ASSOCIATION OF RESEARCH LIBRARIES®

ISSUE BRIEF
Text and Data Mining and Fair Use in the United States[1]

Background

No researcher can read all relevant research articles that are published in her field of interest. Even if she could, she would not be able to detect patterns in the research results that emerge only from large-scale computational analysis, known as text and data mining (TDM). Researchers who want to perform TDM on copyrighted research articles might seek clarity about whether they need permission from journal publishers or whether copyright's fair use doctrine permits TDM on accessible articles. In almost all cases, performing TDM on accessible articles is a fair use. As long as the researcher is not bound by a contract that forfeits her fair use rights, she may proceed with TDM so long as her results do not make the full text, or substantial portions, of the underlying articles publicly available.

Computer programs can crawl millions of articles or other content in digital form (those works that are not "born digital" must be digitized in order to make use of TDM) to derive or organize information from text or data. TDM makes the analysis of vast amounts of information possible that would not otherwise be possible. Through TDM, researchers can discover new knowledge from existing knowledge. TDM can help researchers sort through large amounts of information, identify patterns and trends, and understand individual texts as well as the connections between texts.

TDM almost always involves copying, but not all copying amounts to copyright infringement. If TDM merely involves making temporary copies of text and other data for the purposes of analyzing these, and the durable outputs of the computational analysis are merely facts about what is in the literature, then copyright does not apply at all. The reason is that a copyright owner only has the exclusive right to reproduce the article in "copies" and the law defines a "copy" for copyright purposes as one that lasts for a "period of more than a transitory duration." In a case involving copies made by a computer buffer, a federal court of appeals determined that temporary copies that lasted for only 1.2 seconds were transitory and therefore outside the scope of copyright law.[2]

However, many researchers want to keep a durable copy of the articles that underlie their TDM analysis as a reference to validate their results. In other cases, the durable outputs of TDM may have some copyrighted content, such as images. Copyright does

---

[1] Prepared by Krista L. Cox, director of public policy initiatives, on June 5, 2015.

[2] Cartoon Network LP, LLLP v. CSC Holdings, Inc., 536 F.3d 121 (2d Cir. 2008).

apply to these cases. The issue, then, is whether these uses require a license or qualify as fair uses that do not.

Numerous courts in the United States have upheld the reproduction necessary to perform TDM as fair use, even though the content being copied into the database is copyrighted. Fair use is a flexible limitation and exception that allows copyright law to adapt to changing circumstances and new technologies and helps ensure a balanced copyright system. Thus, while the United States does not have a specific limitation or exception to explicitly allow TDM, fair use has accommodated the creation and growth of TDM as a new research tool.

Fair Use and TDM

Fair use is an equitable doctrine that has long been a critical component to the U.S. copyright system. The doctrine was a common law principle later codified in the 1976 Copyright Law. The House Report discussing the provision noted that it "endorses the purpose and general scope of the doctrine of fair use, but there is no disposition to freeze the doctrine in statute, especially during a period of rapid technological change."[3] Congress therefore intended for courts to continue to apply fair use as an equitable rule, responsive to technological change. Courts have repeatedly emphasized the importance of applying fair use in a flexible manner.[4] When fair use is relied upon, the user does not need to seek permissions.

The fair use statute, codified at 17 U.S.C. 107, reads:

Notwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include:

(1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
(2) the nature of the copyrighted work;
(3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
(4) the effect of the use upon the potential market for or value of the copyrighted work.

---

[3] H.R. Rep. No. 94-1476.

[4] *See, e.g.,* Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1166 (9th Cir. 2007) (". . . [W]e note the importance of analyzing fair use flexibly in light of new circumstances.")

> The fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors.

The four factors are weighed as a whole and a defendant need not win on every factor for a court to rule in favor of fair use. Fair use is a flexible doctrine and whether a particular use is considered fair is determined on a case-by-case basis. As a flexible doctrine, it can accommodate changing and evolving technologies to allow for new uses that may not (or could not) have been conceived of when the Copyright Law was last amended. Where fair use applies, it is unnecessary to seek the permission of the rights holder.

TDM may be used for a variety of purposes, some of which are explicitly referenced in Section 107, such as for scholarship and research. Beginning with a 2003 case involving the incorporation of images in a search engine, in at least eight different cases, courts have found that the creation of a database for TDM and its use amounts to fair use. The purposes have ranged from research by scholars, to use by politicians, to checking for plagiarism. Many of these courts have focused heavily on the benefit that TDM provides to the public, because they "enhance[e] information-gather techniques,"[5] and noted that creation of search engines or databases are highly transformative uses that favor fair use. Importantly, courts have acknowledged that TDM does not serve as a substitute for the original work, but instead these databases offer a new purpose.

Courts have explicitly upheld TDM as fair use in the following instances:

- *Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014) – HathiTrust digitized works for inclusion in a database that enabled data mining and textual analysis and made it easier to identify and locate sources of information.
- *White v. West* (S.D.N.Y. 2014) – Two publishers copied legal filings, including motions and briefs into databases, Westlaw and LexisNexis. Westlaw and LexisNexis added metadata to the copied legal filings that were collected into its databases, creating an interactive legal research tool. The search results included the full text of the legal filings.
- *Fox v. TVEyes* (S.D.N.Y. 2014) – TVEyes recorded the entire contents of television and radio broadcasts. Then, using closed captions and speech-to-text technology, TVEyes created a searchable database of that content. The search results included portions of the transcripts of the programs.
- *Authors Guild v. Google*, 770 F.Supp.2d 666 (S.D.N.Y. 2011) – Google digitally scanned books in the collections of partner libraries and incorporated the works into a searchable database that could be used by scholars and researchers. The search results included "snippets" of text an eighth of a page long. The Southern District of New York explicitly referenced the benefit of Google Books to TDM, noting the project "transformed the book text into data for the purpose of substantive research, including data mining and text mining in new areas,

---

[5] *Kelly v. Arriba-Soft*, 336 F.3d 811 (9th Cir. 2003)

thereby opening up new fields of research. Words in books are being used in a way that they have not been used before."

- *A.V. v. iParadigms, LLC* (4th Cir. 2009) – iParadigms created a database called TurnItIn which allowed teachers to compare a student's work submitted through the site with content available on the Internet, as well as papers previously submitted to the service, in order to determine whether the work had been plagiarized. Despite the commercial nature of the TurnItIn service, the use was considered "highly transformative."
- *Perfect 10 v. Amazon*, 508 F.3d 1146 (9th Cir. 2007) – Google used "thumbnail" versions of copyrighted images in its search engine and included "in-line linking" to the full images which directed the user to the full-size image on the plaintiff's website. The Ninth Circuit found that the purpose as an "electronic reference tool" was highly transformative.
- *Field v. Google*, 412 F.Supp.2d 1106 (D. Nv. 2006) – Google provided copies of an author's original web content in its website cache. The cached links were used for a number of reasons, including archival copies, for web comparisons or identification in a search query.
- *Kelly v. Arriba Soft*, 336 F.3d 811 (9th Cir. 2003) – The search engine company, Arriba Soft, included thumbnails of and in-line linking to images hosted on the photographer's website. Arriba Soft's search engine was used as a tool to help index and improve access to images on the Internet.

Analysis of the Four Fair Use Factors

In assessing the four fair use factors, courts have emphasized the transformative nature of searchable databases, noted the unlikelihood of adverse impact on the original market for the work and upheld fair use.

Factor 1: Purpose and Character of the Use

Courts have found the creation of searchable databases or search engines to be highly transformative and thus the first factor has repeatedly favored fair use. In *Authors Guild v. HathiTrust*, for example, the Second Circuit explained "creation of a full-text searchable database is a quintessentially transformative use [and] the result of a word search is different in purpose, character, expression, meaning and message from the page (and the book) from which it is drawn."[6] Even where the use was commercial, the transformative nature weighs in favor of fair use.[7]

Factor 2: Nature of the Copyrighted Work

Most of the cases that have been decided by the courts with respect to TDM have involved creative works, which are generally afforded greater protection. In such cases,

---

[6] Authors Guild v. HathiTrust, CITE

[7] *See Fox News v. TVEyes* CITE (S.D.N.Y. 2014)

some courts have suggested that this factor slightly favors the plaintiff. However, two courts have found this factor to be either neutral or favor the defendant. Regardless of how courts have looked at this factor, it has not afforded the second factor much weight.

Factor 3: Amount and Substantiality of Portion Used

TDM necessitates verbatim copying of the whole text or work into the database; without copying of the work in its entirety, a researcher could not effectively use TDM because potentially necessary portions of the work would not be analyzed. Courts have found that the verbatim copying of the whole work in databases for TDM is reasonable for the use and that this factor is neutral.

Factor 4: Effect on the Potential Market

Courts have found that the fourth factor tends to favor the user. The highly transformative nature of TDM suggests that the use is less likely to have an adverse impact on the market of the original because it is unlikely that the use would supersede the copyrighted work.

Weighing these four factors together, courts have upheld TDM as highly transformative uses that do not substitute for the original works, even with full verbatim copying of copyrighted works.

TDM and Best Practices in Fair Use

Existing fair use precedent has clearly endorsed the creation of databases for TDM because of its transformative nature. In addition, the *Code of Best Practices in Fair Use for Academic and Research Libraries*[8] addresses the issue of "creating databases to facilitate non consumptive research uses (including search)." The *Code* gives the following guidance:

> **PRINCIPLE:**
> It is fair use for libraries to develop and facilitate the development of digital databases of collection items to enable nonconsumptive analysis across the collection for both scholarly and reference purposes.
>
> **LIMITATION:**
>
> - Items in copyright digitized for nonconsumptive uses should not be employed in other ways (e.g., to provide digital access for ordinary reading) without independent justification, either by a license from the rights holder or pursuant to a statutory exception. Search access to database materials should be limited to portions appropriate to the nonconsumptive research purpose.

---

[8] http://www.arl.org/storage/documents/publications/code-of-best-practices-fair-use.pdf

ENHANCEMENTS:

- The case for fair use will be at its strongest when the database includes information such as rich metadata that augments the research or reference value of its contents.
- Assertions of fair use will be particularly persuasive when libraries cooperate with other institutions to build collective databases that enable more extensive scholarship or reference searching.

While this issue brief covers fair use and TDM in the United States, TDM is an issue of concern in other countries, as well.[9] Internationally, 171 organizations[10] including ARL, have called for the removal of barriers with respect to data, through the Hague Declaration on Knowledge Discovery in the Digital Age.[11] The Hague Declaration calls for clarity around the scope of intellectual property law as well as calling for better infrastructure to allow for content mining.

---

[9] The United Kingdom, for example, addressed TDM through implementation of a new copyright exception to allow researchers to make copies for the purpose of text and data mining for non-commercial research.

[10] As of June 4, 2015.

[11] The Hague Declaration, http://thehaguedeclaration.com/