A Bold Take on Copyright Implications of Text & Data Mining

Author : Pamela Samuelson

Date : April 1, 2021

Michael W. Carroll, Copyright and the Progress of Science: Why Text and Data Mining Is Lawful, 53 **UC Davis L. Rev.** 893 (2020).

Professor Carroll is not the first copyright scholar to have asserted that text and data mining (TDM) is and should be lawful as a matter of copyright law (and he probably won't be the last).[1] The hook that pulled me through the 72 pages of his excellent article was the introduction's announced intention to explain why use of TDM tools to run searches on digital repositories of infringing copies of copyrighted works do not infringe, at least as a matter of U.S. copyright law.

Text and data mining is a multi-stage technical process by which researchers compile and refine large quantities of text and other data so that it can be processed with statistical software to detect patterns that would be difficult or impossible for a human to perceive without the aid of the machine. The article considers the legality of TDM using SciHub as an exemplar. SciHub is a well-known repository of vast quantities of the scientific journal literature. Many scientists want to do TDM research using SciHub, but courts have held that that database is infringing. Although SciHub has more than once been forced to shut down, it has re-emerged every time and can still be found on the Internet.

Well-documented in this article, as well as in the technical literature to which Carroll copiously cites, is the promise of myriad scientific insights that researchers' use of TDM tools could unlock in a wide variety of fields. (For those not already conversant with TDM technologies, this article provides a very useful primer that is neither too nerdy nor too simplistic for lay readers to follow.) If promoting progress in science and useful arts continues to be copyright's constitutional purpose, the logical conclusion follows, Carroll intimates, that copying of in-copyright works to enable TDM research is and should be lawful.

Thanks to the Supreme Court's Eleventh Amendment jurisprudence[2] and the audacity of Google and the University of Michigan when agreeing to allow Google to scan all eight million books in the university's library in exchange for the library's getting back a digital copy, and thanks also to the Authors Guild for its unsuccessful lawsuits charging Google, the University of Michigan and its HathiTrust repository with copyright infringement, we know that digitally scanning in-copyright books for TDM and other non-consumptive purposes is non-infringing.

Carroll methodically works through each type of copying that happens in the course of collecting, formatting, processing, and storing data for TDM purposes. The article works through the relevant copyright case law for each type of copying that TDM involves. The ground over which the article travels will be familiar to many readers, but it provides a useful recap of how the law of digital copying has evolved over the last two decades.

Copyright is not, of course, the only potential obstacle to TDM research. Numerous proprietary publishers of scientific journals offer institutional database subscriptions to universities and other research institutions. However, those digital repositories are not interoperable. Researchers consequently cannot run searches across various databases. Cross-publisher collaborations are rare,

and the license terms on which databases are available may impair researchers' ability to make the full use of TDM tools. Publishers and the Copyright Clearance Center are promoting licensing of TDM as a value-added service and some of these licenses are more restrictive than TDM researchers would want.

One can understand why scientific researchers, even at institutions with institutional database subscriptions, would be attracted to using SciHub for TDM research. It is easier to use than some of the publisher repositories; the SciHub database is far more comprehensive than any of the proprietary databases; and there are no license restrictions to limit researcher freedom to investigate with TDM tools to their hearts' content.

Downloading SciHub seems a risky strategy for TDM researchers who do not want to be targets of copyright infringement lawsuits. Carroll argues that running TDM searches on the SciHub collection hosted elsewhere involves only the kind of transient copying that the Second Circuit found too evanescent to be an infringing "copy" of copyrighted television programming in the *Cartoon Networks* case. The results of the TDM research would be unprotectable facts extracted from the SciHub collection.

This is a bold assertion, which is well-documented. Read it for yourself to see if you agree.

1. *See, e.g.*, Edward Lee, *Technological Fair Use*, 83 **S. Cal. L. Rev**. 797, 846 (2010); Jerome H. Reichman & Ruth L. Okediji, *When Copyright Law and Science Collide: Empowering Digitally Integrated Research Methods on a Global Scale*, 96 **Minn. L. Rev.** 1362, 1368-70 (2012); Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 **NW. U. L. Rev.** 1607 (2009).
2. The Supreme Court has concluded that the Eleventh Amendment bars damage awards against states or state-related institutions. *See* Allen v. Cooper, 140 S.Ct. 494 (2020). The University of Michigan had reason to think that its endowment was safe from any lawsuit that might challenge its deal with Google as copyright infringement.

Cite as: Pamela Samuelson, *A Bold Take on Copyright Implications of Text & Data Mining*, JOTWELL (April 1, 2021) (reviewing Michael W. Carroll, *Copyright and the Progress of Science: Why Text and Data Mining Is Lawful*, 53 **UC Davis L. Rev.** 893 (2020)), https://ip.jotwell.com/a-bold-take-on-copyright-implications-of-text-data-mining/.