Research on Designing Fair Use Rules for Text and Data Mining in the AI Era: Postprint

Authors: Wu Gao, Huang Xiaobin

Date: 2023-04-01T16:02:58+00:00

Abstract

[Purpose/Significance] Based on a comparative analysis of legislative and judicial responses to TDM practices in the United States, United Kingdom, European Union, Japan, Germany, and other jurisdictions, this study proposes a design framework for China's TDM fair use rules. [Method/Process] The study first analyzes the technical characteristics and challenges of text and data mining in the artificial intelligence era, introduces the divergent positions of publishers and libraries regarding TDM fair use, and subsequently, based on a comparative analysis of legal response mechanisms for TDM fair use in representative countries, elaborates on the specific content of China's TDM fair use rules from the perspectives of subject, object, purpose, act, and other conditions for TDM exceptions. [Results/Conclusion] Any institution or individual conducting text and data mining on any work through reproduction, extraction, adaptation, or limited dissemination for scientific research or other reasonable purposes constitutes fair use, provided that the utilization of works from non-legitimate sources requires payment of reasonable compensation.

Full Text

Preamble

Title: Study on the Design of Fair Use Rules for Text and Data Mining in the Age of Artificial Intelligence

Authors: Wu Gao[1],[2], Huang Xiaobin[2]
[1] Guangxi Normal University School of Law, Guilin 541004
[2] Sun Yat-sen University School of Information Management, Guangzhou 510006

Abstract: [Purpose/Significance] Based on a comparative analysis of legislative and judicial responses to TDM behavior in the United States, United Kingdom, European Union, Japan, and Germany, this paper proposes design ideas for

China's TDM fair use rules. [Method/Process] The paper first analyzes the technical characteristics and challenges of text and data mining in the AI era, introduces the divergent positions of publishers and libraries regarding TDM fair use, and then—based on comparative analysis of legal response mechanisms for TDM fair use in representative countries—demonstrates the specific content of China's TDM fair use rules from perspectives including the subject, object, purpose, behavior, and other conditions of TDM exceptions. [Result/Conclusion] Text and data mining conducted by any institution or individual through reproduction, extraction, adaptation, or limited dissemination of any work for scientific research or other reasonable purposes constitutes fair use, though reasonable royalties must be paid for utilizing works from non-legitimate sources.

Keywords: artificial intelligence; text and data mining; TDM; fair use; exception
Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2021.22.001

Artificial intelligence is regarded as one of the three cutting-edge technologies of the 21st century (alongside genetic engineering and nanoscience), with many countries elevating it to a national strategic priority through dedicated planning documents. Humanity has now transitioned from the information age into a data-driven "intelligent era" where artificial intelligence requires big data learning to establish its intelligence. Text and data mining (TDM) serves as a crucial supporting technology for numerous interdisciplinary fields and forms the foundation for applications such as artificial intelligence, blockchain, and cloud services. It holds significant value not only for accelerating scientific discovery, driving technological innovation, creating new business models, and promoting economic development, but also for public cultural sectors, providing robust technical support for libraries and other institutions to offer personalized intelligent services based on TDM. However, since TDM technology likely involves reproducing, extracting, reconstructing, and disseminating copyright-protected works or database content, whether such behavior constitutes lawful activity remains a focal point of debate in both theoretical and practical circles. Most countries and regions have not yet effectively established TDM fair use rules, with only a few such as the EU, UK, Japan, and the US providing legislative and judicial responses. According to CNKI database statistics, domestic literature on TDM copyright issues is scarce, and existing theoretical research tends to focus on introducing legislative experiences from select countries, with few systematic analyses of the specific content of China's TDM fair use rules. The development of new technology inevitably disrupts the traditional interest balance mechanism of copyright law, making it essential to reconstruct this balance under new technological conditions and build new TDM fair use rules adapted to China's national conditions to promote scientific innovation and economic development.

1 Technical Characteristics and Challenges of TDM in the AI Era

1.1 Technical Characteristics of TDM in the AI Era

TDM generally refers to the process of extracting data, organizing information, and discovering knowledge from large volumes of text or data. The 2014 UK Copyright Act defines TDM as using automated analytical techniques to analyze text and data to obtain patterns, trends, and other useful information [1]. The European Parliament's March 26, 2019 Directive on Copyright in the Digital Single Market [2] defines TDM as any automated analytical technique aimed at analyzing text and data in digital format to generate information including but not limited to models, trends, and correlations. TDM exhibits the following technical characteristics:

1.1.1 Subject Specialization

Implementing TDM requires not only developing mining algorithms but also complex processes including data preparation and management, data preprocessing and transformation, algorithm debugging and application, and result analysis and presentation. Individuals generally lack the conditions and capacity to reproduce, extract, process, compare, and analyze massive amounts of text and data. Implementers are typically organizations with certain technical and material resources, such as libraries, universities, enterprises, or other research institutions.

1.1.2 Object Extensiveness

Any digital material—including text, images, video, audio, and data—can become a target for data mining. Since the more materials relevant to a research topic are obtained, the more generalizable and accurate the conclusions drawn through mining technology analysis, TDM adopts a "sample = population" full-data model to acquire all relevant materials to the greatest extent possible.

1.1.3 Process Transformative Nature

The transformative nature of the process refers both to data format conversion—where various types of data (including unstructured, semi-structured, and structured data) must be converted into structured data processable by computers—and to the inevitable reproduction of copyright-protected text data during the mining process. However, the output results do not directly present the original content but rather use it as background material for analysis to discover patterns, trends, and other information.

1.1.4 Result Value

As an important tool for discovering potential value, TDM applications in commerce, education, scientific research, and social management contain enormous economic value and social opportunities, particularly in fields such as medicine, pharmaceuticals, finance, and other areas requiring market analysis. It can improve research efficiency, uncover hidden information, develop new knowledge, enhance research processes and foundations, and explore new fields.

1.2 Infringement Risks in TDM Utilization

The text or data mined by TDM includes any materials not protected by copyright or protected by copyright law. TDM implementation processes often require reproducing materials, though different TDM technologies vary significantly in their reproduction methods or quantities. Most TDM analysis prerequisites involve repeatedly reproducing entire works [3], though sometimes TDM technology only processes target text "one by one" or "individually," reproducing only single or small amounts of words or data each time without retaining or fixing the captured copies, merely counting word or data occurrences. In such cases, TDM behavior generally does not constitute copyright-relevant reproduction. This paper primarily discusses infringement risks arising from implementing TDM technologies that reproduce large amounts of target text. Specific risks include:

1.2.1 Potential Infringement of Reproduction Rights

China's revised Copyright Law of November 2020 defines reproduction rights as the right to produce one or more copies of a work by printing, photocopying, rubbing, recording, video recording, re-recording, re-shooting, digitization, or other means. Chinese legal academia generally holds that constituting "reproduction" under copyright law requires two conditions: the work must be reproduced on a tangible material carrier, and the work must be "fixed" on a tangible carrier [4]. TDM implementation processes generally include four steps: information extraction, semantic analysis, relationship calculation, and knowledge discovery. During the mining process, large numbers of copyrighted works are often involved. Whether reading data into the system or conducting format conversion and data analysis, reproduction acts controlled by copyright holders are involved. Due to the existence of numerous orphan works and systems' inability to effectively identify the rights status of target objects, if actors have not obtained authorization from rights holders or do not meet relevant infringement exemption conditions, TDM behavior likely infringes copyright holders' reproduction rights.

1.2.2 Potential Infringement of Database Rights

For databases with originality, reproduction rights provide protection. Whether copyright protection should be provided for databases without originality varies among countries. To better promote database industry development, the EU pioneered the Database Protection Directive (1996) [5], establishing a new special right—the database right—followed by Germany incorporating non-original databases into neighboring rights protection scope, while the UK enacted separate legislation for database protection. Due to widespread public concern about "information freedom" damage and rights monopolies, the US did not establish a new special rights system but instead adopted anti-unfair competition law for protection. In the EU, as long as producers have made substantial investment in database-related facilities and equipment, they can obtain special protection under database rights. According to Article 7 of the Database Protection Directive, database rights primarily grant producers the right to prohibit extraction

(permanently or temporarily transferring all or a substantial part of a database to another medium by any means) and re-utilization (making all or a substantial part of a database available to the public through reproduction, rental, network, or other transmission methods). During TDM, extraction acts are often unavoidable. If actors have not obtained rights holders' authorization and lack other exemption grounds, they may infringe copyright holders' database rights. It should be noted that since China has not stipulated database rights, TDM infringement risks for non-original databases are relatively low.

1.2.3 Potential Infringement of Other Copyright Holder Rights

From a broad rights perspective, TDM actors may also infringe copyright holders' adaptation rights and dissemination rights. China's Copyright Law does not directly adopt the concept of adaptation rights but breaks them down into translation rights, adaptation rights, compilation rights, etc. From a content perspective, TDM actors likely infringe copyright holders' adaptation rights because TDM technology applications require identifying and transcoding target text. Transcoding behavior involves "changing or arranging the expression form of target objects to form new research samples" [6], and the "transcoding behavior" in TDM processes is largely homogeneous with "adaptation behavior" under China's Copyright Law. Additionally, as the right most obviously affected by technological development, dissemination rights constitute a general term for copyright economic rights emerging after reproduction and adaptation rights, including performance rights, rental rights, broadcasting rights, exhibition rights, and information network dissemination rights. China's Copyright Law does not have a "dissemination rights" concept but instead responds to new technological developments by adding "information network dissemination rights." TDM's final analysis results may be presented either as simple conclusions or as lengthy commercial or academic reports. Regardless of presentation form, they may involve the original work's expression or database rights content. If TDM actors publicly disclose such content through online or offline means, they likely infringe copyright holders' dissemination rights.

1.3 Legal Dilemmas in TDM Utilization

As an emerging technology, although a few countries such as the EU, Japan, and the UK have timely revised laws to incorporate TDM into copyright fair use scope, most countries have not yet made institutional arrangements specifically addressing TDM technology. Existing copyright fair use provisions can only cover very small amounts of TDM behavior meeting specific conditions, leaving the vast majority of TDM behavior facing significant legal application dilemmas. For example, China's new Copyright Law issued in November 2020 lacks TDM-specific clauses, and TDM behavior cannot satisfy the fair use provisions in Article 24, Paragraph 1, Items 1 (personal use), 2 (appropriate quotation), 6 (teaching and scientific research), or 8 (cultural institutions). Even in countries that have issued corresponding rules, due to case-by-case judgment or applicable condition restrictions, whether TDM behavior constitutes fair use also faces legal

uncertainty. Current TDM fair use mechanisms mainly fall into two types: the flexible exception system represented by the US and the statutory exception system represented by the EU, both facing corresponding dilemmas.

1.3.1 Flexible Exception System Faces Case-by-Case Judgment Dilemmas

US copyright statutes do not contain explicit TDM provisions but instead apply the "four factors" of fair use under Section 107 to comprehensively judge whether TDM behavior constitutes fair use. The most typical cases are the Google Books case [7] and the Hathitrust Digital Library case [8]. Courts ultimately determined that TDM behavior implemented by Google and Hathitrust was highly transformative and constituted fair use. Although the US flexible exception system has significantly promoted TDM technology application and development, the system also has major defects. Relying on judicial case-by-case reasonableness judgments prevents providing the public with stable legal expectations of what constitutes fair use and is unfavorable to AI industry development. For instance, in the 2018 "TV Eyes case" [9], the Second Circuit Court partially overturned the district court's decision, concluding that TV Eyes' provision of television broadcast content video clips (not exceeding 10 minutes) for search and browsing services likely constituted a substitute for the original work and did not constitute fair use.

1.3.2 Statutory Exception System Faces Overly Strict Conditions Dilemmas

Traditional EU copyright law lacks clear application space for TDM. For example, Article 5(1) of the EU Information Society Copyright Directive (2001) [10] on temporary reproduction exceptions and Articles 5(2) and (3) on non-mandatory personal use exceptions and scientific research exceptions for non-commercial purposes cannot effectively satisfy most TDM behavior. The EU's Digital Single Market Copyright Directive (2019) proposed two TDM exceptions: one for scientific research purposes and one for text and data mining. While these exceptions can promote TDM industry development, the scientific research purpose exception is limited to research purposes with a narrow application scope, while the text and data mining exception attaches rights holders' reservation of rights declarations, leaving TDM behavior facing potential infringement risks [11].

2 Divergent Positions on TDM Fair Use

2.1 Publishers' Position on TDM Fair Use

As TDM technology continues developing, legal certainty issues have become increasingly prominent. Throughout copyright law history, rights holders' "bundles of rights" have continuously expanded to adapt to new communication technologies. Concerning potential negative impacts of emerging TDM technology on rights holders' interests, publishers generally hope to obtain new right types through legislative amendments to counter new technology applications—this

new right being "the right to mine." International publishers' specific positions on users' TDM fair use include:

2.1.1 Advocating Licensing Agreement Models as Optimal TDM Practice

Since most countries or regions have not enacted specific TDM legal provisions, most publishers advocate resolving TDM legal certainty issues through licensing agreements, as these can meticulously regulate TDM behavior. They argue that the "non-commercial research" concept is vague, TDM exception clauses would undermine incentives for publishers' continuous investment in high-quality content, and there is no solid evidence that lacking TDM exception systems causes economic or competitive disadvantage. Publishers have repeatedly expressed these positions, such as at the 2013 EU stakeholder dialogue on "European Licenses," where participating publishers recommended adopting multi-party cooperative market mechanisms for TDM issues, promising to provide convenience for non-commercial researchers to conduct TDM based on licensing agreement terms [12]. In 2014, the European Publishers Association explicitly opposed introducing TDM fair use clauses in copyright law, arguing that licensing agreements could meet TDM practice needs and reduce infringement and abuse risks. The International Association of Scientific, Technical & Medical Publishers (STM) stated in 2015 that TDM legal certainty could be achieved through licensing agreements and that TDM exception clauses would undermine investment incentive mechanisms ensuring high-quality content production [13]. In November 2016, STM responded to the Digital Single Market Copyright Directive (draft), arguing the proposed exception would produce unintended consequences: (1) introducing public-private partnership concepts could facilitate commercial entities' potential abuse of TDM exceptions; (2) the key to TDM exception application is lawful acquisition of relevant content; (3) for publishers' platforms, distinguishing genuine text miners from those attempting copyright infringement through massive illegal reproduction is difficult, as both access materials automatically, and current directives' wording on technological protection is vague, causing confusion and exposing publishers' content to risk; (4) copies of works or subject materials should be deleted immediately upon completion of extraction. STM also argued that US TDM must proceed according to licensing agreements and can only be considered fair use when no market substitute exists or no market harm is caused to rights holders, while UK and French copyright exceptions for scientific research are stricter than the directive draft in requiring non-commerciality [14].

2.1.2 Advocating TDM Implementation Through Publisher-Provided APIs and Platforms

Many publishers advocate providing TDM services through Application Programming Interfaces (APIs) or mining platforms. To address cross-platform licensing difficulties, thousands of publishers partnered with CrossRef in May 2014 to launch CrossRef TDM services [15], providing a common API and licensing agreement framework. Publishers' rationale for requiring users to utilize API platforms for TDM includes: (1) ensuring system platform operation—allowing

users to use arbitrary third-party software for batch data extraction or downloading would create enormous pressure on their platforms or cause system crashes, affecting normal user use and increasing publishers' breach of contract risks; (2) having relevant legal basis—UK copyright law exception clause explanatory documents state that publishers may take reasonable measures (such as limiting download speeds, controlling user access volume within specific time periods, etc.; Elsevier even believes reasonable measures include requiring TDM implementation through specialized APIs [16]) to maintain network security and stability, provided these measures do not prevent or unreasonably restrict any researcher's ability to benefit from exceptions [17]; (3) distinguishing between two types of behavior—publishers strictly differentiate the connotations of "mining" and "reading," considering them fundamentally different information activities requiring different licensing agreement terms to determine TDM rights scope.

Some publishers have established numerous restrictive conditions for API services. Taking Elsevier's TDM policy as an example: its TDM registration form shows Elsevier's API service imposes strict restrictions on subscribers, including limiting mining scope to XML files only, requiring negotiation with relevant rights holders for authorization to reuse images and other information, explicitly prohibiting any automatic downloading devices or software other than APIs to obtain website content, and strictly limiting access quantity and frequency [18]. Mining results are also restricted: although researchers are allowed to store results in institutional repositories or publish papers, creating, modifying, or translating any derivative works based on corpora is prohibited to avoid competitive threats to products and services. Springer Nature's TDM policy is also representative. Springer partners with the Copyright Clearance Center (CCC) to provide cross-platform TDM solutions [19]. Although Springer's TDM policy is relatively lenient, allowing subscribing institution researchers to use tools like PubMed, Web of Science, or Springer Nature's metadata API to implement TDM, and directly download full-text data content from its platform without mandatory API key registration (though download rates are limited to one request per second, or 150 requests per minute with registered API keys), researchers must take reasonable measures to ensure data security, such as storing data on internal secure servers, prohibiting third-party access, and using data only during TDM projects [20].

2.2 Libraries' Position on TDM Fair Use

TDM concerns whether the public can enjoy the right to freely access knowledge. Faced with rights holders' tightening "encirclement" measures triggered by TDM technology's copyright interest conflicts, libraries—as important groups representing public interests—have taken the initiative to voice their positions, proposing the important claim that "the right to read includes the right to mine," and have undertaken a series of efforts to actively fight for public TDM rights, particularly focusing on resolving legal uncertainty issues facing TDM and clearly articulating the international library community's position, which

has gained recognition from numerous consumer organizations.

2.2.1 Advocating Prompt Improvement of TDM Fair Use Rules

The International Federation of Library Associations and Institutions (IFLA) issued a Statement on Text and Data Mining in 2013 [21], stating that as an important tool for promoting learning and creating new forms, TDM legal certainty can only be achieved through (statutory) exceptions, that licensing agreements cannot serve as a solution for TDM, and that information should be utilized without restriction, which is crucial for enhancing community education and cultural welfare. IFLA believes that without TDM exceptions, researchers face risks from legal uncertainty when conducting important research and data-driven innovation. The Association of European Research Libraries (LIBER) released The Hague Declaration on Knowledge Discovery in the Digital Age [22], proposing that users enjoy rights to privacy, information, and mining, and that policymakers should clarify legal content, explicitly include "mining rights" within reading rights, and ensure content mining does not infringe copyright and neighboring rights. It recommends that universities, research funders, research institutions, and commercial entities formulate policies encouraging content mining research methods.

The library community's call for improving TDM fair use rules through legal systems is based on several reasons: (1) Regulating TDM through licensing agreements has major drawbacks. During licensing agreement negotiations, publishers (representing private interests) and libraries (representing public interests) have unequal status, with publishers often leveraging their monopoly advantages to impose unilateral will through reserved modification clauses or unconscionable terms, unreasonably restricting users' TDM behavior, and long-term agreement validity cannot be guaranteed. Moreover, the exceptionally rich information utilized in TDM far exceeds research databases covered by licensing agreements, making licensing solutions impractical. (2) Establishing TDM exception clauses helps enhance national international competitiveness. Some countries have already incorporated TDM fair use systems into their copyright legal frameworks; for example, the US and Canada regulate TDM behavior based on fair use frameworks, while the EU, UK, and Japan have enacted specialized TDM fair use provisions. These countries ensure efficient TDM implementation through legal revisions, helping their researchers better seize the initiative in scientific innovation and thereby promoting technological innovation and economic development.

2.2.2 Opposing Restrictions on TDM-Related Tool Software

Regarding most publishers' requirements that users employ specialized APIs or mining platforms to restrict TDM, libraries and other information service institutions hold opposing positions for the following reasons: (1) API platforms themselves have numerous drawbacks. LIBER and other organizations believe Elsevier's API registration terms are overly strict, such as only allowing text mining while excluding images, charts, videos, etc., explicitly prohibiting automated programs like robots and spiders, and limiting quotations from original

text fragments to no more than 200 characters. (2) Allowing only API platforms creates monopolies. Prohibiting researchers from using their own or third-party R&D tools is detrimental to maintaining academic freedom, promoting scientific research, and improving research efficiency, and infringes on researchers' privacy rights. (3) Technical restrictions on API platforms violate readers' legitimate reading rights. Information service institutions believe that the right to read includes the right to mine, and since institutions invest substantial funds to purchase corresponding database resources, they should provide readers with unrestricted mining rights [23].

3 Foreign Approaches to TDM Fair Use

3.1 US Judicial Approach

To address copyright dilemmas facing TDM behavior, the US has not resolved the issue through separate legislation but rather by applying principled fair use provisions to judicial practice, ultimately confirming the legality of TDM behavior through individual cases. Since 2003, US courts have recognized TDM-related reproduction as fair use in multiple judgments [24], such as *Kelly v. Arriba Soft* (2003) [25], *Field v. Google* (2006) [26], *Perfect 10 v. Amazon* (2007) [27], *A.V. v. iParadigms* (2009) [28], *Authors Guild v. Google* (2011) [29], *Fox News Network v. TVEyes* (2014) [30], *White v. West* (2014) [31], and *Authors Guild v. Hathitrust* (2014) [32], with the *Authors Guild v. Google* and *Authors Guild v. Hathitrust* cases being most representative. In these cases, US courts applied the "four factors" test for fair use under Section 107 of the Copyright Act, introducing "transformative use" theory to reason that TDM behavior constitutes fair use. The transformative use concept was first proposed by Judge Leval in 1990, referring to innovative use of works for different purposes or in different ways [33]. In the 1994 *Campbell v. Acuff-Rose Music* case, transformative use was first recognized as fair use in the judicial domain [34]. Subsequently, P. Samuelson further refined transformative use into three types based on the *Campbell* case: (1) using new expression methods to transform original works when commenting; (2) appropriately altering original works' meaning or conveyed information to add new content; and (3) using original works for purposes different from their original purpose [35].

In *Authors Guild v. Google*, Google developed the "Google Books Project," which involved full-text digitization scanning of paper books provided by US university libraries, using TDM core technology to develop search and snippet browsing functions to provide efficient and innovative digital retrieval services to the public. Specifically, Google divided each scanned book page into eight parts, presenting only small snippets containing search terms when the public input search queries—this is essentially TDM technology application. The Authors Guild filed a copyright infringement lawsuit against Google in 2005. In November 2009, both parties reached a settlement agreement allowing Google to use copyrighted works through implied licensing, which was later rejected by the court. In November 2011, the Southern District Court of New York issued its

first-instance judgment, finding that the "Google Books Project" had a highly transformative purpose—providing comprehensive and efficient text search services to the public—and constituted fair use. In December 2015, the US Second Circuit Court upheld the judgment, adding: "Commercial profit cannot be an absolute standard for denying fair use; and the snippet retrieval model does not create a substitution effect in the copyright market and will not diminish rights holders' substantive interests" [36]. In *Authors Guild v. Hathitrust*, the Hathitrust Digital Library allowed the public to search for specific words across all digital copies in its institutional repository, but search results only displayed page numbers where keywords appeared and the frequency of those words on each page. The district court applied the fair use "four factors" test, finding that Hathitrust's retrieval service was not simple work usage but derived new academic research methods and pathways with strong "transformative purpose" [37]. The US Second Circuit Court also held that Hathitrust Digital Library's TDM behavior constituted fair use [38]. Thus, US courts have maximally confirmed the legality of TDM behavior from a judicial perspective, representing an "unconditional exception" model. However, this model suits common law countries, requires judges with sufficient judicial experience, involves case-by-case judgment, and suffers from low efficiency and high costs.

3.2 UK Legislation

In 2010, the UK government launched comprehensive copyright system reform. In May 2011, Professor I. Hargreaves, commissioned by the UK government, published the report *Digital Opportunity: A Review of Intellectual Property and Growth* [39], which made 10 recommendations regarding UK intellectual property legislation, including introducing a TDM behavior exception for non-commercial research. In 2014, the UK enacted the Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations [40], introducing a TDM copyright exception in Section 29A (Copying for Text and Data Mining for Non-Commercial Research) of the Copyright, Designs and Patents Act (1988) [41]. The main contents include: (1) A person who has lawful access to a work does not infringe copyright by reproducing it, provided that: reproduction is for computational analysis of any work content and the sole purpose is non-commercial research; and the reproduction includes sufficient acknowledgment of authorship (unless impractical). (2) Transferring reproductions to others or using them for purposes other than non-commercial research without copyright owners' permission constitutes infringement. (3) If reproductions are sold, rented, or offered for sale or rental, such reproduction constitutes infringement. (4) The text and data mining exception prevails over contractual terms.

UK copyright law thus stipulates three preconditions for TDM exception application: (1) Reproduced works must be lawfully acquired, meaning researchers can only conduct data mining and information analysis on copyrighted works if they or their institutions have already purchased relevant resources; otherwise, it constitutes infringement. (2) Reproduction purposes must be non-commercial

research, meaning any commercial TDM-related reproduction must obtain copyright holders' permission. (3) Reproduction must fully indicate work authorship information, meaning users must respect copyright holders' moral rights, and failure to fully indicate authorship constitutes infringement unless impractical. UK copyright law also stipulates corresponding restrictions for TDM exceptions, namely prohibiting transfer of reproductions to others, using them for purposes other than non-commercial research, or using them for "rental or sale." Notably, UK copyright law provides that statutory TDM exceptions prevail over contractual effectiveness, meaning any contract terms limiting or preventing TDM-related reproduction are unenforceable.

3.3 EU Legislation

The EU's 1996 Database Protection Directive and 2001 Information Society Copyright Directive, due to their closed-list approach to copyright exceptions, could not apply to TDM, affecting TDM technology development and application. To remedy this defect and further promote EU copyright law integration, the EU released the Digital Single Market Copyright Directive proposal in 2016, explicitly stipulating a TDM copyright exception for scientific research purposes. In April 2019, the EU formally promulgated the revised Digital Single Market Copyright Directive, adding a "TDM exception for text and data mining purposes" beyond the "TDM copyright exception for scientific research purposes." Article 3(1) stipulates that member states shall provide that research institutions and cultural heritage institutions may reproduce and extract works or other content they have lawful access to for scientific research purposes. Article 4 provides that member states shall provide that for text and data mining purposes, reproducing and extracting lawfully accessed works or other content does not infringe copyright.

For the "TDM copyright exception for scientific research purposes," the EU stipulates the following restrictions: (1) TDM implementation subjects must be research institutions and cultural heritage institutions. Research institutions refer to universities (including their libraries), research institutes, or other entities whose primary purpose is scientific research or teaching activities related to scientific research. Cultural heritage institutions refer to libraries, museums, archives, or audio-visual heritage institutions accessible to the public. (2) TDM implementation objects must be lawfully accessed works or other content. Lawfully sourced resources include print and electronic forms, meaning users' lawful access to works' print and electronic versions, scanning and digitizing paper materials, and downloading electronic materials. (3) TDM implementation purposes must be for scientific research. The EU does not explicitly restrict the nature of scientific research, and research activities conducted by research institutions in cooperation with commercial companies can still apply this exception. However, although this exception covers commercial research purposes, it has certain conditions: research institutions must reinvest profits obtained into their scientific research, and enterprises with decisive influence over the institution

cannot 优先 access relevant research results. (4) TDM implementation methods are limited to "reproduction" and "extraction." Whether "adaptation," "translation," or "compilation" apply to this exception is not explicitly stated. (5) TDM results must be stored with appropriate security levels to prevent users from abusing the TDM exception and protect rights holders' legitimate interests.

Given that the "TDM copyright exception for scientific research purposes" faces many restrictions and cannot be widely applied to TDM applications beyond scientific research (such as complex commercial decision-making, government services, or new application or technology development), and considering that TDM analytical behavior does not constitute "temporary reproduction" explicitly stipulated as fair use in the EU's 2001 Copyright Directive, the EU added the "TDM exception for text and data mining purposes." This exception applies as long as rights holders have not explicitly reserved relevant content usage in appropriate ways, allowing users to reproduce and extract lawfully accessed content without restrictions on subject qualification (not limited to research institutions) or use purpose (not limited to non-profit purposes), to better encourage private enterprise innovation.

3.4 Japanese Legislation

In 2009, Japan amended its Copyright Law, introducing a copyright exception for "reproduction for information analysis" in Article 47-7 [42], generally regarded as a TDM copyright exception. This provision states that for the purpose of information analysis by computer (meaning extracting language, sound, images, or other elements constituting information from numerous works and other large amounts of information, and conducting comparison, classification, and other statistical analyses), works may be recorded on recording media or adapted within necessary limits; however, this does not apply to database works specifically made for information analysis. Japan's "information analysis exception" provides good legal basis for TDM behavior and has progressive significance. This exception's characteristics include: (1) Applicable objects are not limited to text but also include language, images, sound, or other elements, except for database works specifically for information analysis; (2) Usage methods are not limited to specific data analysis methods but also include comparison, classification, or other statistical analysis methods; (3) Usage methods are not limited to reproduction but also include adaptation, meaning reproduction of derivative works created based on the work can also apply this exception. However, it must be noted that Japan's "information analysis exception" also has limitations, such as limiting information analysis tools to computers, being overly confined to existing technology, and ignoring possibilities for other advanced equipment and facilities beyond computers with future technological development.

3.5 German Legislation

In September 2017, Germany promulgated the Act on the Harmonisation of Copyright Law with the Requirements of the Knowledge Society [43], amending copyright law. The act restructured the system limiting copyright holders' rights for educational and scientific purposes, including introducing TDM exceptions and regulating relationships between legal licensing and contractual authorization, attempting to address criticisms that German copyright law could not adapt to technological changes and its complex structure was difficult for laypeople to understand [44]. Article 60d of the German Copyright Law is the TDM provision. Paragraph 1 provides that for automatic analysis of large amounts of source material for scientific research, users should be permitted to reproduce source materials in systematic and automatic ways for non-commercial purposes, and create analyzable corpora through normalization, structuring, and classification; these corpora may be disclosed to specific limited groups conducting cooperative scientific research and independent third parties responsible for monitoring scientific research quality. Paragraph 2 provides that using database works for TDM under Paragraph 1 constitutes normal utilization of the database. Paragraph 3 provides that after research completion, corpora and reproductions of original materials should be permitted to be transferred to libraries, archives, museums, and educational institutions for long-term preservation. Thus, German conditions for TDM behavior constituting fair use are: (1) Use subjects are not limited to specific subjects (such as research institutions or cultural heritage institutions) but include commercial entities (engaging in non-commercial research); (2) Use purposes are limited to non-commercial scientific research; (3) Use objects are not limited to lawfully sourced works but include database works; (4) Use behaviors include creating corpora for analysis and two exemption scenarios: disclosing such corpora to specific scientific researchers or third parties responsible for monitoring research quality; (5) TDM-related materials can be transferred to public institutions like libraries for long-term preservation.

4 Designing China's TDM Fair Use Rules

China's new Copyright Law promulgated in November 2020 added Article 24, Item 13, a catch-all fair use clause for "other circumstances stipulated by laws and administrative regulations," providing some legal space for recognizing TDM behavior as fair use. However, considering this provision's vagueness, the broad demand for TDM, and China's civil law tradition, this paper recommends establishing a specialized TDM copyright exception clause in the Copyright Law. Drawing on recent foreign legislative experience and combining it with China's actual conditions, recommended TDM fair use rules should mainly include the following content:

4.1 Subject Conditions for TDM Exceptions

Although the EU's 2019 Digital Single Market Copyright Directive does not limit subject qualifications for the TDM exception for text and data mining purposes, it limits subjects for the scientific research purpose TDM exception to research institutions and cultural heritage institutions. The US, UK, Japan, Germany, France, and other countries do not limit subject conditions for TDM fair use behavior, and commercial companies can also become subjects of TDM fair use. This paper recommends that applicable subjects for TDM fair use should not be limited to research institutions or public cultural institutions for the following reasons: (1) Subject limitations would prevent full utilization of TDM technology. As technology continues developing, internet or software companies that are better at responding to market changes often have stronger TDM research capabilities than research institutions or cultural heritage institutions. Many public cultural institutions cannot complete TDM alone and need to cooperate with relevant technology companies [45]. Therefore, limiting subjects to research institutions or public cultural institutions would be unfavorable for full application and development of TDM technology. (2) Subject limitations would harm citizens' personal research freedom. Completely excluding enterprises or individuals from becoming subjects of TDM fair use would harm the freedom of scientific research, literary and artistic creation, and other cultural activities protected by China's Constitution, destroying equal opportunities people should have to understand things' patterns. (3) Subject limitations would harm public interests. Limiting subjects to research institutions or public cultural institutions would block other subjects' motivation to contribute to public welfare by publishing TDM results [46], hindering commercial entities' active participation in public welfare activities.

4.2 Object Conditions for TDM Exceptions

Do TDM exception objects mandatorily require lawful sources? Countries have different provisions. Some countries such as the US, Germany, and Japan do not treat the legality of mined objects' sources as a prerequisite, while others like the EU, UK, and France limit TDM exception object conditions to lawfully accessed works. Lawful access generally refers to legally obtaining works through subscribing to books and journals, purchasing databases, or complying with open licensing agreements. The EU Digital Single Market Copyright Directive's Preamble Article 14 states that lawful access should be understood to include content obtained according to open access policies, or through contractual arrangements between institutions and rights holders, or through other lawful means, or accessing online freely available content (where rights holders have not appropriately reserved reproduction rights). UK government documents explicitly state that lawful access means legally enjoying rights to access works, with subscribing to books, journals, or databases and complying with contractual agreements being lawful access pathways [47]. Considering that determining whether mined objects have lawful sources requires substantial time and cost,

and from the perspective of promoting TDM new technology application and development and protecting public interests, this paper recommends that TDM exception objects should not be limited to lawful sources. On this issue, the Max Planck Institute for Innovation and Competition also recommends that research institutions should have the right to implement TDM without lawful access rights but should pay reasonable licensing fees [48]. Additionally, should TDM fair use rules apply to unpublished works? Some scholars explicitly oppose this [49-50]. However, this paper argues that although unpublished works fall within authors' absolute control scope, from the perspective of promoting knowledge dissemination and ensuring public access to works, the law should recognize fair use of unpublished works [51]. TDM behavior itself does not cause substantial substitution for original works and does not harm rights holders' interests. For libraries or archives that collect many precious manuscripts, diaries, correspondence, telegrams, and other resources with important historical and research value, conducting TDM on these unpublished works constitutes productive rather than consumptive use with strong public welfare characteristics. In summary, TDM exception object requirements should not have any restrictions, not being limited to lawful sources or published works, though users could be required to pay reasonable licensing fees for using works from non-legitimate sources.

4.3 Purpose Conditions for TDM Exceptions

The UK limits TDM purposes constituting fair use to non-profit or non-commercial purposes. The EU stipulates more relaxed requirements for TDM purpose conditions. For the "scientific research purpose TDM copyright exception," the EU uses scientific research purpose as a limiting condition but includes commercial research, though requiring research institutions to reinvest profits into scientific research. For the "TDM exception for text and data mining purposes," the EU imposes no corresponding purpose restrictions to better promote TDM technology application and development in fields beyond scientific research. The US determines TDM fair use through the four-factor test, with the US Second Circuit Court explicitly stating in the Google Books case that commercial profit cannot be an absolute standard for denying fair use. Japanese legislation also does not explicitly limit TDM purpose conditions. Synthesizing major countries' legislative situations, this paper recommends that China's TDM fair use purposes should be limited to "for scientific research and other reasonable purposes" for the following reasons: (1) Compared with non-commercial purposes, scientific research covers a broader scope, including both non-commercial and commercial research. Scientific research has essential public interest attributes, and TDM-generated new works do not substantially substitute original works. TDM behavior itself does not harm rights holders' interests, and commercial scientific research has its value. (2) Adding "other reasonable purposes" provides safeguards for TDM applications in other fields to better realize public interests. As a new data analysis technology, TDM has important application value in fields beyond scientific research, such as

major commercial decision-making, government services (e.g., public health risk assessment and decision-making, food safety supervision and control [52]), and new technology or procedure development.

4.4 Behavior Conditions for TDM Exceptions

TDM utilization behaviors generally include "reproduction," "extraction" (or "adaptation"), and "dissemination." Countries have different provisions regarding applicable behaviors for TDM exceptions. The UK limits TDM exception behaviors to reproduction; the EU limits them to reproduction and extraction; Japan limits them to reproduction and adaptation; Germany provides for reproduction and dissemination, but dissemination is limited to two specific objects (specific scientific researchers or third parties monitoring research quality). The US is unique: although it has no specialized legislation on TDM exception behaviors, relevant judicial cases show US courts generally limit TDM exception behaviors to reproduction and limited dissemination (not constituting substantial substitution of original works). For example, in the Google Books case, the court held that Google's snippet presentation of original works (a form of dissemination) was insufficient to substitute for original works and constituted fair use; but in the TVEyes case, the Second Circuit Court in 2018 overturned the 2014 district court's conclusion that defendant's snippet presentation constituted fair use, finding that providing users with snippet browsing services not exceeding ten minutes allowed users to understand original works' core ideas through these snippets, creating market threats to original works and not constituting fair use. In summary, considering that reproduction and extraction or adaptation are the most fundamental and essential aspects of text and data mining applications, and that completely prohibiting "dissemination" would not serve public interests, this paper recommends that TDM exception behaviors include "reproduction," "extraction or adaptation," and "limited dissemination" (dissemination to specific objects such as scientific researchers or dissemination to the public where content is insufficient to substitute for original works).

4.5 Other Conditions for TDM Exceptions

Other conditions for TDM exceptions include: (1) Security preservation measures. TDM results need secure storage for subsequent data analysis or result verification. Many countries' legislative examples strictly regulate subsequent storage behavior for TDM, which China should draw upon. For example, the EU requires that work copies made for TDM should be stored with appropriate security levels for scientific research or verification of research results. Germany requires that corpora and reproductions of original materials should be deleted after research completion and no longer made available to the public, but allows their transfer to libraries, archives, museums, and educational institutions for long-term storage. In the US Google Books case, Google emphasized its implementation of strict security preservation measures to prevent digital book leakage and dissemination. (2) Technological protection measure circumvention

exceptions. In the digital era, increasing numbers of people employ technological measures to protect their works. If users cannot lawfully circumvent technological measures, TDM behavior cannot be implemented. Globally, many countries have stipulated multiple statutory exceptions for technological protection measure circumvention, especially designating libraries, archives, and other institutions as subjects of lawful circumvention, such as in the US, France, Germany, and Australia, which to some extent benefits public institutions implementing TDM. However, these statutory exceptions cannot fully satisfy TDM application and development needs, so some regions or countries have specifically legislated that technological protection measures should not hinder TDM implementation. For example, Article 3(3) of the EU Digital Single Market Copyright Directive provides: "Rights holders may take measures to ensure the security and integrity of networks and databases bearing works or other copyright-protected content, but such measures should not exceed what is necessary to achieve this objective." The Max Planck Institute for Innovation and Competition also recommends that necessary technological protection measures taken by rights holders should not unnecessarily hinder text and data mining [47]. Regarding China, the 2020 revised Copyright Law only provides five statutory circumstances for technological protection measure circumvention exceptions—for classroom teaching or scientific research, for blind persons' benefit, for official duties, for security testing, and for encryption or reverse engineering research—without even designating libraries and other institutions as lawful circumvention subjects, which far from satisfies TDM needs. Therefore, China especially needs to stipulate technological protection measure circumvention exception clauses specifically for TDM. (3) Prevalence over contractual terms. To prevent database providers and other rights holders from using their monopoly positions to exclude TDM applications through contractual monopoly clauses, China needs to legislatively provide that any contract terms excluding TDM are invalid, as seen in similar provisions in the UK's Copyright, Designs and Patents Act Section 29A(5) and Germany's Copyright and Related Rights Act Section 60g(1). (4) Sufficient copyright information labeling. The UK requires that reproductions for TDM must include sufficient acknowledgment of work authorship unless labeling is not feasible, such as when annotating dispersed text and data. Meanwhile, TDM purposes, occurrence and ending times and locations, and TDM objects should be fully explained in written or electronic form, except involving national security, emergency states, or other special circumstances [51]. (5) Use of mining results. It is recommended that TDM results can be publicly disclosed for free but subject to certain restrictions, such as non-commercial purposes, prohibition of private transfer without rights holders' permission, and prohibition of related transaction behaviors (sale or rental, or promising sale or rental, or disclosure for sale or rental).

In summary, this paper recommends that China's TDM fair use rules mainly include: (1) Any institution or individual, including research institutions, public cultural institutions, commercial entities, and other organizations, conducting text and data mining on works through reproduction, extraction, adaptation, or

limited dissemination for scientific research or other reasonable purposes does not require copyright holders' permission and does not require payment of remuneration, but should indicate the author's name or designation and work title, and must not affect the work's normal use or unreasonably harm copyright holders' legitimate interests. (2) The above text and data mining behavior may circumvent technological protection measures but must not provide others with technologies, devices, or components for circumventing technological protection measures, and must not infringe rights holders' other legally enjoyed rights; use of works from non-legitimate sources is not prohibited, but reasonable licensing fees must be paid; text and data mining results should be securely stored and may be publicly disclosed for non-commercial purposes without harming rights holders' interests; any contract agreement aimed at preventing or restricting utilization behavior permitted under this clause is invalid.

References

[1] UK Intellectual Property Office. Guidance of exceptions to copyright [EB/OL]. [2021-04-13]. https://www.gov.uk/guidance/exceptions-to-copyright#text-and-data-mining-for-non-commercial-research.

[2] The European Union. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the digital single market and amending directives 96/9/EC and 2001/29/EC (Text with EEA relevance.) [EB/OL]. [2021-04-13]. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2019.130.01.0092.01.

[3] BORGHIM, KARAPAPA S. Copyright and mass digitization: a cross-jurisdictional perspective [M]. New York: Oxford University Press, 2013.

[4] Wang Qian. Copyright Law [M]. Beijing: China Renmin University Press, 2015.

[5] The European Union. Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [EB/OL]. [2021-04-13]. https://wipolex.wipo.int/zh/text/126788.

[6] Dong Fan, Guan Yonghong. On the construction of copyright exception rules for text and digital mining technology application [J]. Hebei Law Science, 2019, 37(9): 148-160.

[7] Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015) [EB/OL]. [2021-04-13]. https://www.lexisnexis.com/.

[8] Authors Guild, Inc. v. Hathitrust, 755 F, 3d. 87 (2d Cir. 2014) [EB/OL]. [2021-04-13]. https://www.lexisnexis.com/.

[9] Fox News Network, LLC v. TVEyes, Inc., 883 F.3d 169 (2d Cir. 2018) [EB/OL]. [2021-04-13]. https://www.lexisnexis.com/.

[10] The European Union. Directive 2001/29/EC of the European Parliament

and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society [EB/OL]. [2021-04-13]. https://www.wipo.int/edocs/lexdocs/laws/en/eu/eu049en.pdf.

[11] ROSATI E. Copyright as an obstacle or an enabler? a European perspective on text and data mining and its role in the development of AI creativity [J]. Asia Pacific law review, 2019, 27(2): 198-219.

[12] European Commission. Licences for Europe stakeholder dialogue [EB/OL]. [2021-04-13]. https://ec.europa.eu/commission/presscorner/detail/en/MEMO_{13} frequentlyaskedquestions.

[13] Ru Lijie, Gu Liping, Tian Pengwei. International publishers' restrictions on text and data mining [J]. Library Construction, 2016(7): 27-33.

[14] Ten things to know about text mining and the proposed copyright directive COM(2016) 593 final [EB/OL]. [2021-04-13]. https://www.stm-assoc.org/2016_{{{11}}*{{24}}*{{{2016}}*{{11}}*{{{STM}}*{{Ten}}*{{{things}}*{{to}}*{{{know}}*{{about}*

[15] Text and data mining [EB/OL]. [2021-04-13]. https://www.crossref.org/education/retrieve-metadata/rest-api/text-and-data-mining/.

[16] How does Elsevier's text mining policy work with new UK TDM law? [EB/OL]. [2021-04-13]. https://www.elsevier.com/connect/how-does-elseviers-text-mining-policy-work-with-new-uk-tdm-law.

[17] Intellectual property office. Exceptions to copyright: research (2014) [EB/OL]. [2021-04-13]. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachm

[18] Elsevier provisions for text and data mining (TDM) [EB/OL]. [2021-04-13]. https://www.elsevier.com/___{data}/assets/pdf_{file}/0012/102234/TDM-sign-up-short-form.pdf.

[19] Springer partners with copyright clearance center to power text and data mining solution [EB/OL]. [2021-04-13]. https://www.springer.com/springer-partners-with-copyright-clearance-center-to-power-text-and-data-mining-solution/.

[20] Springer Nature TDM policy [EB/OL]. [2021-04-13]. https://www.springernature.com/gp/researchers/text and-data-mining.

[21] IFLA statement on text and data mining [EB/OL]. [2021-04-13]. https://www.ifla.org/publications/node/8225.

[22] LIBER. The Hague declaration on knowledge discovery in the digital age [EB/OL]. [2021-04-13]. https://thehaguedeclaration.com/wp-content/uploads/sites/2/2015/04/Liber_{{{Declaration}}*{{A4}}*{2015}.pdf.

[23] LIBER calls on Elsevier to withdraw TDM policy [EB/OL]. [2021-04-13]. https://libereurope.eu/article/liber-calls-on-elsevier-to-withdraw-tdm-policy/.

[24] Luo Jiao, Zhang Xiaolin. Copyright law and policy recommendations supporting text and data mining [J]. Journal of Library Science in China, 2018,

44(3): 21-34.

[25] Kelly vs. Arriba Soft, 336 F.3d 811 (9th Cir. 2003) [EB/OL]. [2021-04-13]. https://www.lexisnexis.com/.

[26] Field vs. Google, 412 F. Supp. 2d 1106 (2006) [EB/OL]. [2021-04-13]. https://www.lexisnexis.com/.

[27] Perfect 10 vs. Amazon, 508 F.3d 1146 (9th Cir. 2007) [EB/OL]. [2021-04-13]. https://www.lexisnexis.com/.

[28] A.V. v. iParadigms, LLC, 562 F.3d 630, 634 (4th Cir. 2009) [EB/OL]. [2021-04-13]. https://www.lexisnexis.com/.

[29] Authors Guild v. Google, 770 F. Supp. 2d 666 (2011) [EB/OL]. [2021-04-13]. https://www.lexisnexis.com/.

[30] Fox News Network, LLC v. TVEyes, Inc., 43 F. Supp. 3d 379 (2014) [EB/OL]. [2021-04-13]. https://www.lexisnexis.com/.

[31] White v. West Pub'g Corp., 29 F. Supp. 3d 396 (2014) [EB/OL]. [2021-04-13]. https://www.lexisnexis.com/.

[32] Authors Guild, Inc. v. Hathitrust, 755 F.3d 87 (2d Cir. 2014) [EB/OL]. [2021-04-13]. https://www.lexisnexis.com/.

[33] LEVAL P. Toward a fair use standard [J]. Harvard law review, 1990, 103(5): 1105-1136.

[34] Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569 (1994) [EB/OL]. [2021-04-13]. https://www.lexisnexis.com/.

[35] SAMUELSON P. Possible futures of fair use [EB/OL]. [2021-04-13]. https://ssrn.com/abstract=2584180.

[36] Authors Guild, Inc. v. Google, Inc. 804 F.3d 202 (2d Cir. 2015) [EB/OL]. [2021-04-13]. https://www.lexisnexis.com/.

[37] Authors Guild, Inc. v. Hathitrust, 902 F. Supp. 2d 445 (2012) [EB/OL]. [2021-04-13]. https://www.lexisnexis.com/.

[38] Authors Guild v. Hathitrust, 755 F.3d. 87 (2d Cir. 2014) [EB/OL]. [2021-04-13]. https://www.lexisnexis.com/.

[39] HARGREAVES I. Digital opportunity: a review of intellectual property and growth [EB/OL]. [2021-04-13]. https://www.gov.uk/government/uploads/system/uploads/attachment_{data} finalreport.pdf.

[40] The copyright and rights in performances (research, education, libraries and archives) regulations [EB/OL]. [2021-04-13]. https://www.legislation.gov.uk/uksi/2014/1372/contents/made.

[41] The copyright, designs and patents act 1988 [EB/OL]. [2021-04-13]. https://wipolex.wipo.int/en/legislation/details/18023.

[42] Japanese intellectual property law [M]. Translated by Yang Heyi. Beijing: Peking University Press, 2014.

[43] Act on the harmonisation of the copyright law with the requirements of the knowledge society (Copyright Knowledge Society Act) [EB/OL]. [2021-04-13]. https://wipolex.wipo.int/en/legislation/details/18029.

[44] German reform on the use of copyright protected works in the fields of education and research will come into force soon [EB/OL]. [2021-04-13]. http://copyrightblog.kluweriplaw.com/2018/01/15/german-reform-use-copyright-protected-works-fields-education-research-will-come-force-soon/.

[45] Wang Wenmin, Gao Jun. Copyright exception rules for library information analysis in the age of artificial intelligence [J]. Library Tribune, 2020, 40(9): 60-68.

[46] Song Yaxin. Copyright exception for text and data mining—taking the EU Copyright Directive amendment proposal as perspective [J]. Intellectual Property, 2017(10): 109-116.

[47] UK intellectual property office. Exceptions to copyright: research [EB/OL]. [2021-04-13]. https://www.gov.uk/guidance/exceptions-to-copyright#text-and-data-mining-for-non-commercial-research.

[48] Max Planck institute for innovation & competition. Position statement of the Max Planck institute for innovation and competition on the proposed modernisation of European copyright rules part b exceptions and limitations (Art. 3-Text and data mining) [EB/OL]. [2021-04-13]. https://pure.mpg.de/rest/items/item_{2383669}8/component/file_{2409840}/content.

[49] Tang Sihui. Research on copyright exceptions for text and data mining in big data environment—taking the EU DSM Copyright Directive proposal as perspective [J]. Intellectual Property, 2017(10): 109-116.

[50] Dong Fan, Guan Yonghong. On the construction of copyright exception rules for text and digital mining technology application [J]. Hebei Law Science, 2019, 37(09): 148-160.

[51] Liang Zhiwen. Fair use of unpublished works under China's copyright law and its legislative model [J]. Law Science, 2008(3): 101-108.

[52] Zhao Li. Drawing on The Hague Declaration on Knowledge Discovery in the Digital Age—taking content mining as core [J]. Library, 2015(9): 22-26.

Author Contributions:
Wu Gao: Responsible for initial draft writing and revision.
Huang Xiaobin: Responsible for paper revision and content supplementation.

Note: Figure translations are in progress. See original paper for figures.