the right to refuse in advance utilization of their works for text and data mining for non-research purposes and reserve rights to reject private entities' reproduction and extraction of their works. This means that if miners intend to conduct text and data mining for commercial purposes, they shall abide by restrictions set by copyright owners; otherwise, they will face potential infringement risks. Overall, these text-and-data-mining exception clauses of the European Union provide institutional guarantees for development of the digital economy and strike a balance between copyright protection and technological innovation.

5. Suggestions on Incorporating Text and Data Mining into the Copyright Fair-Use System

5.1. Constructing a Conditional-Exception Model for Fair Use

To promote rapid growth of emerging-technology industries, China urgently needs to renovate its existing copyright-law system so as to better cope with work utilization in the application of big-data and intelligent technologies. This poses new challenges for China's copyright fair-use system. Incorporating

text-and-data-mining technologies into fair use is both necessary and feasible. Through in-depth comparison and analysis of legal systems for fair use of text and data mining across different countries worldwide, we can find that there are two major incorporation strategies commonly adopted internationally. The first is to follow the “unconditional-exception” model. Its core practice is to introduce general judgment criteria for copyright fair use without imposing restrictions on specific utilization. The second is to learn from the “conditional-exception” model, which adds special exception clauses and limits utilization to non-commercial uses. For the first model, although Chinese judicial practice has witnessed cases starting to apply general clauses to judge fair use, considering that China is not a case-law country, complete reliance on judges to judge rationality of data-mining acts according to general clauses may fail to deliver sufficient legal guarantees for sound development of text-and-data-mining technologies. Furthermore, general clauses for fair use are rather principled, granting judges relatively broad discretion in concrete cases, which may exert adverse impacts on the existing copyright-interest balance. For the second model, special exception clauses targeting text and data mining are added through legislative channels to give explicit legal responses to such utilization acts. Compared with the first model, the second model is more conducive to ensuring legal exemption for fair use of text and data mining and reducing uncertainties in defending against infringement by invoking principled clauses in practice. Therefore, formulating special exception clauses to construct and improve China’s copyright fair-use system constitutes a more appropriate choice.

When constructing the fair-use system for text and data mining, China shall fully consider its existing legal framework for fair use and explore a path that conforms to China’s legislative reality while featuring Chinese characteristics. Hence, this paper proposes adding special fair-use clauses targeting text and data mining in legislation. Such clauses shall explicitly affirm that work utilization aiming to create new knowledge, including text-and-data-mining acts, shall be deemed fair use. Meanwhile, these special clauses shall be combined with general fair-use principles to evaluate and judge concrete work-utilization acts in practice. In this way, while protecting copyright, necessary legal space can be reserved for development of new technologies such as text and data mining to adapt to demands of the information age. This will facilitate theoretical innovation in copyright and provide solid legal foundations for application of new technologies.

According to patterns of law revision in China, neither the *Copyright Law* nor its *Implementing Regulations* are revised frequently. The *Copyright Law* completed its third revision in 2020 and generally will not be revised again in the short term. Therefore, text-and-data-mining acts may be added as fair-use provisions when the *Implementing Regulations of the Copyright Law* are revised next time.

5.2. Design of Specific Rules for Incorporating Text and Data Mining into the Copyright Fair-Use System

5.2.1. Broadening the Scope of Mining Subjects

First, an overly-narrow scope of subjects in text and data mining can neither promote development of text-and-data-mining technologies nor encourage innovation. More inclusive provisions shall be formulated to bring more

institutions and individuals within the scope of mining subjects. For example, besides research institutions and cultural-heritage institutions, commercial-research institutions, educational institutions, public libraries and other entities may also be incorporated. When engaging in text and data mining, these institutions and individuals may also utilize works for rational purposes such as scientific research and education. Copyright owners may sign licensing agreements with relevant institutions or individuals to permit them to conduct text and data mining under specific conditions. Such licensing mechanisms can protect copyright owners’ rights and interests while offering more subjects opportunities for work utilization. Restrictions on subjects adopted by the European Union and other jurisdictions bear certain rationality. Limiting subject scope excludes commercial-research subjects and helps safeguard rights and interests of rights-holders. Nevertheless, no organization can be confined to only one type of activity. For instance, research institutions may also generate profits, while commercial institutions may also conduct public-welfare activities. Therefore, subjects shall not be restricted, and more actors shall be allowed to participate in research and application of text and data mining. This conforms to the original intention for setting exception clauses for text and data mining.

5.2.2. Adopting Technical Measures to Protect Data Security

Copies generated in data-mining processes, including copies stored in external storage devices and temporary copies in memory, face risks of leakage and dissemination. This is especially true given potentially long data-mining cycles and workflows. Therefore, to ensure rationality of reproduction acts in data mining, effective technical measures shall be adopted to protect security of these data copies and prevent prejudice to copyright owners’ interests caused by data leakage. Secure storage of data copies constitutes the key to legitimacy of temporary reproduction acts in data mining. Legislation in many countries sets explicit provisions on storage of data copies. For example, the European Union stipulates that copies of works generated for data mining shall be stored with an appropriate security level; otherwise, infringement will be constituted. Germany has similar provisions requiring appropriate security precautions for copies generated based on data mining to prevent unauthorized access. In the *Authors Guild v. Google* case in the United States, Google demonstrated its rigorous technical preservation measures to protect all digital copies of books from leakage and dissemination. Ultimately, Google’s unauthorized reproduction acts were affirmed as non-infringing. This case highlights the importance of adopting appropriate technical measures to protect data copies and the role of such measures in judging whether reproduction acts in data mining constitute fair use.

5.2.3. Restricting Usage from Extending to Subsequent Dissemination of Copied Works

To establish text and data mining as an exception to copyright infringement, concrete conditions for work utilization in text and data mining shall be clearly stipulated. This paper proposes adding an exception clause when China revises the *Implementing Regulations of the Copyright Law*, permitting reproduction of others’ works for text and data mining in academic-research processes. Under such circumstances, reproduced works shall be strictly confined to utilization in execution of text and data mining. If reproduced works are put to subsequent commercial uses, or fees are

charged for full-text browsing and downloading services for works in databases, such acts will exceed the scope of infringement exceptions and require explicit authorization from original authors or rights-holders. In the Google case, book-retrieval functions were confined to academic spheres, and ordinary readers were only permitted to read small snippets for free. Such “transformative use” was affirmed as fair use because it resembled acts generally accepted by rights-holders in the print-book era, namely readers browsing books in libraries or bookstores. If services provide complete copies of books, it will be difficult to avoid infringement liability.

6. Conclusion

Text and data mining is an important product arising from development of the big-data industry and scientific innovation. It boasts broad application prospects and great value. Nevertheless, text-and-data-mining acts also face substantial infringement risks, including infringements upon the right of reproduction, the right of adaptation and the right of communication through information network. To address new challenges for fair use, various countries have introduced diverse policies to regulate text and data mining, including the unconditional-exception model represented by the United States and Japan and the conditional-exception model represented by the European Union. Text-and-data-mining acts possess legitimacy. Hence, it is necessary to construct China’s conditional-exception model for text and data mining on the basis of extraterritorial models, add targeted fair-use clauses, and combine these special clauses with general

fair-use principles to evaluate and judge concrete work-utilization acts in practice. Finally, exploration shall be conducted to design specific rules for incorporating text and data mining into the copyright fair-use system, including broadening subject scope, adopting technical measures for data protection, and restricting usage from extending to subsequent dissemination of works.

References

- [1] Cai, T. T. (2022). Research on copyright fair use of text and data mining (Master’s thesis). East China University of Political Science and Law.
- [2] Dong, F., & Guan, Y. H. (2019). On the construction of copyright exception rules for the application of text and data mining technologies. *Hebei Law Science*.
- [3] Dong, X. T. (2022). Research on the copyright fair-use system of text and data mining (Master’s thesis). Hebei University of Economics and Business.
- [4] Copyright Law of the People’s Republic of China. Article 24.
- [5] Berne Convention for the Protection of Literary and Artistic Works.
- [6] WIPO Copyright Treaty.
- [7] Agreement on Trade-Related Aspects of Intellectual Property Rights.
- [8] Authors Guild, Inc. v. HathiTrust, 755 F.3d 87 (2d Cir. 2014).
- [9] Authors Guild v. Google, 770 F. Supp. 2d 666 (S.D.N.Y. 2011).