Copyright and the Progress of Science: Why Text and Data Mining Is Lawful

Michael W. Carroll[*]

This Article argues that U.S. copyright law provides a competitive advantage in the global race for innovation policy because it permits researchers to conduct computational analysis — text and data mining — on any materials to which they have access. Amendments to copyright law in Japan, and the European Union's recent addition of limitations on copyright to legalize some TDM research, implicitly acknowledge the competitive benefits provided by the fair use provision of U.S. copyright law.

Focusing only on U.S. law, this Article makes two general contributions to the literature on fair use: (1) in cases involving archiving, the user's security precautions are relevant under the first fair use factor and should not be treated as an unenumerated factor or as part of the market harm analysis; and (2) good faith should not be a factor in fair use analysis, but even if courts do consider good faith, TDM research conducted on infringing sources, such as Sci-Hub, is still lawful because the research provides transformative benefits without causing harm to the markets that matter. This Article also revisits the issue of temporary copies to argue that certain steps in TDM research do not make copies that "count" under U.S. law and that it is possible to design cloud-based TDM research that does not implicate U.S. copyright law at all. This Article addresses the needs of many audiences including policymakers, courts, university counsel, research libraries, and legal scholars who seek a thorough legal analysis to support this argument.

 Professor of Law and Faculty Director, Program on Information Justice and Intellectual Property, American University Washington College of Law. Thanks to Peter Jaszi, Matthew Sag, and Jonathan Band for helpful insights, to Samantha Primeaux, Tamara Celine Winegust, and Alan deLevie for research assistance, and to Dillon Jackson and Lucas Urgoiti of the *UC Davis Law Review* for editing above and beyond the call of duty. Nothing in this Article should be construed as legal advice.

TABLE OF CONTENTS

INTRODUCTION 895
I. TEXT AND DATA MINING 899
A. *The Social Value of Text and Data Mining* 901
B. *A Developing Field of Inquiry* 903
C. *An Exemplary TDM Project — DARPA's Big Mechanism* 905
D. *Summary* 907
II. JUDICIAL TREATMENT OF COMPUTATIONAL AND OTHER RESEARCH-RELEVANT USES 908
A. *Fair Use and Scientific Publishing* 908
1. The Roles of Research and Licensing 912
2. Clarifying the Role of Transformative Use 916
B. *The Copies that Count* 922
1. Temporary Copies v.1.0 923
a. *The Video Game Cases* 924
b. *RAM Copies* - MAI *Revisited* 926
2. Temporary Copies v.2.0 929
3. Congress Has Not Impliedly Amended the Fixation Requirement 933
III. TEXT AND DATA MINING IS LEGAL UNDER U.S. COPYRIGHT LAW 935
A. *Copying Journal Articles to Conduct and Validate TDM Research Is Fair Use* 936
1. The Paradigmatic Use — Compiling Data for Research and Retaining It for Reproducibility 940
a. *Copying to Enable Computational Research Is a Transformative Purpose* 941
b. *The Second and Third Factors Favor the Use* 945
c. *Copying to Conduct and Validate Research Does Not Affect the Markets that Matter* 946
d. *Data Security Is Relevant and Favors This Use* 950
2. Copying from an Infringing Source Necessary for TDM Research Is Still a Fair Use 951
a. *Sci-Hub* 951
b. *Text and Data Mining Sci-Hub Is Lawful* 954
B. *Most Copies for Computation Are Transitory* 959
CONCLUSION 963

INTRODUCTION

Can computers help scholars and scientific researchers better understand and analyze the published literature through computational analysis? Many researchers think so, and the field of so-called "text and data mining" ("TDM") research is fast evolving.[1] TDM research has broad application and is built upon uses embedded in our daily use of the internet. For example, the steps necessary to provide internet search engine services are commonly used forms of text and data mining of websites.

A simple and broad description of the TDM research discussed in this Article is as a multi-step process involving first the compilation of a dataset of text-based and related works into a format amenable to software-based statistical and related forms of pattern analysis. Researchers make multiple copies of the data during the TDM process. They make copies when they: (1) collect and compile the data; (2) format the data for computational processing; (3) process the data in a computer's active memory; and (4) store or archive the data to enable reanalysis or to enable validation through reproducing the analysis. In general, the outputs of this analysis report correlations, patterns or other relationships found in the information that has been mined, but little or none of the text, images or other forms of expression in the data appear in the TDM results.

Most TDM research relies on published books, articles, and other works covered by copyright law as the raw "data" used in computational processing. A central issue for researchers is whether, or how, copyright law applies to their work. This Article argues that TDM research is lawful in the United States because fair use enables the transformative benefits of TDM research and because copyright also has internal limits on the copies that "count."

This Article focuses only on the application of U.S. law to TDM research, recognizing that its topic is the subject of global policy competition to enable the next wave of scientific and technical innovation. Japan amended its copyright law to enable TDM research,[2] and recent changes in copyright law in the European Union that require member states to permit certain forms of TDM research also implicitly acknowledge this competition.

---

[1] This Article reluctantly uses "text and data mining" to designate computational research that might better be termed "computational research" or "content mining." *See infra* note 19 (citing sources and explaining reasons).

[2] *See* JEAN-PAUL TRIAILLE ET AL., DE WOLF & PARTNERS, STUDY ON THE LEGAL FRAMEWORK OF TEXT AND DATA MINING 10-12 (2014), https://www.fosteropenscience.eu/sites/default/files/pdf/3476.pdf (describing Article 47-7 of the Japan Copyright Act).

The roots of that latter change began in the United Kingdom, when former Prime Minister David Cameron initiated an innovation-focused review of copyright law in the United Kingdom.[3] The review report concluded that E.U. copyright law would not allow the United Kingdom to enable TDM research through a fair use provision because of the inflexibility of European Union law, but that U.K. law should provide a specific limitation on copyright to enable TDM research.[4] Parliament accepted this recommendation and adopted a new limit on copyright in 2014 to permit TDM research on a non-commercial basis.[5] The European Union recently followed the United Kingdom's example by adopting a European copyright directive that requires all member states to adopt a less robust user's right to engage in TDM research.[6]

This Article focuses on the United States' side of this policy competition to enable and to promote text and data mining as a means of gaining an innovative edge on a global scale. This Article addresses a range of audiences interested in the broader policy competition and cooperation in copyright law and those interested in text and data mining in particular. For most audiences in the United States, this Article delivers some welcome, and some less welcome, news. The good news is that U.S. copyright law does provide a user's right to do research through TDM.[7] Fair use is only part of the reason. Limits on the exclusive right to reproduce the copyrighted work adopted in the Copyright Act of 1976, as amended, also permit the transient copies made during computational processing without the need to resort to fair use.[8]

---

[3] *See* Andrew Orlowski, *Cameron's 'Google Review' Sparked by Killer Quote that Never Was*, REGISTER (Mar. 21, 2012, 1:01 PM), https://www.theregister.co.uk/2012/03/21/cameron_google_source/ [https://perma.cc/L9P6-XMWY].

[4] *See* IAN HARGREAVES, DIGITAL OPPORTUNITY: A REVIEW OF INTELLECTUAL PROPERTY AND GROWTH 42-43, 99 (2011), https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/32563/ipreview-finalreport.pdf.

[5] The Copyright and Rights in Performances (Research, Education, Libraries, and Archives) Regulations 2014, SI 2014/1372, art. 3, ¶ 2 (Eng.).

[6] *See* Directive 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, 2019 O.J. (L 130) 92, 113-14 [hereinafter DSM Directive].

[7] Treating fair use as a legal right rather than a legal defense is considered contentious by some. This use is explained and defended *infra* notes 72–78 and accompanying text.

[8] *See infra* Part II.

The less favorable news is that users' rights can be waived by contract.[9] Publishers of most scientific and scholarly journals that rely on a subscription revenue model rather than an open access publication model generally use this contractual power to limit researchers' ability to engage in text and data mining by imposing restrictions on access and use of their content in exchange for making this content available.[10] Although some publishers have cooperated to enable some cross-publisher TDM research through the Crossref consortium, from the researcher's perspective this solution is still only patchwork and technologically unnecessarily cumbersome. Crossref's Text and Data Mining services apply only to articles from publishers that have chosen to participate in the program. The researcher must first obtain the digital object identifier ("DOI") for each article they would like to analyze. They must then review publishers' varying text and data mining licenses and draw up a "white list" of licenses the researcher is willing to accept. Finally, the researcher must submit the list of DOIs and the license white list to Crossref to obtain access to the full text of the identified articles.[11]

For courts, counsel to universities and journal publishers, attorneys at U.S. government funding agencies, private funders of research, and university librarians, this Article's analysis explains in some detail why TDM is lawful without a license in the United States. It is this author's hope that this analysis will encourage librarians in particular to negotiate with vigor to eliminate or to reduce contractual restrictions on researchers' rights to engage in TDM.

This Article also contributes to the literature on fair use with two doctrinal arguments. First, when a use requires archiving multiple copyrighted works, courts appropriately should take account of the user's data security measures as part of the analysis under the first fair use factor, the purpose and character of the use, rather than as a freestanding unenumerated factor or under the fourth factor concerning

[9] *See, e.g.*, Bowers v. Baystate Techs., Inc., 320 F.3d 1317 (Fed. Cir. 2003) (holding enforceable a software license agreement requiring user to waive fair use rights to reverse engineer the software).

[10] *E.g.*, *Text and Data Mining*, ELSEVIER, https://www.elsevier.com/about/our-business/policies/text-and-data-mining (last visited Nov. 1, 2019) ("We have adopted a license-based approach which automatically enables researchers at subscribing institutions to text mine for non-commercial research purposes and to gain access to full text content in XML for this purpose.").

[11] *See Text and Data Mining for Researchers*, CROSSREF, https://support.crossref.org/hc/en-us/articles/214298826-Text-and-Data-Mining-for-Researchers (last visited Nov. 14, 2019).

the economic harm to the copyright owner.[12] Doing so appropriately contextualizes the risk analysis without giving it more weight than it deserves.

The second argument may attract greater controversy. This Article concludes that a researcher maintains the right to conduct computational research on the literature even when the material is copied from an infringing source. This argument has two subparts: (1) a user's good faith is irrelevant to the fair use analysis; and (2) even if good faith were relevant, a TDM researcher would be acting in good faith even when knowing that her sources are infringing because of the net social benefits of conducting TDM research.[13]

In particular, researchers have the right to use Sci-Hub,[14] which contains a very large infringing collection of the scientific literature, to text and data mine its corpus. Sci-Hub has been the target of a number of copyright infringement suits,[15] and this Article acknowledges that the claims that underlie the default judgments in those cases are meritorious under the Copyright Act of 1976.[16] But, transient copies made to mine that corpus either do not exercise the copyright owners' rights under Section 106(1)[17] or are covered by fair use, and maintaining a reference copy of the data mined for reproducibility purposes would also still be a fair use.[18]

This Article proceeds as follows: Part I provides a moderately technical description of text and data mining technologies and uses. This field is evolving rapidly, and the goal of this Part is to set forth the principles of computational research without cataloguing the wide range of tools in use or in development. Part II first analyzes the applicable legal precedent concerning copyright and scientific publications as well as the application of fair use to analogous forms of computational services. Recognizing the range of audiences interested in this issue, this Part provides brief summaries of the relevant fair use caselaw before synthesizing these cases and engaging in the scholarly literature around certain computational, or "non-expressive," uses of copyrighted works. This Part then revisits the issue of temporary copies to examine why transitory copies do not count as "copies" under U.S. law. Part III first argues that fair use permits a TDM researcher to make

[12] *See* 17 U.S.C. § 107 (2019).

[13] *See infra* Part III.A.2 (explaining why fair use does not require a lawful copy).

[14] *See infra* Part III.A.2 (explaining what Sci-Hub is and how it works).

[15] *See infra* Part III.A.2 (discussing litigation against Sci-Hub).

[16] 17 U.S.C. §§ 101-810 (1976) (current version at 17 U.S.C. §§ 101-1401 (2019)).

[17] *See infra* Part III.B.

[18] *See infra* Part III.A.1.

even non-transitory copies during processing, and also to archive the data that were processed, because of the beneficial and transformative purpose of TDM research and its negligible impact on the copyright owner's relevant economic interests. This Part then argues that many copies made during processing either currently are, or in the future will be, transitory copies that do not implicate copyright law at all, and that this conclusion may support cloud-based TDM research in the future. Part V concludes.

I. TEXT AND DATA MINING

Text and data mining[19] has broad applications that reach beyond scholarly and scientific research. The application that has garnered the most public attention has been privacy-invasive research conducted by social media platforms.[20] A less striking application, but one more important to internet users, is the process by which internet search engines index the World Wide Web and provide search services.[21] This Article focuses on computational research that is in a more nascent stage but which holds great promise for scientific, medical, and scholarly discoveries.

While favorable to technology companies, the recognition or creation of a right to engage in computational use and analysis of copyrighted works has created a kind of culture clash within publishing and other copyright-intensive industries. Traditional distributors view their

[19] This Article uses the term "text and data mining" or TDM as the term for computational analysis of publications and datasets because this is the term used in global policy conversations concerning the right to use copyrighted works for computational research. *See, e.g.*, Sergey Filippov & Paul Hofheinz, *Text and Data Mining for Research and Innovation: What Europe Must Do Next*, LIBSON COUNCIL (May 30, 2016), https://lisboncouncil.net/publication/publication/134-text-and-data-mining-for-research-and-innovation-.html [https://perma.cc/GBV3-M77J]; *see also* Marti A. Hearst, *Untangling Text Data Mining*, ASS'N FOR COMPUTATIONAL LINGUISTICS 3, https://www.aclweb.org/anthology/P99-1001.pdf (early article explaining possibilities for computational analysis of textual data). However, practitioners have identified seven different forms of computational research that are covered by the TDM terminology. *See* GARY MINER ET AL., PRACTICAL TEXT MINING AND STATISTICAL ANALYSIS FOR NON-STRUCTURED TEXT DATA APPLICATIONS 31-32 (2012).

[20] *See, e.g.*, Karen Weise & Sarah Frier, *If You're a Facebook User, You're Also a Research Subject*, BLOOMBERG (June 21, 2018, 10:13 AM), https://www.bloomberg.com/news/articles/2018-06-14/if-you-re-a-facebook-user-you-re-also-a-research-subject [https://perma.cc/8655-D2Q6] (describing Facebook's internal research and collaborations with academic researchers to investigate aspects of user behavior).

[21] *See* Dave Davies, *How Search Engine Algorithms Work: Everything You Need to Know*, SEARCH ENGINE J. (May 10, 2018), https://www.searchenginejournal.com/how-search-algorithms-work/252301/#close [https://perma.cc/2HX9-7ZTD].

professionally-created, curated works as individually valuable, and they generally express a view that in the digital age, the consuming public undervalues the skill, judgment, and expense involved in bringing these works to the public.[22]

But to technologists seeking to extract meaning or information from these works, it is all data.[23] Once digitized, these works become the raw material for computational analysis, and the relevant skills and judgment that they emphasize is the formulation and testing of algorithms that are most effective and efficient in carrying out the purpose of computational analysis.[24]

The most poignant episode in this clash in the United States has been the suits by commercial authors and publishers against Google[25] and the HathiTrust Digital Library[26] for digitizing and rendering searchable printed books. The legal issues presented by these cases are discussed in Parts II and III *infra*, but it is also important to recognize in these disputes a clash of perspectives and values implicated by digitizing and indexing the print culture of the twentieth century.

These authors and publishers thought it an outrage that Google and its partner libraries made digital copies of twenty million books to build a search service without seeking a license from them. The scale of such an undertaking was considered audacious.[27]

To the computer scientist, however, the scale of the project is quite modest. Twenty million digitized books require less than ten terabytes of data.[28] The computational challenge of processing that amount pales

---

[22] *See, e.g.*, Robert Levine, *'It's a System that is Rigged Against the Artists': The War Against YouTube*, BILLBOARD (May 5, 2016), https://www.billboard.com/articles/business/7356794/youtube-criticism-labels-artists-managers-payouts [https://perma.cc/Z3U4-2QAR] (arguing that musicians are underpaid for their contributions to YouTube's value).

[23] *See, e.g.*, CHUNLEI TANG, THE DATA INDUSTRY: THE BUSINESS AND ECONOMICS OF INFORMATION AND BIG DATA 2-3 (2016) (explaining emergence of data analytics in the economy).

[24] *See id.* at 11-12.

[25] Authors Guild, Inc. v. Google, Inc., 804 F.3d 202 (2d Cir. 2015).

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

[27] *See, e.g.*, Roxana Robinson, Editorial, *How Google Stole the Work of Millions of Authors*, WALL ST. J. (Feb. 7, 2016, 4:26 PM), https://www.wsj.com/articles/how-google-stole-the-work-of-millions-of-authors-1454880410 [https://perma.cc/DV8Y-DJBB] ("In 2004 Google sent its moving vans to the libraries and carted off some 20 million books. It copied them all, including books in copyright and books not covered by copyright. It asked no authors or publishers for permission, and it offered no compensation for their use.").

[28] *See* Tai Coromondel, *How Much Electronic Data Storage Would it Take to Store All Books that Ever Existed?*, QUORA (Nov. 28, 2015), https://www.quora.com/How-much-

in comparison to the resource requirements for indexing and algorithmically assessing the relevance of more than sixty trillion web pages that take over 100 petabytes[29] of data.[30] The law on both sides of the Atlantic and Pacific now sides with the user's perspective to some degree.

A. The Social Value of Text and Data Mining

For researchers in many fields, computational research of the published literature holds out great promise. Scholars in most fields, and particularly in the sciences, suffer from information overload. Too many potentially relevant journal articles are published each day for a scholar to find, read, and analyze.[31] Computational analysis can help the scholar sort through this information to identify those articles most relevant to the scholar. More importantly, a computer can independently process (read) all of these data to mine for patterns, concordances, and other relationships in the data that are, or potentially are, relevant to the scholar's field of inquiry.[32]

For some of the reasons explained below, TDM research remains in its early stages of development. The potential value it may add to scientific and scholarly research is quite significant. If TDM technologies simply aided a researcher's assessment of which articles were most relevant to her research question(s), they would provide a significant service that would save researchers collectively significant amounts of precious time.[33]

But, researchers developing TDM technologies have much more ambitious goals. Most TDM algorithms are related to technologies sometimes called colloquially "big data" or more precisely "machine

electronic-data-storage-would-it-take-to-store-all-books-that-ever-existed [https://perma.cc/S7BU-B2E3]. *See generally* Leonid Taycher, *Books of the World, Stand Up and Be Counted! All 129,864,880 of You*, GOOGLE BOOK SEARCH BLOG (Aug. 5, 2010, 8:26 AM), http://booksearch.blogspot.co.nz/2010/08/books-of-world-stand-up-and-be-counted.html [https://perma.cc/4Q8J-DNL6].

[29] One petabyte is 1024 terabytes.

[30] *See* Paul K. Young, *How Large is the Google Search Index, as of Mid-2016?*, QUORA (Sept. 27, 2016), https://www.quora.com/How-large-is-the-Google-search-index-as-of-mid-2016 [https://perma.cc/89ZR-ZK99].

[31] *See, e.g.*, Elisabeth Pain, *How to Keep Up with the Scientific Literature*, SCIENCE MAG. (Nov. 30, 2016, 4:00 PM), http://www.sciencemag.org/careers/2016/11/how-keep-scientific-literature [https://perma.cc/M9W5-PM5D] (interviewing scientists who struggle with information overload).

[32] *See id.* (describing tool for automating search).

[33] *See id.* (describing how slow readers are in sorting through relevant articles).

learning."[34] In the sciences, TDM algorithms are designed to analyze large amounts of data to identify patterns and correlations that can either directly or indirectly help to explain causal relations associated with the natural phenomena under investigation.

The promise of these technologies is that they may be able to identify patterns that would otherwise emerge only after years of trial-and-error experimentation or that may never be recognized. In its most ambitious form, a TDM technology would use semantic processing to be able to analyze these patterns to make judgments or conclusions that are as good or better than a skilled researcher's about the meaning and significance of these findings.[35] For experimentation that can be done *in silico* in virtual machine environments, TDM technologies could be written to identify patterns, formulate hypotheses about the causal relations that these patterns suggest, and then to formulate experiments that could be run, read, and refined through multiple iterations until a specified level of confidence in findings is reached.[36]

In the nearer term, these technologies can provide important inputs into how researchers formulate their research questions. For example, if a researcher were seeking to understand the role of a particular gene or set of genes in relation to a disease pathway, the researcher might run an analysis of all articles reporting results of experiments of any kind with this gene or these genes.[37] The results of this analysis may identify an unexpected correlation between the gene(s) and some other part of the body that would otherwise appear to be unrelated to the disease under study. This finding would likely lead the researcher to test whether any relation existed. In the best case, the results would show that the relation exists and how this relation works. This would unlock a solution that had been holding researchers back in the development of a relevant compound or gene therapy.[38]

An even larger promise for TDM research is that findings from TDM could open entirely new lines of research. In the above example,

---

[34] *See, e.g.*, Byoung-Tak Zhang, *Machine Learning Methods for Text / Web Data Mining*, https://bi.snu.ac.kr/Tutorials/ml4textweb00.pdf [https://perma.cc/D22Y-9AUW] (last visited Apr. 10, 2018) (describing methods for using machine learning to improve TDM).

[35] *See, e.g.*, JOHN D. KELLEHER ET AL., FUNDAMENTALS OF MACHINE LEARNING FOR PREDICTIVE DATA ANALYTICS: ALGORITHMS, WORKED EXAMPLES, AND CASE STUDIES (2015) (describing possibilities for predictive analytics on textual data).

[36] *See id.* at 126-30 (providing examples).

[37] *See, e.g.*, Ayush Singhal, Michael Simmons & Zhiyong Lu, *Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine*, PLOS COMPUTATIONAL BIOLOGY, November 2016.

[38] *See* KELLEHER ET AL., *supra* note 35, at 126-30.

imagine that the finding of a relation between genes and muscle function in an unexpected part of the body not only identified new drug targets but also revealed a more generalizable phenomenon about the interaction between certain types of genes and muscles. TDM research could help further develop this line of research.

When considering the potential knowledge that can be discovered using TDM, focus should not solely be placed on discrete deliverables such as finding the gene that is responsible for a certain disease. Text and data mining can be used to discover new lines of inquiry. In other words, TDM not only helps researchers find the right answers, but it can also help them find the right questions. For example, a TDM application may not be able to make a scientifically-concrete finding.

However, if this application reports that a certain unforeseen hypothesis has a 40% chance of being valid, researchers now have solid foundation from which to justify the direction of resources into this previously-unknown line of inquiry. Through the use of ontologies and other formalized systems of organizing knowledge, unforeseen associations between entities can be discovered.[39] For example, while there may be no literature that links A to C, TDM can help researchers discover links from A to B and from B to C. It could thus be said that the link from A to C could not have been made but for TDM.

B. A Developing Field of Inquiry

Text and data mining tools are used by scholars in all fields. While most attention has been focused on the promise of using TDM in the biological sciences to discover new lines of research that ultimately improve human health, other fields of inquiry, including the emergence of digital humanities as a distinct field, also rely on text and data mining.[40] There are different methodologies for conducting TDM research. The steps in a traditional hypothesis-driven model which relies on deductive reasoning are: (1) identify the general research

[39] *See, e.g.*, Steve Hardin, *Text and Data Mining Meets the Pharmaceutical Industry: Markus Bundschus Speaks*, 43 BULL. ASS'N INFO. SCI. & TECH. 42 (2017), https://onlinelibrary.wiley.com/doi/full/10.1002/bul2.2017.1720430314 (explaining how semantic analysis is used).

[40] *See* Brief of Digital Humanities and Law Scholars as Amici Curiae in Support of Defendant-Appellees and Affirmance at 2, Authors Guild, Inc. v. Google, Inc., 804 F.3d 202 (2d Cir. 2015) (No. 13-4829-cv) [hereinafter Brief of Digital Humanities] (explaining importance of TDM research for digital humanities scholars and arguing for its legality under copyright law); Alex H. Poole, *The Conceptual Ecology of Digital Humanities*, 73 J. DOCUMENTATION 91, 92-93 (2017) (describing 10 modes of digital humanities research).

question to be answered; (2) program queries, algorithms, or other forms of processing to be conducted, including defining the format for the outputs of such procession; (3) make copies of the digital data to be analyzed and combine them into a file or files to be analyzed; (4) format the data to enable or improve the effectiveness of computational processing; (5) run the programmed algorithms, which involves making temporary copies of the data in a computer's active memory; (6) store the outputs that result from step 5; and (7) store the data files from steps 3 and 4.[41]

There are too many variations on the above steps to do justice to researchers' creativity, but another general approach is to reason inductively by using various algorithms on data without any preconceptions about whether statistically significant patterns may emerge and then analyzing the results for such patterns.

From the perspective of copyright law, many of the differences among research approaches and fields of inquiry are irrelevant. The steps to which copyright law may apply are the making of copies to be analyzed, the reformatting of that data for analysis, potentially the temporary copies made during analysis, the outputs of the analysis if enough copyrightable expression is contained in those outputs, and the archival copies of the data.

This Article focuses primarily on the use of scholarly and scientific publications, usually journal articles, as the data being analyzed by the researcher. Most of the legal analysis that follows would apply equally to TDM conducted on any publications or on datasets as well, but nuances could emerge in cases in which the reformatting of the data could be considered the preparation of a derivative work or if the outputs of the analysis contain significant amounts of copyrightable expression. To avoid overgeneralizing, this Article specifies the steps in the TDM process to which the legal analysis applies.

TDM researchers also experience subscription publishers' use of copyright law to limit access and use of their publications as a barrier to their access to publications as source data. Copyright law does not provide an exclusive right of access as such, but copyright owners generally use the exclusive rights on how their works are used to also set the terms of access. Access to the full text of biomedical journal

[41] *Cf.* NATIONAL ACADEMIES OF SCIENCE, ENGINEERING AND MEDICINE, OPEN SCIENCE BY DESIGN: REALIZING A VISION FOR 21ST CENTURY RESEARCH 108-11 (2018) (generally describing the hypothesis-driven research process); Sarah Lucy Cooper, *The Collision of Law and Science: American Court Responses to Developments in Forensic Science*, 33 PACE L. REV. 234, 301 (2013).

articles has been challenging for researchers,[42] who have instead applied their tools only to article abstracts available in the National Institutes of Health's Public Library of Medicine. A recent study demonstrates that access to full text journal articles improves the outcomes and utility of TDM tools.[43] Relevant to the discussion below about researchers' need to keep reference copies of the articles mined, the authors of this study stated in their data availability statement that "[d]ue to copyright and legal agreements the full text articles cannot be made available."[44] While the authors were able to share the digital object identifiers for the fifteen million articles that they had mined, another researcher would be unable to duplicate this research without access to the full-text articles in the dataset. The promise of text mining to speed scientific progress has been unfulfilled in part because of limits on access to full-text articles. Like other forms of machine learning, TDM technologies develop through repeated experiments with large amounts of data.

C. An Exemplary TDM Project — DARPA's Big Mechanism

Text mining of scientific articles refers to the processes by which information is located, extracted, and interpreted from hundreds of articles and synthesized into causal models.[45] The Defense Advanced Research Projects Agency ("DARPA") established the Big Mechanism program in 2014 to read and interpret information and data regarding cancer biology, due to the impracticability of researchers keeping up with the vast amount of scientific literature.[46] As one observer explains, "[t]he Big Mechanism program aims to develop technology to read research abstracts and papers to extract pieces of causal mechanisms, assemble these pieces into more complete causal models, and reason

[42] *See* Petr Knoth & Nancy Pontika, *Aggregating Research Papers from Publishers' Systems to Support Text and Data Mining: Deliberate Lack of Interoperability or Not?*, *in* CROSS-PLATFORM TEXT MINING AND NATURAL LANGUAGE PROCESSING INTEROPERABILITY 1-2 (Richard Eckart de Castilho et al. eds., 2016), http://oro.open.ac.uk/46870/1/INTEROP-1.pdf.

[43] *See* David Westergaard et al., *A Comprehensive and Quantitative Comparison of Text-Mining in 15 Million Full-Text Articles Versus Their Corresponding Abstracts*, PLOS COMPUTATIONAL BIOLOGY, February 2018, at 1.

[44] *Id.* at 1.

[45] *See Big Mechanism*, NAT'L CTR. FOR TEXT MINING, http://www.nactem.ac.uk/big_mechanism/ (last visited Nov. 5, 2019) [https://perma.cc/97KT-MVDA]. Hat tip to Michael Madison for pointing me to this project.

[46] *See* Paul R. Cohen, *DARPA's Big Mechanism Program*, PHYSICAL BIOLOGY, July 2015, at 1. The Big Mechanism program assembles causal assertions from many sources into large models, allowing scientists to understand complex systems. *See id.* at 7.

over these models to produce explanations."[47] Big Mechanism focuses on systems that have a multiplicity of elements and relationships not easily comprehended by people, concentrating specifically on Ras cancers.[48] Although the Big Mechanism program has a narrow focus on cancer biology, the researchers conducting this research believe that this technology could extend into other areas of scientific research.[49]

DARPA's Big Mechanism operates by using mechanistic models, rather than syntactic or semantic models, to make claims about mechanisms, measurements, and observations.[50] These models relate inputs by shallow reading, which "discovers the entities, relations, events, and processes in text," without considering what the text contains about prior models.[51]

Big Mechanism machines read source texts and then "the content of these texts will suggest revisions to prior models. Machine reasoning about whether and how to modify models is called assembly."[52] Afterward, Big Mechanism will produce explanations and predictions; "for example, finding similar models, or finding generalities across models, or identifying potentially druggable proteins or pathways."[53]

[47] Joshua Elliott, *Big Mechanism*, DARPA, https://www.darpa.mil/program/big-mechanism (last visited Nov. 5, 2019).

[48] *See* Cohen, *supra* note 46, at 1.

[49] *See, e.g.*, *id.* at 7 (noting that "the BMP is not a big data program, a cancer therapy program, or even a cancer biology program" but rather "a program designed to develop technologies to help scientists understand very complicated systems, generally").

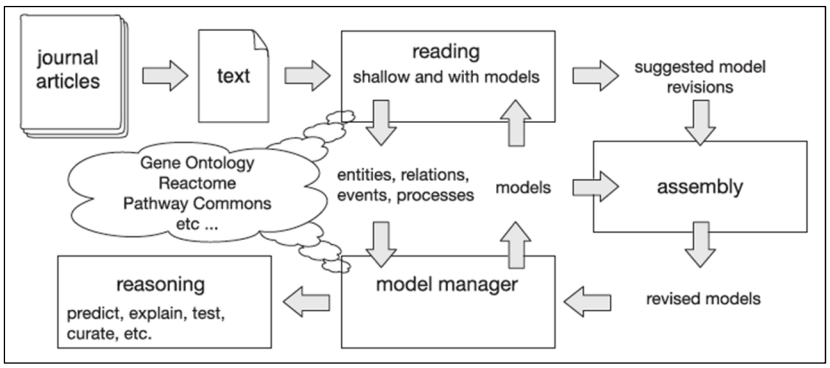
[50] *See id.* at 1-2. For a visual representation of how Big Mechanism reads and interprets journal articles, see Figure 1 *infra*.

[51] *Id.* at 3.

[52] *Id.*

[53] *Id.*; Jia You, *DARPA Sets Out to Automate Research*, 347 SCIENCE 465, 465 (2015).

Figure 1. A rough architecture for Big Mechanism systems[54]



D. Summary

Text and data mining technologies promise to unlock new lines of research and to reveal new correlations and patterns in the scientific and scholarly literature that would not be otherwise discoverable. To work effectively, these technologies require access to large numbers of full-text journal articles. Such access is readily obtained for journals that publish their content with open access, meaning that the journal is freely available for download and reuse.[55] Currently, open access publications comprise about 28% of all published articles, and the trend is that this proportion is increasing.[56] Publishers of subscription-based journals restrict access to their publications, although some of them have developed licenses for text and data mining.[57] Cross-publisher collaborations to provide access to subscription-based content are incomplete, technologically cumbersome, and do not provide

[54] Cohen, *supra* note 46, at 3.

[55] *See, e.g.*, *The Right to Read Is the Right to Mine*, PLoS, https://www.plos.org/text-and-data-mining (last visited Nov. 14, 2019).

[56] *See* Heather Piwowar et al., *The State of OA: A Large-Scale Analysis of the Prevalence and Impact of Open Access Articles*, NAT'L CTR. FOR BIOTECH. INFO. (Feb. 13, 2018), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5815332/.

[57] *See Copyright Clearance Center Launches Text Mining Solution*, COPYRIGHT CLEARANCE CTR., http://www.copyright.com/copyright-clearance-center-launches-text-mining-solution/ [https://perma.cc/5FPJ-FKAD] (last visited Apr. 10, 2018) (offering researchers the right to download files in XML format from participating publishers); *see also* ELSEVIER, *supra* note 10.

researchers with comprehensive access to the published literature for purposes of conducting TDM research.[58]

II. JUDICIAL TREATMENT OF COMPUTATIONAL AND OTHER RESEARCH-RELEVANT USES

Copyright law strikes a balance between exclusive rights granted to authors and the rights of users reflected either in the definitions of copyright's scope or in specific limitations or exceptions to the author's exclusive rights. For most TDM researchers engaged in current research practices, fair use is the user's right that provides the legal justification for their use of copyrighted works. But, there is also a right to make transitory copies of works that are definitionally outside the scope of copyright law in the United States. Currently, this limit applies to the temporary copies made during processing in some instances. It is foreseeable that some forms of TDM research could rely entirely on this limit on temporary copies in the future, with increases in computing power expected and increasing reliance on storing data with cloud services. For these reasons, this Part first reviews the legal sources relevant to whether TDM research is protected by fair use and then turns to the current state of the law with respect to temporary copying of copyrighted works.

A. Fair Use and Scientific Publishing

Computational research is a scientific endeavor. Basic principles of scientific practice require that researchers should be able to test the validity of empirical claims by using the underlying inputs into the research to reproduce its results.[59] Reproducibility has been a broader

[58] *See, e.g.*, *Text and Data Mining Services*, CROSSREF, https://support.crossref.org/hc/en-us/articles/215750183-Crossref-Text-and-Data-Mining-Services (last visited Nov. 14, 2019) (providing aggregated metadata search and download process that requires researchers to independently identify each individual article to be mined and is limited to participating publishers' journals); *RightFind® XML for Mining*, COPYRIGHT CLEARANCE CTR., http://www.copyright.com/business/xmlformining/ (providing access to XML-formatted versions of journal articles for text and data mining limited to content submitted by participating publishers) (last visited Nov. 14, 2019).

[59] *See, e.g.*, Victoria Stodden et al., *An Empirical Analysis of Journal Policy Effectiveness for Computational Reproducibility*, 115 PROCS. NAT'L ACAD. SCI. 2584, 2584 (2018) ("For computational and data-enabled research, [reproducibility requirements] ha[ve] often been interpreted to mean making available the raw data from which results were generated, the computer code that generated the findings, and any additional information needed such as workflows and input parameters.") (alteration in original).

concern in the scientific community in recent years.[60] Reliance on software models that are unavailable for reproducibility testing has exacerbated this problem. Most researchers engaged in text and data mining do so as scientists who follow these basic scientific practices.[61] Once they have defined the problem(s) or question(s) they seek to research, they need to make copies of the relevant sources, reformat those sources, make temporary copies necessary for running relevant algorithms and then keep a reference copy of their data to make their research reproducible and, often, to satisfy the data management terms and conditions of their funding agreements.[62]

A researcher who makes and retains copies of the data to be analyzed and who makes non-transitory temporary copies during processing would exercise the copyright owners' rights to make reproductions of their works.[63] Sharing these copies with other researchers would exercise the exclusive right to distribute copies to the public.[64] A researcher engaged in these acts would be liable for copyright infringement unless they are covered by the doctrine of fair use.[65] Part III of this Article analyzes this issue, and this Part focuses on the applicable precedent necessary for that analysis.

Fair use originated as a judicially-created limit on the scope of copyright that was formalized into a four-factor doctrine that Congress codified in Section 107 of the Copyright Act of 1976.[66] Section 107 provides that "[n]otwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies . . . for purposes such as . . . scholarship, or research, is not an infringement of copyright."[67] This provision taken

---

[60] *See* DAVID RANDALL & CHRISTOPHER WELSER, NAT'L ASS'N OF SCHOLARS, THE IRREPRODUCIBILITY CRISIS OF MODERN SCIENCE: CAUSES, CONSEQUENCES, AND THE ROAD TO REFORM 7-8 (2018), https://www.nas.org/storage/app/media/Reports/Irreproducibility%20Crisis%20Report/NAS_irreproducibilityReport.pdf (providing details about the inability of researchers to reproduce the results reported in most published scientific articles).

[61] *See* Filippov & Hofheinz, *supra* note 19, at 2.

[62] *See, e.g.*, SPARC, *Browse Article and Data Sharing Requirements by Federal Agency*, http://datasharing.sparcopen.org/compare?ids [https://perma.cc/3P47-XUFV] (last visited Nov. 5, 2019) (discussing resource collecting data management and sharing requirements of U.S. government funding science funding agencies); *cf.* DSM Directive, *supra* note 6, art. 3, at 113 (permitting copies made for TDM to "be retained for the purposes of scientific research, including for the verification of research results").

[63] *See* 17 U.S.C. § 106(1) (2019).

[64] *See id.* § 106(3).

[65] *See id.* § 107.

[66] *Id.*

[67] *Id.*

alone could be read to mean copying scholarly and research purposes is categorically, or at least presumptively, a fair use. However, the Supreme Court has declared that no uses are presumptively fair uses.[68] Instead, to determine whether a use is a fair use, four factors must be considered. They are:

> (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.[69]

Fair use balances the author's exclusive rights in their works with a user's right to make certain uses of those works without a license. For purposes of this Article, the question is whether fair use provides a right to research. The nomenclature of "user's rights" in copyright law is contested. Without fully detailing the arguments, this Article treats fair use as a user's right advisedly.

Formally, fair use is one of a series of limitations on, or exceptions to, the copyright owner's rights set forth in Section 106(a), which grants these rights "subject to sections 107 through 122."[70] These sections identify a set of uses that would otherwise fall within the scope of § 106(a), but which Congress has decided are "not an infringement of copyright."[71] Those who object to the "user's rights" designation for fair use argue that merely because a use is "not an infringement" does not mean that the user has an affirmative legal right to make the use.

This is sophistry for two reasons. First, in the Anglo-American legal tradition, liberty means that acts that are not prohibited by law are lawful.[72] This is in contrast to the civil law tradition of some countries in which the legal code provides a comprehensive statement of applicable rights and duties.[73] Therefore, in the Anglo-American

[68] *See* Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 594 (1994) (stating that fair use factors are to be balanced without presumptions).

[69] 17 U.S.C. § 107.

[70] *Id.* § 106.

[71] *Id.* § 107.

[72] *See* Rogers v. Tennessee, 532 U.S. 451, 467 (2001) (Stevens, J., dissenting).

[73] *See, e.g.*, John Y. Gotanda, *Recovering Lost Profits in International Disputes*, 36 GEO. J. INT'L L. 61, 71-73 (2004).

tradition, the user has a positive right to engage in conduct that is "not an infringement" of the exclusive rights under copyright law so long as this conduct is not otherwise prohibited by some other source of law.

Second, the possibility that other sources of law can limit a fair use is quite narrow because fair use is grounded in a well-known fundamental right to freedom of expression. Because copyright empowers a court to enjoin a user's expression and to seize and destroy books and other forms of expression,[74] the exclusive rights under copyright would be an infringement of a user's free speech rights.[75] Copyright law coexists with free speech because the Constitution authorizes Congress to give authors "the exclusive Right to their respective Writings."[76] Therefore, the user's right to free speech and the author's exclusive rights coexist in balance: a user has a First Amendment right to express herself using another's expression unless this expression is protected by one or more of the exclusive rights provided by the Copyright Act of 1976. The Court has recognized that user's free speech rights under the First Amendment impose structural limits on Congress's copyright-granting power, but the Court is satisfied that existing statutory limits, such as the idea/expression dichotomy and fair use, sufficiently guard user's free speech rights such that further elaboration is not currently needed.[77] Because the fair use doctrine provides that a fair use is "not an infringement" of the author's rights set forth in Section 106(a), fair uses fall within the user's rights protected by the First Amendment.[78]

Within this framework, the relevant issue for this Article is whether a researcher has a fair use right to reproduce a group of books, scientific journal articles, datasets, or other research outputs for purposes of computationally analyzing them and keeping these copies for further computational research or to share with other researchers who seek to reproduce or extend this computational research. The courts have previously ruled on whether a researcher may make copies of journal

---

[74] *See* 17 U.S.C. § 503 (authorizing impoundment and destruction of infringing copies as remedies for infringement).

[75] *See, e.g.*, Mark A. Lemley & Eugene Volokh, *Freedom of Speech and Injunctions in Intellectual Property Cases*, 48 DUKE L.J. 147, 169 (1998) (arguing for narrow use of injunctions in copyright cases to limit conflicts with the First Amendment).

[76] U.S. CONST. art. I, § 8, cl. 8.

[77] *See* Golan v. Holder, 565 U.S. 302, 328 (2012); Eldred v. Ashcroft, 537 U.S. 186, 219 (2003).

[78] *Golan*, 565 U.S. at 328; *Eldred*, 537 U.S. at 219; *cf.* CCH Canadian Ltd. v. Law Soc'y of Upper Canada, [2004] 1 S.C.R. 339, para. 48 (Can.) (holding that "the fair dealing exception . . . is a user's right. In order to maintain the proper balance between the rights of a copyright owner and users' interests, it must not be interpreted restrictively").

articles for research purposes, but the factual and legal contexts of those disputes differ in certain respects from storing and sharing copies for text and data mining purposes.

1. The Roles of Research and Licensing

Courts and advocates are likely to rely on these precedents for most of the steps in the fair use analysis. In particular, parties to a dispute would likely read these cases differently with respect to whether a copyright owner's willingness to license text and data mining undermines a user's reliance on fair use. The courts have found research to be a favored use, but they have also attended to the issue of licensing.

In *Williams & Wilkins Co. v. United States*,[79] a publisher of medical journals sued the United States for copyright infringement, claiming that the photocopies of journal articles made by the staff of the National Library of Medicine ("NLM") for intramural researchers at the National Institutes of Health ("NIH"), and for researchers requesting copies through interlibrary loan agreements, exceeded the limits of fair use.[80] NLM purchased two subscriptions to the journals in suit, keeping one copy of each issue in the library and routing the other copy among researchers.[81] NLM had adopted an internal policy that limited the number of copies of articles that individual researchers and external researchers could request through their libraries.[82]

Although the case arose under the Copyright Act of 1909, the court applied the same judicially-created factors that Congress later codified in 17 U.S.C. § 107 to hold that the copying practices that conformed with NLM's internal policy were fair use.[83] The court determined that the publisher had not offered evidence to show that NLM's photocopying was causing it economic harm, other than its stated willingness to license article photocopying, and that an injunction against NLM's photocopying would interfere with medical research.[84]

The court emphasized that most of the copying was done to aid the noncommercial research purposes of medical and scientific researchers and that "it is settled that, in general, the law gives copying for scientific

[79] 487 F.2d 1345 (Ct. Cl. 1973), *aff'd by an equally divided Court*, 420 U.S. 376 (1975).

[80] *See id.* at 1346-47.

[81] *See id.* at 1347-48.

[82] *See id.* at 1349 (detailing use restrictions on photocopying service).

[83] *See id.* at 1352-53.

[84] *See id.* at 1354.

purposes a wide scope."[85] Photoduplication had been a long-accepted practice until an increase in its scale during the 1960s prompted publisher complaints.[86] The court relied on plaintiff's concession and on the trial testimony of requesting researchers and librarians that in the absence of photocopying, the supply of reprints was wholly inadequate to meet researchers' demand and that medical research would be seriously impeded.[87]

Having determined that plaintiff failed to demonstrate any harm to its subscription revenues, the court also rejected the plaintiff's proposed remedy of a prospective reasonable royalty on the grounds that the 1909 Act did not authorize compulsory licensing as a remedy.[88] The court also cast doubt on the viability of plaintiff's photocopying licensing system and on the question whether voluntary collective licensing would be possible in the absence of new legislation. While this decision drew a lone dissent in the appellate court, the Supreme Court found the case more difficult, affirming the decision through a 4-4 split vote.[89]

Following the enactment of the Copyright Act of 1976,[90] journal publishers acted to make their nascent collective licensing system a reality through the founding of the Copyright Clearance Center ("CCC") in 1977.[91] CCC offered journal subscribers a photocopying license to cover the copying of articles from these journals.[92]

Seeking a test case that would yield a contrary result to *Williams & Wilkins,* the publishers found one in *American Geophysical Union v. Texaco, Inc.*[93] Through a stipulation, the parties narrowed the issues for trial to whether photocopying articles from one of the plaintiff's journals by a single Texaco scientist chosen at random was a fair use. The case went to trial over the copying of eight articles from the selected journal.[94] Texaco had initially purchased a single subscription of the journal for its New York facility, later increasing that to three subscriptions.

---

[85] *Id.* While most of the copying was for university or governmental researchers, the record demonstrated that about 12% of the external copying requests came from private organizations, primarily drug companies. *See id.* at 1349.

[86] *See id.* at 1356.

[87] *See id.*

[88] *See id.* at 1359-60.

[89] *See* Williams & Wilkins Co. v. United States, 420 U.S. 376, 376 (1975).

[90] Copyright Act of 1976, Pub. L. No. 94-553, 90 Stat. 2541.

[91] *See* Copyright Clearance Ctr. v. Comm'r, 79 T.C. 793, 794 (1982) (stating that CCC was incorporated in July 1977).

[92] *See id.* at 798 (describing CCC's operation).

[93] 60 F.3d 913 (2d Cir. 1994).

[94] *See id.* at 915.

The two material factual differences from *Williams & Wilkins* were that the plaintiffs offered photocopying licenses through the newly-created CCC licensing system and that the defendant was a commercial entity whose research was directed toward increasing its profits. Emphasizing the importance of this latter fact, the Second Circuit stated: "We do not deal with the question of copying by an individual, for personal use in research or otherwise (not for resale), recognizing that under the fair use doctrine or the *de minimis* doctrine, such a practice by an individual might well not constitute an infringement."[95]

In analyzing the four fair use factors, the court determined that the purpose and character of the use was for the scientists to archive articles for future reference. Texaco had argued that this was a transformative use because the copies allowed its scientists to access an article's content in the laboratory.[96] The court indicated that such a use may well be transformative, but the record did not provide enough support and that archiving photocopies for personal convenience did not make the use transformative.[97] As a result, the court held that this factor favored the plaintiff because the use was not transformative and that the ultimate purpose of the scientist's research was to contribute to Texaco's profits.[98] The second and third factors were largely immaterial to the court's analysis.[99]

With respect to the effect on the plaintiff's market, the court held that the photocopying may have slightly diminished subscription revenue but more significantly the court held that this factor favored the plaintiff because of the impact on licensing revenue.[100] Not all claims of lost licensing revenue will succeed. "Only an impact on potential licensing revenues for traditional, reasonable, or likely to be developed markets should be legally cognizable" under the fourth factor.[101] The court determined that the creation of the CCC had established a "workable market" that tipped this factor in the plaintiff's direction.[102]

On this point, the parties' agreement to limit the factual issues for trial worked against the defendant. By having chosen one scientist and one

[95] *Id.* at 916.

[96] *See id.* at 918-20.

[97] *See id.* at 923-24.

[98] *See id.* at 924-25.

[99] *See id.* at 925-26 (giving little independent weight to these factors and emphasizing that the amount of use under the third factor is, in effect, a further elaboration of the nature and purpose of the use under the first factor).

[100] *See id.* at 928-31.

[101] *Id.* at 930.

[102] *See id.*

journal as representative of Texaco's photocopying practices, the court emphasized that its holding on the fourth factor turned on the fact that CCC had the authorization to offer a photocopying license for that journal. The court explicitly reserved the question of how the analysis would turn out were this license not available.[103]

In a vigorous dissent, Judge Jacobs would have found that the photocopying was transformative because it was an intermediate step in carrying out research[104] and that the CCC did not provide a workable licensing market.[105] In particular, he pointed out that the court erred in grounding its holding on the availability of a CCC license for the particular works in suit when the parties' stipulation also demonstrated that: "(a) institutions such as Texaco subscribe to numerous journals, only 30 percent of which are covered by a CCC license; (b) not all publications of each CCC member are covered by the CCC licenses; and (c) not all the articles in publications covered by the CCC are copyrighted."[106]

Recognizing the merits of many of Judge Jacobs' objections, the court emphasized the narrowness of its holding: "Our ruling is confined to the institutional, systematic, archival multiplication of copies revealed by the record — the precise copying that the parties stipulated should be the basis for the District Court's decision now on appeal and for which licenses are in fact available."[107]

This decision's emphasis on "systematic" and "institutional" copying led some authors and publishers to read the court to have determined that if these facts are present, such copying cannot be fair use.[108] However, the law is to the contrary. The next Part discusses how the Second Circuit and the Ninth Circuit have held that fair use protects certain forms of systematic and institutional copying if it is done for a transformative purpose or when there is no economically feasible licensing market for a non-transformative use.

---

103 *See id.* at 931.

104 *See id.* at 935 (Jacobs, J., dissenting) (characterizing the scientist's photocopying as enhanced note-taking for the transformative purpose of providing an input to new research).

105 *See id.* at 937.

106 *Id.*

107 *Id.* at 931 (majority opinion).

108 *Cf.* Brian T. Ster, *Photocopying and Fair Use: Exploring the Market for Scientific Journal Articles*, 30 IND. L. REV. 607, 622 (1997) (noting that "even a CCC license would not have guaranteed protection to Texaco, or any other institutional user, from copyright infringement for photocopying a given article").

2. Clarifying the Role of Transformative Use

In a series of cases involving digital technologies, the federal courts have held that fair use permits conducting computational analysis and creating a digital archive to enable search services, and this reasoning supports archiving by TDM researchers. The Ninth Circuit held in *Kelly v. Arriba Soft Corp.*[109] and in *Perfect 10, Inc. v. Amazon.com, Inc.*[110] that systematic and institutional copying of images for the transformative purpose of providing a commercial image search service is a fair use. The Fourth Circuit reasoned analogously that archiving student research papers to provide a plagiarism search service also was a transformative use.[111]

With respect to the fourth fair use factor, in *Arriba Soft*, the court held that Kelly's market for his full-size professional photographs was not harmed by Arriba Soft's lower-resolution thumbnail versions of the photographs.[112] In *Perfect 10*, the plaintiff sought to distinguish its claim against image search services on the grounds that Google's and Amazon's thumbnail images competed with its market to license its photographs for display on mobile devices.[113] The court disagreed, holding that the plaintiff had presented no evidence to show that mobile users had relied on the thumbnails provided by image search services as a substitute for purchasing a licensed download.[114] Having determined that the use, though commercial, was highly transformative even if it presented a risk of market substitution, the court held that the fourth factor favored neither party and therefore the use was fair.[115]

The Second Circuit has also held that institutional and systematic copying is fair use in two cases arising from Google's Books project, for which it copied the full text of millions of books from academic libraries to provide an online search tool.

For its Books project, Google entered into agreements with partner libraries and other non-profit institutions through which it would scan the full contents of millions of volumes in their respective collections in exchange for providing each library with a digital copy of each book

109 Kelly v. Arriba Soft Corp., 336 F.3d 811, 822 (9th Cir. 2003).

110 Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1169 (9th Cir. 2007).

111 *See* A.V. *ex rel.* Vanderhye v. iParadigms, LLC, 562 F.3d 630, 640 (4th Cir. 2009).

112 *Arriba Soft Corp.*, 336 F.3d at 821-22.

113 *Perfect 10, Inc.*, 508 F.3d at 1168.

114 *See id.*

115 *See id.*

scanned.[116] *Authors Guild, Inc. v. HathiTrust*[117] involved a suit against a collaboration among eighty libraries and other institutions to combine the digital copies each had received from Google to create the HathiTrust Digital Library ("HDL"), with a collection of more than ten million works.[118]

The HDL made three uses of these copies. First, it provided a search service that allowed a patron to identify relevant works responsive to her query. The search results displayed only the page number(s) of responsive works on which the search term(s) appeared. Second, for patrons with certified print disabilities, HDL members could make the full text of works available through adaptive technologies. Third, a member could make a replacement copy for its collection if it originally held the title, that copy had been lost, destroyed, or stolen, and a replacement copy was unavailable at a fair price.[119]

The court affirmed the district court's holding that the first two uses were fair use and vacated the judgment with respect to the third because no evidence had been provided to indicate that an HDL member had made a replacement copy of any of the plaintiffs' copyrighted works.[120] The court held that institutional and systematic copying to provide full-text search is a transformative use and that this diminishes the role of the fourth factor.[121] The fourth factor focuses solely on "the harm that results because the secondary use serves as a substitute for the original work."[122] As a result, "under Factor Four, any economic 'harm' caused by transformative uses does not count because such uses, by definition, do not serve as substitutes for the original work."[123] The court rejected

---

116 *See* Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 90 (2d Cir. 2014) (describing arrangement).

117 *Id.*

118 *See id.* at 90.

119 *See id.* at 91-92.

120 *See id.* at 104.

121 *See id.* at 97 (characterizing building the search index as a "quintessentially transformative use"). It is noteworthy that the court reached this conclusion after having received briefing about the potential use of the HDL for text and data mining. *See, e.g.*, Reply Memorandum in Support of the Libraries' Motion for Summary Judgment on Fair Use and Lack of Infringement under Section 106 of the Copyright Act, Authors Guild, Inc. v. HathiTrust, 755 F.3d 87 (2d Cir. 2014) (No. 11 Civ. 6351); Brief of Digital Humanities and Law Scholars as Amici Curiae in Partial Support of Defendants' Motion for Summary Judgment, Authors Guild, Inc. v. HathiTrust, 755 F.3d 87 (2d Cir. 2014) (No. 11 Civ. 06351).

122 *HathiTrust*, 755 F.3d at 99 (citing Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 591 (1994)).

123 *Id.* (citing Bill Graham Archives v. Dorling Kindersley Ltd., 448 F.3d 605, 614 (2d Cir. 2006)).

the plaintiffs' claim that permitting wholesale copying to provide full-text search would inhibit the emergence of a market to license such copying.[124]

While the court disagreed that providing copies for use with adaptive technologies was a transformative use,[125] it nonetheless determined that expanding access for persons with print disabilities is fair because it is a favored use and the effect on the market is negligible.[126]

In *Authors Guild, Inc. v. Google, Inc.* (*Google Books*),[127] the court turned to whether Google's systematic and institutional copying of books to provide full-text search that yields snippets (i.e., quotations) of text containing the search term(s) is fair use. The court first refined the concept of transformativeness under the first factor, writing that "transformative uses tend to favor a fair use finding because a transformative use is one that communicates something new and different from the original or expands its utility, thus serving copyright's overall objective of contributing to public knowledge."[128]

The copying to provide a search service is transformative because it does not provide the same information as the copied work but instead provides information about these works. The fact that Google's motivation is commercial does not distinguish this case from *HathiTrust* because "[o]ur court has . . . repeatedly rejected the contention that commercial motivation should outweigh a convincing transformative purpose and absence of significant substitutive competition with the original."[129] The quotations in the snippets were not longer than necessary to provide a searcher with enough context to gauge whether the work is relevant and Google took reasonable steps to prevent a searcher from piecing an entire, or even a substantial portion, of a searched work together through repeated searches.[130]

With respect to the fourth factor, the court emphasized that it is not to be considered in isolation: "*Campbell* stressed the close linkage between the first and fourth factors, in that the more the copying is done

[124] *See id.* at 100 ("Thus, it is irrelevant that the Libraries might be willing to purchase licenses in order to engage in this transformative use (if the use were deemed unfair). Lost licensing revenue counts under Factor Four only when the use serves as a substitute for the original and the full-text-search use does not.").

[125] *See id.* at 101.

[126] *See id.* at 102-03.

[127] Authors Guild, Inc. v. Google, Inc. (*Google Books*), 804 F.3d 202 (2d. Cir. 2015).

[128] *Id.* at 214.

[129] *Id.* at 219 (citing Cariou v. Prince, 714 F.3d 694, 703 (2d Cir. 2013), *cert. denied*, 571 U.S. 1018 (2013)).

[130] *See id.* at 222-23.

to achieve a purpose that differs from the purpose of the original, the less likely it is that the copy will serve as a satisfactory substitute for the original."[131]

Responding to the absence of probative evidence in the record, the court rejected plaintiffs' two principal theories of market harm: that snippet views would suppress book sales[132] and that Google's unlicensed search-and-snippet-view service undermined existing unpaid licensing markets.[133] The court acknowledged that some books sales might suffer because a snippet would satisfy a focused factual inquiry that might otherwise lead to a sale,[134] but because the author's copyright does not extend to factual information, this substitution effect would not be related to the value of the author's copyrighted expression.[135] Google does not provide snippets for reference works for which snippets may have a more pronounced substitution effect.[136] In relation to the social value provided by Google's transformative search service, these minor substitution effects were insufficient to tip the fourth factor in plaintiffs' favor.[137]

Turning to the licensing claim, the court held that even if plaintiffs offered paid licenses to use digitized works to provide a search service, copyright law does not grant the copyright owner an exclusive right to provide information about a work of authorship, so there would be no need for such a license.[138] To the extent that there are other licensed uses to provide views or portions of digitized books, such as Amazon's Search Inside the Book service, the court found these to be distinguishable because they provide significantly more expressive content than snippets do.[139] Nor is analogizing the potential to license snippets to a licensing market for ringtones persuasive because snippets vary based upon the search terms; whereas, a ringtone provides a mini-performance of the most economically significant portion of a piece of recorded music.[140]

In both *HathiTrust* and *Google Books*, the copyright owners argued that the court should also treat as part of its market-harm analysis the

131 *Id.*
132 *See id.* at 224.
133 *See id.* at 226-27.
134 *See id.* at 224.
135 *See id.*
136 *See id.* at 210.
137 *See id.* at 224-25.
138 *See id.* at 226.
139 *See id.*
140 *See id.* at 226-27.

risk that the use could lead to large-scale infringement.[141] Judge Leval found this argument to be "theoretically sound,"[142] but in both cases the evidence was that the full-text copies used for search by both the HathiTrust Digital Library and by Google were subject to reasonable security measures with no accompanying evidence that breach of these systems was impending or even likely.[143]

More recently, a district court relied on these cases to hold that systematic copying of television newscasts for the purpose of providing a commercial "clipping service" for journalists, politicians, and others with an interest in the news of the day was fair use as to the copying to provide a search service.[144] That portion of the decision was not appealed, but the Second Circuit held in *Fox News Network, LLC v. TVEyes, Inc.*[145] that the service's distribution and performance of ten-minute video clips to its clients exceeded the bounds of fair use.[146]

TVEyes records all available television news programs from approximately 1,400 channels along with the text transcripts of these programs provided by closed captioning or speech-to-text software. Users, such as politicians, journalists, marketing executives, and others with an interest in monitoring how news on particular topics is portrayed, pay $500 per month to subscribe to TVEyes' service.[147]

A subscriber's search of the text version of the database yields results showing thumbnail images of, and identifying information about, responsive clips. If the subscriber clicks on the link to such a clip, a video begins playing fourteen seconds before the search term is mentioned and the subscriber can continue watching for up to ten minutes. Subscribers can download these clips and can email them to others without technical restriction. Subscribers also can archive clips

[141] *See id.* at 227; Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 100-01 (2d Cir. 2014).

[142] *Google Books*, 804 F.3d at 227.

[143] *See id.* at 227-28 (finding that "Google has made a sufficient showing of protection of its digitized copies of Plaintiffs' works to carry its burden on this aspect of its claim of fair use . . ."); *HathiTrust*, 755 F.3d at 100-01 (finding "no basis . . . on which to conclude that a security breach is likely to occur, much less one that would result in the public release of the specific copyrighted works belonging to any of the plaintiffs in this case").

[144] Fox News Network, LLC v. TVEyes, Inc., 43 F. Supp. 3d 379, 392-93 (S.D.N.Y. 2014).

[145] 883 F.3d 169 (2d Cir. 2018).

[146] *See id.* at 174.

[147] *See id.* at 175.

for later retrieval; otherwise TVEyes records over prior programming after thirty-two days.[148]

The service is sold only to businesses, who enter into an agreement that restricts their use of clips for "internal purposes only." Subscribers are not limited in the number of searches that they can run, but during the course of the litigation, TVEyes implemented a technology to prevent subscribers from stitching an entire newscast together from ten-minute segments.[149]

The court's fair use analysis divided TVEyes' service into two uses — those necessary to support the search function and those related to the "watch" function. Following the reasoning of the search cases discussed above, the district court held that the copying and archiving necessary to provide a search service was a fair use.[150] Fox conceded this point by choosing not to appeal this portion of the judgment.[151]

With respect to the provision of clips for viewing, the court held that this use was transformative because it "enables nearly instant access to a subset of material — and to information about the material — that would otherwise be irretrievable, or else retrievable only through prohibitively inconvenient or inefficient means."[152] The reasoning in this case is not fully in concert with circuit precedent and that of the other cases concerning computational analysis and should therefore be read as an outlier. For example, the court characterized providing the watch function only as a means of enhancing efficiency by saving clients the time and effort of constantly monitoring news programming themselves without recognizing that a further purpose of the function was to enable clients to compare whether politicians and other newsmakers were taking inconsistent positions on issues over time or before different audiences.[153] The court also recharacterized the

[148] *See id.*

[149] *Id.*

[150] Fox News Network, LLC v. TVEyes, Inc., 43 F. Supp. 3d 379, 392-93 (S.D.N.Y. 2014).

[151] *See* 883 F.3d at 176.

[152] *Id.* at 177.

[153] *See id.* at 177-78. There is some irony in this holding. The court's opinion in *TVEyes* was written by Judge Jacobs, who had dissented in the *Texaco* case discussed *supra* notes 93–107. The majority opinion in that case was written by Judge Newman, who also was on the *TVEyes*' panel. *See* 883 F.3d at 172. Twenty-four years previously, Judge Jacobs had argued in dissent in *Texaco* that the scientist's photocopying was transformative because it contributed to research. *See* Am. Geophysical Union v. Texaco, Inc., 60 F.3d 913, 932 (2d Cir. 1994) (Jacobs, J., dissenting). Having lost that argument then, Judge Jacobs somewhat overread *Texaco* in *TVEyes* when writing for the

Supreme Court's fair use analysis in *Sony Computer Entertainment, Inc. v. Connectix Corp.*[154] as having determined that viewers who taped television shows were doing so for the transformative purpose of improving the efficiency of delivering content, which unduly minimizes the impact of a technology that created a new audience for the television programming.[155]

* * * * *

On balance, the relevant judicial principles that a court would apply to a TDM researcher's reproducibility copies would favor the use. Recent cases have focused attention on the first fair use factor, the purpose and character of the use, and have been particularly solicitous when the use is transformative because it is for a new purpose. Although the court in *Texaco* took a narrow view of transformativeness, the internal limits of that decision combined with subsequent clarifications of the transformativeness inquiry provide broader room for TDM research, whether done by academic or commercial researchers.

B. The Copies that Count

Some forms of TDM research either currently do or, in the future, may rely only on making temporary copies of copyrighted works during the computational research step and then keeping durable outputs of that analysis that do not contain any expression that is substantially similar to the works that were analyzed. For this mode of TDM research, copyright law would not apply at all unless making the temporary copies in a computer's memory exercises the copyright owner's exclusive right of reproduction.[156] This Section addresses whether the copies made into a computer's active memory during the course of processing "count" for copyright purposes.

Not all copies count for purposes of copyright infringement. The Copyright Act confers upon the copyright owner the exclusive right to "reproduce the copyrighted work in *copies* or phonorecords."[157] Only reproductions that are in "copies" count. The Act defines "copies" as "material objects, other than phonorecords, in which a work is *fixed* by any method now known or later developed, and from which the work

court that the fact that a use facilitates research does not by itself make the use transformative. *See TVEyes, Inc.*, 883 F.3d at 178 n.4.

154 203 F.3d 596 (9th Cir. 2000).

155 *See* 883 F.3d at 177.

156 *See* 17 U.S.C. § 106(1) (2019).

157 *Id.* (emphasis added).

can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device."[158] Therefore, only reproductions that are "fixed" count. A copyrighted work is "'fixed' in a tangible medium of expression when its embodiment in a copy or phonorecord, by or under the authority of the author, is sufficiently permanent or stable to permit it to be perceived, reproduced, or otherwise communicated *for a period of more than transitory duration*."[159]

When these statutory dots are connected, the limit on the copies that count is that the copyrighted work must be reproduced in statutory "copies," which require (1) an embodiment — the work is permanent or stable enough to be perceived, reproduced or communicated; and (2) sufficient duration — this embodiment must last "for a period of more than transitory duration."[160]

Most computational analysis of scholarly articles requires a copy in active memory that lasts for less than one second.[161] From the statutory text, one would reasonably conclude that these evanescent computational copies endure for too short a period to count for purposes of Section 106(1). However, some publishers assert that a researcher requires a license to mine their publications.[162]

This legal position is based on a combination of a misunderstanding about the facts of text and data mining and a misreading of the case law. While the early cases interpreting the fixation requirement have been discussed by commentators in the past,[163] a brief review is in order to address more recent characterizations of these decisions.

1. Temporary Copies v.1.0

During the analog era, judicial construction of the fixation requirement was rare because analog technologies usually do not generate evanescent copies. As competition in the market for video games grew in the late 1970s and early 1980s, the fixation requirement became more salient.[164] The computer program files that instructed

---

158 17 U.S.C. § 101 (emphasis added).

159 *Id.* (emphasis added).

160 *See, e.g.*, Cartoon Network LP v. CSC Holdings, Inc., 536 F.3d 121, 127 (2d Cir. 2008).

161 *See* MINER ET AL., *supra* note 19, at 32.

162 *See, e.g.*, *supra* note 10 (providing example of publisher license).

163 *See, e.g.*, Joseph P. Liu, *Owning Digital Copies: Copyright Law and the Incidents of Copy Ownership*, 42 WM. & MARY L. REV. 1245 (2001); Aaron K. Perzanowski, *Fixing RAM Copies*, 104 NW. U. L. REV. 1067 (2010).

164 *See* Perzanowski, *supra* note 163, at 1093 n.135.

video game machines to display audiovisual works during game play were stored in read-only memory ("ROM") on a disc in the machine.[165] During game play, the program was loaded into active memory, from which it rendered an audiovisual work that combined stored graphic images for the background and characters with dynamic responses to player inputs.[166]

a. The Video Game Cases

When competitors created games that presented a similar audiovisual experience generated by computer programs that were not similar to the original game's program, copyright owners in the original game sued for infringement based on the audiovisual elements of the game as a distinct work of authorship.[167] One common defense was that what happened on the screen was not a copy that counted because the dynamic presentation of the audiovisual game components failed to meet the requirement either that the original work was fixed in order to acquire copyright protection or that the allegedly infringing audiovisual work was not fixed as a "copy."[168]

In *Midway MFG. Co. v. Artic International, Inc.*,[169] defendant Arctic sold circuit boards that could be attached to plaintiff's Galaxian and Pac-Man arcade game units. One circuit board sped up the game play for Galaxian; the other produced a "Puckman" game (the original title of Pac-Man as published in Japan).[170] Midway's copyright claims were limited to the audiovisual works on the screen. Artic argued that these works were not fixed because they appeared only briefly on the screen in either the game's "attract" or "play" modes, and therefore Artic was not making copies that count. The court rejected this argument, holding that the "fixation requirement . . . does not require that the work be written down or recorded somewhere exactly as it is perceived by the

165 *See* Midway Mfg. Co. v. Artic Int'l, Inc., 547 F. Supp. 999, 1007-08 (N.D. Ill. 1982), *aff'd*, 704 F.2d 1009 (7th Cir. 1983) (discussing audiovisual work in connection with ROM).

166 *See, e.g.*, Williams Elecs., Inc. v. Artic Int'l, Inc., 685 F.2d 870, 874 (3d Cir. 1982) (rejecting the contention that "there is a lack of 'fixation' because the video game generates or creates 'new' images each time the attract mode or play mode is displayed, notwithstanding the fact that the new images are identical or substantially identical to the earlier ones").

167 *See, e.g.*, Stern Elecs., Inc. v. Kaufman, 669 F.2d 852, 853 (2d Cir. 1982).

168 *See, e.g.*, *Williams Elecs., Inc.*, 685 F.2d at 873-74.

169 547 F. Supp. 999 (N.D. Ill. 1982), *aff'd*, 704 F.2d 1009 (7th Cir. 1983).

170 *Id.* at 1004-05.

human eye."[171] Rather, all that was required was that the work "is capable of being 'reproduced … with the aid of a machine or device."[172] Thus, although "new" sights and sounds were technically created every time the game was manipulated by a user, the audiovisual portions of the videogame were nevertheless copyrightable because they were fixed as a computer program, which permanently resided in ROM embedded in the game console.

In *Stern Electronics, Inc. v. Kaufman*,[173] the Second Circuit similarly held that the audiovisual components of the Scramble video game were copyrightable because the program hosting the game was "permanently embodied in a material object, the memory devices."[174] Because all the information required to display any component of the game permanently existed in a hard drive embedded in the unit, each sound and image was permanently capable of perception with the aid of a machine, and the audio-visual portions of the game were thus fixed.[175] Importantly, the court found that "fixation" did not arise because of the display, but rather because the components facilitating the display permanently resided on the game's hard drive, as a program.[176]

Finally, the Third Circuit in *Williams Electronics, Inc. v. Artic International, Inc.*,[177] held that no "new" images or sounds were created, because the information used to create those images and sounds permanently resided on the same machine as was used to create those images and sounds.[178] Agreeing with the Second Circuit in *Stern Electronics*, the court found that the computer program embedded in the ROM was protected under copyright because that was how the audio-visual portions of the game were fixed.[179]

Three trends thus emerge from these video game cases. First, even though displays of the audiovisual components are arguably "transitory," if the source of the display's components are copied into storage, the durable component copies comprise copies of the audiovisual work that count. Second, the computer programs containing data facilitating audiovisual displays were considered "fixed"

171 *Id.* at 1007.
172 *Id.* at 1007-08.
173 669 F.2d 852 (2d Cir. 1982).
174 *Id.* at 856.
175 *Id.* at 856-57.
176 *Id.* at 856.
177 685 F.2d 870 (3d Cir. 1982).
178 *Id.* at 874.
179 *See id.*

because they *permanently resided* on the console's ROM. Last, the courts accepted permanent embodiment in ROM *as a method of fixation*.

b. RAM Copies - MAI *Revisited*

Reliance on the fixation of the ROM copy to find fixation of the original work did not address issues that began to arise in the early 1990s in the corpus of what was then called "computer law." When the courts had to confront whether a copy of a computer program was fixed while it was being run, they appeared to ignore the durational limit and treat all digital copies as actionable reproductions.

In *MAI Systems Corp. v. Peak Computer, Inc.*,[180] the Ninth Circuit held that turning on a computer was an act of copyright infringement. The copyright owner, MAI, had granted its client a license to use its software but the license did not extend to third parties acting on the client's behalf. This license limitation was an intentional effort to be the exclusive maintenance provider for the client.[181] The technicians working for Peak had been former employees of MAI. In its copyright claim, MAI asserted that when the Peak technicians ran the operating system, they necessarily loaded a copy of the program that counted under § 106(1) into the computer's random-access memory ("RAM") to perform a diagnostics test.[182]

The principal issue was whether the copy of MAI's software was fixed when run by the Peak technician. The Ninth Circuit said that it was. It stated its holding twice in the opinion,[183] and one of those statements was understood to mean that nearly all temporary digital copies were fixed and therefore counted under the Copyright Act.[184]

The fixation issue arose on interlocutory review of a partial summary judgment and a permanent injunction in favor of the plaintiff.[185] In its summation of its copyright infringement analysis, the *MAI* court

[180] 991 F.2d 511 (9th Cir. 1993).

[181] *See id.* at 517.

[182] *See id.* at 518.

[183] *See id.* at 518-19.

[184] *See, e.g.*, Religious Tech. Ctr. v. Netcom On-Line Commc'n Servs., Inc., 907 F. Supp. 1361, 1368 (N.D. Cal. 1995) ("*MAI* [*Systems*] established that the loading of data from a storage device into RAM constitutes copying because that data stays in RAM long enough for it to be perceived."). *But see* Cartoon Network LP v. CSC Holdings, Inc. (*Cablevision*), 536 F.3d 121, 128 (2d Cir. 2008) (construing "*MAI Systems* and its progeny as holding that loading a program into a computer's RAM *can* result in copying that program" while declining to "read *MAI Systems* as holding that, as a matter of law, loading a program into a form of RAM *always* results in copying").

[185] *See MAI Systems*, 991 F.2d at 516-19.

appeared to have embraced a broad view of copyright liability by stating that "since we find that the copy created in the RAM can be 'perceived, reproduced, or otherwise communicated,' we hold that the loading of software into the RAM creates a copy under the Copyright Act."[186] Since nearly every digital copy can be transmitted or copied within microseconds, this view of the law would have treated all transient copies made by digital technologies as copies that count under the Copyright Act.

However, some courts and commentators have mischaracterized the holding in *MAI* by overlooking the procedural posture of the case.[187] The statement of the holding incorporates its analysis of the relevant facts that demonstrate recognition and respect for the durational component of fixation. The defendant had argued in its motion that the RAM copy was not fixed, but, according to the court, the defendant pointed to no facts in the record to support this argument.[188]

According to the court:

> It is also uncontroverted that when the computer is turned on the operating system is loaded into the computer's RAM. As part of diagnosing a computer problem at the customer site, the Peak technician runs the computer's operating system software, allowing the technician to view the systems error log, which is part of the operating system, thereby enabling the technician to diagnose the problem.[189]

From this uncontroverted fact and *the court's independent review of the record*, the court drew the legal conclusion that "by showing that Peak loads the software into the RAM and is then able to view the system error log and diagnose the problem with the computer, MAI has adequately shown that the representation created in the RAM is 'sufficiently permanent or stable to permit it to be perceived, reproduced, or otherwise communicated *for a period of more than transitory duration*.'"[190] Notwithstanding the above analysis, a range of courts cited *MAI* for the proposition that all temporary copies loaded into RAM are fixed without regard to duration.[191]

---

[186] *Id.* at 519.

[187] *See* Perzanowski, *supra* note 163, at 1073-75 (discussing courts and commentators who have misread *MAI Systems*).

[188] *MAI Systems*, 991 F.2d at 518.

[189] *Id.*

[190] *Id.* (emphasis added).

[191] *See, e.g.*, Stenograph L.L.C. v. Bossard Assocs., Inc., 144 F.3d 96, 101-02 (D.C. Cir. 1998) (stating that "RAM reproduction constitutes a 'copy'"); Iconix, Inc. v.

MAI sparked a heated response and debate in the scholarly literature and in public policymaking.[192] This debate was less about the durational element and more about the premise that *any* RAM copies should count for copyright purposes. On one hand, and most controversially, the Clinton Administration's roadmap for the digital environment treated the RAM copy doctrine as if it were settled law.[193] Some scholars, who recognized the RAM copy doctrine to change the balance between copyright owners and users, supported it on policy grounds in light of other changes made by digital technologies that empower users.[194]

In strong opposition, scholars such as Jessica Litman,[195] Peter Jaszi,[196] James Boyle,[197] Pamela Samuelson,[198] Niva Elkin-Koren,[199] and Joseph

Tokuda, 457 F. Supp. 2d 969, 994 (N.D. Cal. 2006) (interpreting *MAI* to mean loading software into RAM infringe); Playboy Enters., Inc. v. Webbworld, Inc., 991 F. Supp. 543, 551 (N.D. Tex. 1997) (same); Sega Enters. Ltd. v. MAPHIA, 948 F. Supp. 923, 931-32 (N.D. Cal. 1996) (same); Religious Tech. Ctr. v. Netcom On-Line Commc'n Servs., Inc., 907 F. Supp. 1361, 1368 (N.D. Cal. 1995) (stating that under *MAI* "the loading of data from a storage device into RAM constitutes copying because that data stays in RAM long enough for it to be perceived"); CSU Holdings, Inc. v. Xerox, 910 F. Supp. 1537, 1541 (D. Kan. 1995) (interpreting *MAI* to mean loading software into RAM infringe); Tricom, Inc. v. Elec. Data Sys. Corp., 902 F. Supp. 741, 745 (E.D. Mich. 1995) (same).

192 *See* Perzanowski, *supra* note 163, at 1073-80.

193 *See* BRUCE A. LEMAN, ASSISTANT SEC'Y OF COMMERCE AND COMM'R OF PATENTS AND TRADEMARKS, INFORMATION INFRASTRUCTURE TASK FORCE, INTELLECTUAL PROPERTY AND THE NATIONAL INFORMATION INFRASTRUCTURE: THE REPORT OF THE WORKING GROUP ON INTELLECTUAL PROPERTY RIGHTS 92 (1995).

194 *See, e.g.*, I. Trotter Hardy, *Computer RAM "Copies": A Hit or Myth? Historical Perspectives on Caching as a Microcosm of Current Copyright Concerns*, 22 U. DAYTON L. REV. 425, 457-58 (1997) (treating sympathetically the case for making copyright a law of access and use).

195 *See* Jessica Litman, *Fetishizing Copies*, *in* COPYRIGHT LAW IN AN AGE OF LIMITATIONS AND EXCEPTIONS 107, 118-19 (Ruth L. Okediji ed., 2017); Jessica Litman, *The Exclusive Right to Read*, 13 CARDOZO ARTS & ENT. L.J. 29, 31-32 (1994).

196 *See, e.g.*, Peter Jaszi, *Taking the White Paper Seriously*, *in* COPYRIGHT AND THE NII: RESOURCES FOR THE LIBRARY AND EDUCATION COMMUNITY 97, 99 (Patricia Brennan ed., 1996) (contesting the alleged clarity of the RAM copy doctrine and stating that "[t]he net result of this interpretation . . . is to make any act of 'reading' a digital work acquired online or through a network, including the everyday act of browsing information accessible over the Internet, a potential infringement of proprietary rights").

197 *See* James Boyle, *Intellectual Property Policy Online: A Young Person's Guide*, 10 HARV. J.L. & TECH. 47, 83-94 (1996).

198 Pamela Samuelson, *The NII Intellectual Property Report*, 37 COMM. ACM 21, 22-23 (1994); Pamela Samuelson, *The Copyright Grab*, WIRED (Jan. 1, 1996, 12:00 PM), https://www.wired.com/1996/01/white-paper/ [https://perma.cc/YT3L-WJ63].

199 *See* Niva Elkin-Koren, *Copyright Law and Social Dialogue on the Information Superhighway: The Case Against Copyright Liability of Bulletin Board Operators*, 13 CARDOZO ARTS & ENT. L.J. 345, 354 n.47 (1995).

Liu,[200] all have advanced arguments that such an approach deprives users of their traditional right to read, view, or watch copies of copyrighted works to which they had access. In effect, the RAM copy doctrine created a new exclusive right to access the copyrighted work that Congress had never intended. This is a powerful critique. The analysis below shows that, ironically, users retain a right to read when done by their machines.

Anthony Reese argues that the RAM copy doctrine is inconsistent with the legislative history of the Copyright Act of 1976, and that the public display right rather than the reproduction right was intended to cover works that appear on a computer screen.[201] Other scholars have emphasized that even if RAM copies count, they are unlikely to be infringing because of implied license, fair use or other limits on liability.[202] Recent clarifications of the scope of fair use address some, but not all, of these scholars' concerns about most RAM copies as copies that count.

2. Temporary Copies v.2.0

If the period of time necessary for a technician to check a computer's error logs is more than transitory, how short a period is transitory? According to the Second Circuit, at least in one context, 1.2 seconds is only transitory.

In *Cartoon Network, LP v. CSC Holdings, Inc.* (*Cablevision*),[203] copyright owners in television programming sued a cable company for offering a remote digital video recorder ("DVR") service through which

---

200 *See* Joseph P. Liu, *Owning Digital Copies: Copyright Law and the Incidents of Copy Ownership*, 42 WM. & MARY L. REV. 1245, 1258-60 (2001).

201 *See* R. Anthony Reese, *The Public Display Right: The Copyright Act's Neglected Solution to the Controversy Over RAM "Copies,"* 2001 U. ILL. L. REV. 83, 140 ("It seems absolutely clear, however, from the structure and legislative history of the 1976 Act, that the drafters did not consider projecting an image onto a screen to be reproducing the work in a copy but rather they considered it a display of the work.") (citation omitted); *see also* Jonathan Band & Jeny Marcinko, *A New Perspective on Temporary Copies: The Fourth Circuit's Opinion in* Costar v. Loopnet, 2005 STAN. TECH. L. REV. P1, P5 (2005); Perzanowski, *supra* note 163, at 1076 n.46 (quoting legislative testimony by then Register of Copyrights supporting a right to read digitally).

202 *See* Mark A. Lemley, *Dealing with Overlapping Copyrights on the Internet*, 22 U. DAYTON L. REV. 547, 566-67 (1997) (arguing that RAM copies are likely impliedly licensed); Jule L. Sigall, Comment, *Copyright Infringement Was Never This Easy: RAM Copies and Their Impact on the Scope of Copyright Protection for Computer Programs*, 45 CATH. U. L. REV. 181, 217-19 (1995) (arguing that making a RAM copy is fair use).

203 Cartoon Network LP v. CSC Holdings, Inc. (*Cablevision*), 536 F.3d 121 (2d Cir. 2008).

customers could record and play back television programs on dedicated servers on Cablevision's premises.[204] The parties' respective litigation strategies put fixation at issue in an unusual case.

The plaintiffs and defendant stipulated not to raise certain issues in order to focus the case on whether Cablevision was directly liable for its actions. The plaintiffs agreed to forgo relitigating indirect liability for the recordings made by Cablevision's customers[205] in exchange for Cablevision forgoing its fair use argument.[206] The litigation focused on how Cablevision's technology copied the entirety of the programming it received into a buffer as it was being transmitted. As the programming data streamed in, it remained in memory for 1.2 seconds until it was overwritten by the next segment of data.[207] During this interval, Cablevision's computers would process which customers had recorded particular shows and would transmit copies of the relevant data to the designated storage units for those customers. At issue, then, was whether Cablevision was directly infringing the plaintiffs' right to reproduce the copyrighted works in copies and whether Cablevision was publicly performing these works when its computers streamed the data to the customers' homes when the customers pushed the play button on their remote control.[208]

By placing Cablevision's buffer copies squarely at issue, the case directly addressed when a copy counts in the digital context. The court reaffirmed that the statutory language discussed above "directs us to ask not only 1) whether a work is 'embodied' in that medium, but also 2) whether it is embodied in the medium 'for a period of more than transitory duration.'"[209]

The buffer copies met the embodiment requirement because the entire copyrighted work passed through the buffer even though only small portions were embodied at any given time.[210] With respect to the duration requirement, the court stated that "[w]hile our inquiry is necessarily fact-specific, and other factors not present here may alter the

---

[204] *See id.* at 124.

[205] The Supreme Court held that the manufacturer of a home video recorder is not indirectly liable for any infringing copies made by its customers because most copies are made for the non-infringing purpose of time-shifting the programming. *See* Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 417-18 (1984). Plaintiffs chose not to relitigate this issue with respect to Cablevision's remote DVR.

[206] *Cablevision*, 536 F.3d at 124.

[207] *Id.* at 124-25.

[208] *See id.* at 125-26.

[209] *Id.* at 129.

[210] *See id.* ("Cablevision does not seriously dispute that copyrighted works are 'embodied' in the buffer.").

duration analysis significantly, these facts strongly suggest that the works in this case are embodied in the buffer for only a 'transitory' period, thus failing the duration requirement."[211] *Cablevision*'s two-part analysis for determining whether a copy has been fixed has been reaffirmed or followed by other courts.[212]

Two other cases bear on the interpretation of copyright's duration element for the copies that count, as further discussed *infra* in Part III. *Cablevision*'s context-specific approach to the "transitory duration" element resonates with the Fourth Circuit's *dicta* in *CoStar Group, Inc. v. Loopnet, Inc.*[213] The court held that a web hosting service was not directly liable for infringing photographs stored on its site because it did not act volitionally.[214] To reinforce its reasoning that the defendant was providing an automated process, the court also questioned whether temporary copies made by the defendant's machines during an Internet transmission even met the durational element of fixation.[215] In a passage cited with approval by the Second Circuit,[216] the Fourth Circuit elaborated: "'Transitory duration' is thus both a qualitative and quantitative characterization. It is quantitative insofar as it describes the

[211] *Id.* at 130.

[212] *See, e.g.*, Capitol Records, LLC v. ReDigi Inc., 910 F.3d 649, 657-58 (2d Cir. 2018) (holding that service that enabled resale of digital music files created infringing stored copies even if buffer copies made in process did not last for more than a transitory duration); Soc'y of the Holy Transfiguration Monastery, Inc. v. Archbishop Gregory of Denver, Colo., 689 F.3d 29, 55 (1st Cir. 2012) (holding that copies stored and displayed continuously on website were fixed); IMAPizza LLC v. At Pizza Ltd., 334 F. Supp. 3d 95, 120 (D.D.C. 2018) (holding that file of photograph transmitted across the internet not sufficiently fixed until stored on a computer outside the United States); Grady v. Iacullo, No. 13-CV-00624-RM-KMT, 2017 WL 1176415, at *4 (D. Colo. Mar. 29, 2017) (holding that browsing thumbnail images on the internet creates temporary downloads that last for more than a transitory duration); Capitol Records, LLC v. Escape Media Grp., Inc., No. 12-CV-6646(AJN), 2015 WL 1402049, at *40 (S.D.N.Y. Mar. 25, 2015) (holding that plaintiffs had failed to submit evidence showing that buffer copies made during upload process or music streaming process lasted for more than a transitory duration); Live Face on Web, LLC v. Emerson Cleaners, Inc., 66 F. Supp. 3d 551, 555 (D.N.J. 2014) (holding that plaintiff had sufficiently alleged copying of sufficient duration of software downloaded onto a user's cache or computer memory).

[213] CoStar Group, Inc. v. Loopnet, Inc., 373 F.3d 544 (4th Cir. 2004).

[214] *See id.* at 555-56 ("[W]e hold that the automatic copying, storage, and transmission of copyrighted materials, when instigated by others, does not render an ISP strictly liable for copyright infringement under §§ 501 and 106 of the Copyright Act.").

[215] *See id.* at 551 ("While temporary electronic copies may be made in this transmission process, they would appear not to be 'fixed' in the sense that they are 'of more than transitory duration,' and the ISP therefore would not be a 'copier' to make it directly liable under the Copyright Act.").

[216] *See Cablevision*, 536 F.3d at 129.

period during which the function occurs, and it is qualitative in the sense that it describes the status of transition."[217]

Last is an early case involving text and data mining, *Ticketmaster Corp. v. Tickets.com, Inc.*[218] Tickets.com's computers used webcrawling software to temporarily copy Ticketmaster's web pages to extract uncopyrightable factual data about upcoming concerts and events. In light of the internet and computing speeds in the late 1990s, this process took 10-15 seconds.[219] Relying on an erroneously broad reading of *MAI*, the court held that "the copying is transitory and temporary and is not used directly in competition with [Ticketmaster], but it is copying and it would violate the Copyright Act if not justified."[220]

The court then correctly held that the copying was justified by the fair use doctrine as necessary intermediate copying to obtain uncopyrightable information.[221] The *Ticketmaster* court's holding with respect to copies made during the extraction phase of TDM is supported by the reasoning of a line of cases involving reverse engineering of software. The Ninth Circuit first recognized the right to copy software to reverse engineer it in *Sega Enterprises, Ltd. v. Accolade, Inc.*[222] and subsequently broadened this right to include systematic copying to facilitate the research process in *Sony*.[223] Other courts have followed the Ninth Circuit's reasoning.[224] The fair use doctrine is discussed in detail in Part III.A, *infra*, in relation to the compiling and archiving of datasets by TDM researchers. But, it is important to note here that even if the copies made during extraction "count" because they are in memory for a non-transitory period of time, it is a fair use to make those copies for the reasons given in that Part.

---

[217] *LoopNet, Inc.*, 373 F.3d at 551.

[218] No. 99CV7654, 2000 WL 1887522 (C.D. Cal. Aug. 10, 2000), *aff'd*, 2 F. App'x 741 (9th Cir. 2001).

[219] *Id.* at *2.

[220] *Id.* at *3.

[221] *See id.*

[222] 977 F.2d 1510 (9th Cir. 1993); *cf.* Atari Games Corp. v. Nintendo of Am. Inc., 975 F.2d 832 (Fed. Cir. 1992) (agreeing that intermediate copying is fair use but holding the defendant liable because of similarities in its final software product).

[223] Sony Comput. Entm't, Inc. v. Connectix Corp., 203 F.3d 596 (9th Cir. 2000).

[224] *See* DSC Commc'ns Corp. v. DGI Techs., Inc., 81 F.3d 597, 601 (5th Cir. 1996); Bateman v. Mnemonics, Inc., 79 F.3d 1532, 1539 n.18 (11th Cir. 1996); Mitel, Inc. v. Iqtel, Inc., 896 F. Supp. 1050, 1056-57 (D. Colo. 1995), *aff'd on other grounds*, 124 F.3d 1366 (10th Cir. 1997); *see also* Pamela Samuelson & Suzanne Scotchmer, *The Law and Economics of Reverse Engineering*, 111 YALE L.J. 1575, 1608-13 (2002) (noting that "*Sega v. Accolade* has been followed in virtually all subsequent cases").

In general, a researcher can take comfort in the Second Circuit's reinvigoration of a durational limit on the copies that count, because copies made for extraction during the mining process are usually transitory. However, the discussion in Part III.B *infra* analyzes the court's indication that measuring this durational limit on copyright is a fact-specific, context-dependent inquiry.

3. Congress Has Not Impliedly Amended the Fixation Requirement

A final issue that needs clarification is whether Congress impliedly amended the definition of fixation in Section 101 of the Copyright Act in 1998 in Section 512(a) of the Digital Millennium Copyright Act, or in 2018 in the Music Modernization Act. The courts have correctly interpreted the DMCA to be tailored to internet service providers without intending to broadly change fundamental provisions of the law, and they should do the same with the MMA.

Section 512(a) could be read to treat even buffer copies made by internet service providers as copies that count and to reverse the judicial gloss requiring volitional conduct for a person to directly exercise the copyright owner's exclusive rights. The reason is that Section 512 is a limit on a copyright owner's remedies for infringement, implying that the conduct within its safe harbor provision is infringing.[225] Such a reading would require a court to treat the provider of basic internet transmissions as infringing copyright when a user sends an infringing file through the service. An aggressive reading would hold that the provider is *directly* liable because its buffer copies count as infringing reproductions. Such a reading would contradict the Second Circuit's subsequent analysis of fixation in *Cablevision* and would overturn the volitional limit on direct liability first recognized in *Religious Technology Center v. Netcom On-Line Communication Services, Inc.* (*Netcom*).[226]

The courts have correctly rejected such a reading, recognizing that Section 512(a)'s limits apply to remedies for indirect infringement for providers of internet service. With respect to the continued vitality of the volitional conduct limit on direct liability, there is some judicial disagreement about how to characterize and to apply this limit, but no court of appeals has held that Section 512(a) overturned the volitional

[225] *See* 17 U.S.C. § 512(a) (2019) (located within Chapter 5 that covers remedies for infringement).

[226] 907 F. Supp. 1361 (N.D. Cal. 1995).

limit on liability for direct infringement first recognized in *Netcom*.[227] Therefore, it is the person who volitionally causes temporary copies to be made who is the relevant actor and not the provider of an automated service that actually many, if not all, such copies.

Similarly, no court of appeals has read Section 512(a) to amend the transitory duration limit on the copies that count. On the contrary, in *BMG Rights Management (US) LLC v. Cox Communications, Incorporated*,[228] the Fourth Circuit held that Cox's repeat infringer policy had not been reasonably implemented, and therefore Section 512(a) did not shield Cox from BMG's claims of *indirect* liability for contributory infringement.[229] Had BMG thought it could successfully hold Cox directly liable, it surely would have pressed such a claim. Therefore, even if the researcher rather than a provider of a cloud service is the relevant actor with respect to the copies made during TDM processing, only copies that are more than transitory count. And, if they count, they would likely be held to be fair use copies for the reasons set forth in Part III.A *infra*.

The courts should similarly understand that the Music Modernization Act of 2018's provision of a blanket license of the *reproduction right* for interactive music streaming services does not mean that Congress intended to overturn *Cablevision*'s reading of the transitory duration limit on the copies that count. While the MMA's definitional sections gesture at the buffer copies made by interactive streaming services as being within the license's covered activities — implying that unlicensed buffer copies might be infringing — a close reading of the MMA supports the continued vitality of *Cablevision*'s interpretation of fixation.

My forthcoming work dives deep into the meaning(s) of the MMA,[230] but for present purposes the ambiguity is in how the MMA defines the activities qualifying for the blanket license and how the MMA defines the scope of that license. The "covered activit[ies]" that qualify for the

[227] *See* BWP Media USA, Inc. v. Polyvore, Inc., 922 F.3d 42, 58-64 (2d Cir. 2019) (per curiam) (a panel opinion holding that summary judgment was inappropriate on volitional conduct with each member of the panel concurring separately in the result to provide distinct interpretations of the volitional conduct limit in copyright law); *see also id.* at 47-54 (Walker, J., concurring) (collecting cases affirming and applying the volitional conduct limit on direct liability).

[228] 881 F.3d 293 (4th Cir. 2018).

[229] *See id.* at 305.

[230] Michael W. Carroll, *Regulatory Copyright in the Music Industry* (work-in-progress) (on file with author).

license include "interactive stream[s]" that are defined, in part, as making "digital phonorecord deliver[ies]" ("DPDs").[231]

The MMA revised the definition of a DPD to be "each individual delivery of a phonorecord by digital transmission . . . that results in a *specifically identifiable reproduction* by or for any transmission recipient . . . , regardless of whether the digital transmission is also a public performance . . . , and includes a permanent download, a limited download, or an interactive stream."[232] The best reading of the "specifically identifiable" limit on covered copies is that it *ratifies* rather than reverses *Cablevision*'s holding on buffer copies, which are only portions of a work quickly overwritten by other portions such that the buffer never contains a "specifically identifiable" copy intended for a particular user.

This understanding of the MMA should not be altered by the fact that Congress chose to state the scope of the blanket license more broadly to "include[] the making and distribution of server, *intermediate*, archival, and *incidental* reproductions of musical works that are reasonable and necessary for the digital music provider to engage in covered activities."[233] The best reading of the terms "intermediate" and "incidental" in this context is to cover stable, non-transitory copies that may need to be made, for example, to transfer a music file from one server to another, or as steps in an archival process but *not* to include buffer copies that last for only a transitory duration.

III. TEXT AND DATA MINING IS LEGAL UNDER U.S. COPYRIGHT LAW

This Article asserts that text and data mining is legal in the United States because fair use permits copying and archiving data to enable and validate TDM research. The fair use justification for the copies made during TDM research includes the temporary copies made in computer memory during the "mining" phase to the extent necessary to justify this aspect of the research use. But, this Article also argues that in many cases these temporary copies do not even count for copyright purposes, and this is practically significant for certain types of cloud-based TDM research.

This Article also recognizes that if a user enters into a contract in which she promises not to exercise her fair use rights, such a contract is valid in the absence of copyright misuse, unconscionability or other

[231] 17 U.S.C. § 115(e)(7) (2019).

[232] *Id.* § 115(e)(10) (emphasis added).

[233] *Id.* § 115(d)(1)(B)(ii) (emphasis added).

limits on contractual enforcement.[234] However, because, in the United States, researchers and their libraries do not need to agree to publishers' TDM "licenses" to comply with copyright law, these agreements are merely contracts offered in exchange for access to their articles and not copyright licenses. Therefore, only contractual remedies are available for violations of these agreements in the United States, even if such agreements may be necessary as copyright licenses elsewhere.

A. Copying Journal Articles to Conduct and Validate TDM Research Is Fair Use

As scientists, TDM researchers need to copy journal articles and datasets as a necessary step in their research and to allow others to validate their results regardless of whether they conduct their research in academic or industrial settings. Whether they may do so tests modern copyright law's commitment to promote the progress of science.[235]

Contemporary scholarship questions this commitment. Most skeptical are Jerome Reichman and Ruth Okediji, who argue that copyright law's restrictions are adverse to the interests of practicing scientists.[236] While I agree with their critique at the international level, and as applied to the laws of the European Union, I think that U.S. copyright law is more science-friendly, which gives the United States a competitive advantage. As a result, this Article concludes that TDM researchers have a fair use right to keep reproducibility copies of the research articles and data used in their computational research. This right has been hard won.

As the discussion in Part II.A, *supra*, demonstrates, the law of fair use has fluctuated in its science-friendliness, but the recent trend is in favor of treating computational research as fair use. Matthew Sag was early to recognize this trend,[237] along with Ed Lee,[238] and Sag has advocated in

---

[234] *See infra* note 251 and accompanying text.

[235] *See* U.S. CONST. art. I, § 8, cl. 8 (providing Congress with power to grant copyrights to authors to "[p]romote the progress of science").

[236] *See generally* 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) (arguing that a range of international and national laws treat normal scientific practice as infringing).

[237] *See generally* Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 NW. U. L. REV. 1607 (2009) [hereinafter *Copyright and Copy-Reliant Technology*] (arguing that fair use protects non-expressive uses).

[238] *See* Edward Lee, *Technological Fair Use*, 83 S. CAL. L. REV. 797, 846 (2010) (arguing that courts have found fair use when the use involves "(1) verbatim copies of copyrighted works in their entirety at creation in order to create a database, (2) verbatim

favor of treating computational research in the humanities as fair use.[239] Writing in 2012, Reichman and Okediji expressed some skepticism about whether this trend generalized to the federal courts' willingness to treat computational research as fair use. I shared Sag's optimism at the time.[240] The Second Circuit's subsequent decisions in *HathiTrust* and *Google Books* put to rest any doubts about fair use generally favoring computational research. As James Grimmelmann noted, "copyright has concluded that reading by robots doesn't count."[241]

But along the way, there has been some conflation of terminology and analysis in my view. Much of this literature concludes that fair use permits "non-expressive" or "non-consumptive" uses of copyrighted works. These are uses by machines other than to communicate the expressive content of a copyrighted work to a human audience. In the TDM context, it appears that this scholarly analysis has focused on whether making temporary copies for the purposes of extracting non-copyrightable information from copyrighted works is fair use. I agree that fair use permits making such copies, but I further argue that fair use also is needed to justify assembling and keeping the dataset used in TDM research. In addition, as Part III.B. *infra* argues, many of the temporary copies made during the mining phase of the research do not even count for copyright purposes.

Fair use justifies making and keeping a database of copyrighted works necessary to enable computational analysis and to reproduce the results of such research. The "non-expressive" label is less helpful when applied to this use, but the case law provides strong support for this aspect of computational research to also be fair use. Whether it is an image search engine's compilation of a photographic database, a plagiarism detection service's compilation of a database of student term papers, a search engine's digitization and compilation of a database of published books, or a clipping service for television news' compilation of a database of local news programming, the courts have unanimously concluded that these are all fair uses.[242] The computational analysis

or more limited copies of relevant works during operation and use of the database, but (3) a more limited output of the works to the user or public").

[239] *See* Brief of Digital Humanities, *supra* note 40, 2-4.

[240] *See* U.S. LIBRARY OF CONG., ORPHAN WORKS AND MASS DIGITIZATION ROUNDTABLES 99-101, 106-07, 145-48 (Mar. 11, 2014) (comments of Michael W. Carroll), *available at* https://www.copyright.gov/orphan/transcript/0311LOC.pdf.

[241] James Grimmelmann, *Copyright for Literate Robots*, 101 IOWA L. REV. 657, 658 (2016).

[242] *See supra* Part II.A (discussing and citing these cases).

done by researchers in projects such as Big Mechanism described in Part I, *supra*, are closely analogous from a copyright perspective.

It is also worth noting that the fair use analysis that justifies TDM research on the data also is closely related to the issue of whether using such works as training data for machine learning systems or other forms of artificial intelligence is a fair use. Benjamin Sobel argues that using copyrighted works to train systems to create competing copyrighted works calls for a more refined definition of the markets that matter.[243] Engaging with this argument is beyond the scope of this Article other than to say that reshaping the relation between transformative uses and the market analysis in fair use risks creating serious collateral damage for other fair uses. Amanda Levendowski argues that fair use is needed to cover training data in order to unearth and correct implicit bias in machine learning and other artificial intelligence systems,[244] a conclusion that Sobel and the analysis in this Article also support.

As the discussion of *Google Books* and *HathiTrust*, *supra*, make clear, many copyright owners chafe at these results for two reasons. One is that many intellectual property rightsholders have a simple legal policy algorithm that they run against types of use that have not yet been subject to litigation: "if value, then right."[245] On this logic, since researchers are deriving value from TDM research that uses copyrighted works, copyright's scope should be interpreted to bring such uses within the rightsholder's exclusive rights to ensure that they can control and obtain economic benefits from such uses. This would be of particular interest with respect to TDM research carried out by industrial researchers.

The second reason is that the right to control such uses of their works would require that researchers ask rightsholders for permission to conduct such research. Such a right of control would provide copyright owners with leverage to demand certain conditions on how and where their works are used to monitor use and to limit the risk of downstream infringement that has become so much easier in the digital environment. For publishers, these "data" are the full corpus of their publications. They understandably have concerns about copies of this

[243] *See* Benjamin L. W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45, 46 (2017).

[244] *See* Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579, 629-30 (2018); *see also* Sobel, *supra* note 243, at 95-96.

[245] *See* Rochelle Cooper Dreyfuss, *Expressive Genericity: Trademarks as Language in the Pepsi Generation*, 65 NOTRE DAME L. REV. 397, 405 (1990) (coining the "if value, then right" formulation of this argument).

corpus being reproduced and stored on the machines and systems of numerous researchers.[246] Some of the larger publishers have created licensing arrangements designed to balance the competing demands of reproducibility and control over the content.[247]

Recognizing these interests, this Article nonetheless concludes that U.S. law appropriately addresses them while ensuring that the rights to conduct TDM research is a user's right and not the rightsholder's right. Courts have recognized that the balance struck between authors' and users' rights in copyright law provide authors with exclusive rights over some, but not all, uses of their works that generate value. As a result, the if-value-then-right formula overreaches.

The court in *Lewis Galoob Toys, Inc. v. Nintendo of Am., Inc.*[248] stated one of the more forceful judicial rejections of the formula. Nintendo argued that the scope of the exclusive right to prepare derivative works needed to reach the defendant's Game Genie technology — which allowed players of Mario Brothers and other popular games to alter the speed of the game, how many lives a player could have and other data values that impact the player's experience in the game. The technology merely altered these values while the game was being played but created no altered copies of the game. Nintendo's legal position was simply, "if value, then right" — "the existence of a $150 million market for the Game Genie indicates that its audiovisual display must be fixed."[249] The court rejected this reasoning: "Nintendo's argument also proves too much; the existence of a market does not, and cannot, determine conclusively whether a work is an infringing derivative work."[250]

In the United States, the fair use doctrine permits researchers to make and archive a full copy of their research data and to share this archive with other researchers who seek to reproduce their results. However, if

[246] *E.g.*, ELSEVIER, ELSEVIER PROVISIONS FOR TEXT AND DATA MINING (TDM) (2017), https://www.elsevier.com/__data/assets/pdf_file/0012/102234/TDM-sign-up-short-form.pdf [https://perma.cc/8KGF-2PQF] (providing access to content for TDM research under condition that "[y]ou must permanently delete all Elsevier content or Elsevier data which you stored pursuant to your use of the APIs except for the TDM output and the Snippets. Notwithstanding the foregoing, you are permitted to retain a private copy of the corpus, or excerpts thereof, for reasons of data archiving requirements and to make this corpus available for internal institutional uses or for peer review, funding or ethics purposes (but not for further external distribution by these agencies or reviewers).").

[247] *E.g.*, *Text and Data Mining*, *supra* note 10 (providing license for researcher to download articles for TDM research).

[248] 964 F.2d 965 (9th Cir. 1992).

[249] *Id.* at 968.

[250] *Id.* at 969.

a researcher or their librarian agrees to limit their use of the reference data, these agreements would be enforceable under contract law.[251]

This Part applies the general teachings about fair use and computational research and explains why the fair use doctrine in U.S. law protects a researcher's ability to: (1) copy any full-text journal article that the researcher can access; (2) make any non-transitory temporary copies necessary to computationally process these articles; (3) store a copy of the dataset of full-text journal articles used for research; and (4) share this dataset of articles with other researchers for research purposes so long as the dataset is not made generally available to the public over the internet or otherwise.

In light of the relevant fair use decisions described in Part II, the researchers engaged in text and data mining practices that would exercise another's rights under copyright are making fair uses of those works. For purposes of the analysis that follows, this Article assumes the following facts for the paradigm case: (1) the researcher has copied multiple journal articles, datasets, or other research outputs for the purpose of conducting computational research ("the TDM data"); (2) the researcher has reformatted these articles, likely from a PDF to an XML or other structured data format; (3) the results of the researcher's computational processing of the TDM data contain at most only small amounts of the copyrightable expression found in these inputs; (4) the researcher stores archives the TDM data for future reference, either to provide another researcher the means to reproduce the research or to conduct further research; (5) the TDM data are not publicly distributed and are kept under reasonably secure conditions to thwart copying or distribution without the researcher's knowledge.

This Article then discusses whether the analysis materially changes if the research is done by a researcher in a for-profit firm or if the source material for the reproducibility copies is comprised of infringing copies. Even with these additional considerations, each of the fair use factors are in the researcher's favor.

1. The Paradigmatic Use — Compiling Data for Research and Retaining It for Reproducibility

A court asked to rule on whether the uses in the base case are fair uses would likely encounter the following arguments and rule in favor of the researcher. The first factor focuses on the "purpose and character" of

[251] *See, e.g.*, Jacobsen v. Katzer, 535 F.3d 1373, 1381-83 (Fed. Cir. 2008) (enforcing software license that requires user to waive fair use rights).

the use.[252] This two-step inquiry asks whether the use is "transformative" and whether the use is commercial. In the first step, "[t]he central purpose of this investigation is to see . . . whether the new work merely 'supersede[s] the objects' of the original creation . . . or instead adds something new, with a further purpose or different character, altering the first with new expression, meaning, or message; it asks, in other words, whether and to what extent the new work is 'transformative.'"[253]

a. Copying to Enable Computational Research Is a Transformative Purpose

A researcher's purpose for making, processing, and keeping reproducibility copies is transformative because the copies are used to conduct TDM research and to validate the results, or to enable related computational research, and not to provide a substitute to the intended human readers of these articles. While the *Texaco* court held that a researcher's keeping copies of journal articles for future research use was not a transformative use,[254] *Texaco* is distinguishable on this point. Assuming that the *Texaco* court correctly decided that a researcher's keeping copies of journal articles for future reference *by the researcher* was a non-transformative "archival" use,[255] a researcher's archiving of articles as a *dataset* to validate computational research is a wholly new use. More recent decisions support this distinction.

Courts have repeatedly held that institutional and systematic copying of entire works for the purpose of serving as a database to provide a search service for these works is a transformative use.[256] Professor Matthew Sag recognized this trend, deeming these forms of copying "non-consumptive" or "non-expressive" uses because the copying is not to provide others with access to the expressive content of the works in the database.[257]

---

[252] 17 U.S.C. § 107 (2019).

[253] Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994).

[254] *See* Am. Geophysical Union v. Texaco, Inc., 60 F.3d 913, 924 (2d Cir. 1994).

[255] *See* Fox News Network, LLC v. TVEyes, Inc., 883 F.3d 169, 186 (2d Cir. 2018). As the discussion of subsequent fair use case law indicates, were the question to arise in a different circuit, a court may well agree with Judge Jacobs' dissent and hold that the researcher's purpose for copying the journal articles as an input into future research is a transformative use.

[256] *See supra* notes 109–151 and accompanying text.

[257] *See, e.g.*, Sag, *Copyright and Copy-Reliant Technology*, *supra* note 237, at 1625, 1629 (analyzing cases); Matthew Sag, *Predicting Fair Use*, 73 OHIO ST. L.J. 47, 56-57 (2012) (showing that a finding of transformativeness predicts a holding of fair use);

The intermediate copying cases also support TDM under the first factor. In those cases, courts held that copying software for the purposes of extracting public domain factual information about how operating system software functioned in order to produce interoperable software was a favored use under the first fair use factor.[258]

Publishers may argue that as TDM becomes a more well-understood and anticipated research practice, the purpose of journal publishing has shifted to serve both researchers themselves and their machines conducting computational processing. They may point to the existence of their proffered licensing terms for TDM as evidence that serving this need is part of their market under the fourth factor and therefore part of their purpose in publishing under the first factor.

This argument is unpersuasive. The courts have thus far rebuffed attempts to rely on proffered license terms to constrain the scope of transformative uses under the first factor. There's an important point that some courts have emphasized about transformativeness. It is a shorthand phrase for a more complex analysis about the non-substitutional contribution that the use makes. In other words, the use is productive because it "instead adds something new, with a further purpose or different character, altering the first with new expression, meaning, or message."[259]

Were one to take this standard as setting up a mere logic game, a copyright owner could argue that the publication of the original work had two aims: to serve its primary expressive purpose and to also enable derivative uses for which are subject to licensing. Therefore, the argument goes, a use is not transformative if the copyright owner is willing to offer a license for the use because it is not for a "further purpose."

But, this is not a game. The transformativeness inquiry aims to achieve copyright's fundamental balance between public and private interests by recognizing that users have legitimate reasons for making

Matthew Sag, *The Google Book Settlement and the Fair Use Counterfactual*, 55 N.Y.L. SCH. L. REV. 19, 54 (2010) (discussing, in the context of the Google Book settlement, "the creation of a 'Research Corpus' for non-consumptive and non-commercial research by certain qualified users").

258 *See* Sony Comput. Entm't, Inc. v. Connectix Corp., 203 F.3d 596, 601, 608 (9th Cir. 2000); Lewis Galoob Toys, Inc. v. Nintendo of Am. Inc., 964 F.2d 965, 970 (9th Cir. 1992); Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1514, 1518 (9th Cir. 1992); *see also* Pamela Samuelson, *Fair Use for Computer Programs and Other Copyrightable Works in Digital Form: The Implications of* Sony*,* Galoob*, and* Sega, 1 J. INTELL. PROP. L. 49, 86, 95-96 (1993) (identifying a range of uses that rely on intermediate copying that would be fair use).

259 Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994).

productive uses of others' works when they are making positive contributions that do not compete with the author's original expressive contribution.[260] A use is transformative because it "communicates something new and different from the original or expands its utility, thus serving copyright's overall objective of contributing to public knowledge."[261] The courts have rejected copyright owners' attempts to break the spirit of transformative use and bring it under a licensing harness, stating that "a copyright holder cannot prevent others from entering fair use markets merely 'by developing or licensing a market for parody, news reporting, educational or other transformative uses of its own creative work.'"[262]

For this reason, the fact that some journal publishers are willing to offer licenses for text and data mining of their publications does not deprive computational analysis of its transformative character. Scientific research articles, for example, aim primarily to express research findings and analysis to fellow researchers and other human readers who seek to understand what was tested, why, and to what effect. Computational analyses that extract non-copyrightable facts about correlations or patterns that can be found only through large-scale processing add new meaning and new utility to these publications. For this reason, Judge Leval wrote in *Google Books* regarding the nGrams text and data mining tool, "[w]e have no doubt that the purpose of this copying is the sort of transformative purpose described in *Campbell* as strongly favoring satisfaction of the first factor."[263]

When a researcher archives the corpus of publications that were mined for reproducibility purposes, she is also making a transformative use. The purposes of archiving for validation purposes or to enable further transformative text and data mining differ materially from the original expressive purposes of publishing these articles. This is a non-substitutional use that is akin to the transformative use of creating an

260 *See* Neil Weinstock Netanel, *Making Sense of Fair Use*, 15 LEWIS & CLARK L. REV. 715, 759-67 (2011) (reviewing cases and describing the central role of the transformativeness designation in fair use analysis).

261 Authors Guild, Inc. v. Google, Inc. (*Google Books*), 804 F.3d 202, 214 (2d Cir. 2015).

262 Bill Graham Archives v. Dorling Kindersley Ltd., 448 F.3d 605, 615 (2d Cir. 2006) (quoting Castle Rock Entm't, Inc. v. Carol Publ'g Grp., 150 F.3d 132, 145 n.11 (2d Cir. 1998)).

263 *Google Books*, 804 F.3d at 217.

archive that is necessary to support search services or other validation services.[264]

Texaco is not to the contrary. A researcher's making and keeping reproducibility copies differs from the researcher's filing photocopies of articles in her files for later reading, which was held not to be a transformative use in *Texaco* because the purpose of the photocopies was to be read in the same manner as the original publication.[265] Reproducibility copies are archived for different purposes than to be read directly by humans who seek to understand the authors' expressive purpose for writing the article.

Whether the *Texaco* majority's reasoning on this point would still carry the day is also subject to some doubt. Judge Jacobs wrote in dissent that the researcher's photocopies were transformative because they were inputs into ongoing research.[266] One might have argued differently that the researcher's photocopies were a form of time shifting analogous to the use approved as fair use in *Sony Corporation of America v. Universal City Studios, Inc*.[267] Just as a viewer could retrieve and watch a recorded program at a later time, a researcher could retrieve and read an article at a later time. In the absence of a comparison to *Sony*, Judge Newman dismissed a form of this argument that focused on format shifting at the time, characterizing the copying and filing as merely supplying a convenience but not providing a new purpose.[268] But, one could also characterize home taping as supplying a convenience.

If the analogy holds, then perhaps the law in the Second Circuit has shifted. For, while Judge Newman, writing for the court, disagreed with Judge Jacobs in *Texaco* in 1994, twenty-four years later (in 2018) he joined Judge Jacobs' opinion in *TVEyes*,[269] which reinterpreted the copying of television programs for later viewing in *Sony* as a transformative use.[270] If copying to time-shift television programming is a transformative use, then perhaps copying to time-shift journal articles is as well.

---

[264] *See* A.V. *ex rel.* Vanderhye v. iParadigms, LLC, 562 F.3d 630, 639 (4th Cir. 2009) (holding that copying student research papers to supply a plagiarism detection service was transformative).

[265] *See* Am. Geophysical Union v. Texaco, Inc., 60 F.3d 913, 918-20 (2d Cir. 1994).

[266] *See id.* at 935 (Jacobs, J., dissenting).

[267] *See* 464 U.S. 417, 456 (1984) (holding copying for the purpose of time-shifting the viewing of a television program to be a fair use).

[268] *See Texaco, Inc.*, 60 F.3d at 923-24.

[269] Fox News Network, LLC v. TVEyes, Inc., 883 F.3d 169, 172 (2d Cir. 2018).

[270] *See id.* at 177-78.

b. The Second and Third Factors Favor the Use

Text and data mining also fares well under the second fair use factor, the nature of the copyrighted work.[271] As the courts held in *Williams & Wilkins*[272] and in *Texaco*,[273] the focus of journal articles on reporting factual information tips the balance in the user's favor. This, however, is not saying much. Scholarly analysis of the role the second factor has played in fair use decisions demonstrates that it has become a statutorily required step that does little or no persuasive work in fair use analysis,[274] although it could.[275]

The third factor, the amount and substantiality of the work used, also favors wholesale copying for the purpose of computational analysis. This factor has to be analyzed in relation to the first and fourth factors.[276] The inquiry focuses on whether the amount used is appropriate for the purpose(s) analyzed under the first factor and if it is what the economic effect(s) of this use are on the copyright owner's traditional markets.[277] When the use is transformative under the first factor, and the amount used is appropriate, then it is likely that the use is fair because courts give little weight to the impact on lost licensing opportunities for transformative uses.[278]

With respect to the third factor, the analogy to the search cases is fairly direct. For the same reasons that wholesale institutional and systematic copying of entire works is necessary to extract the data required to provide a search service for books in print, images on the internet, or television news programming,[279] such copying also is necessary in order to computationally mine the content of the relevant parts of the scientific or scholarly literature. As in those cases, a court

[271] *See* 17 U.S.C. § 107 (2019).

[272] Williams & Wilkins Co. v. United States, 487 F.2d 1345, 1359 (Ct. Cl. 1973), *aff'd by an equally divided Court*, 420 U.S. 376 (1975).

[273] *Texaco, Inc.*, 60 F.3d at 925.

[274] *See, e.g.*, Barton Beebe, *An Empirical Study of U.S. Copyright Fair Use Opinions, 1978-2005*, 156 U. PA. L. REV. 549, 610-11 (2008) (finding that second fair use factor is not determinative).

[275] *See generally* Robert Kasunic, *Is That All There Is? Reflections on the Nature of the Second Fair Use Factor*, 31 COLUM. J.L. & ARTS 529 (2008) (arguing that the second factor should be a more important part of the balance in fair use analysis).

[276] *See* Authors Guild, Inc. v. Google, Inc. (*Google Books*), 804 F.3d 202, 221-23 (2015) (determining that Google's snippets were not longer than necessary to provide meaningful search results).

[277] *See id.* at 221.

[278] *See id.* at 224.

[279] *See* Fox News Network, LLC v. TVEyes, Inc., 43 F. Supp. 3d 379, 383 (S.D.N.Y. 2014), *aff'd in part and rev'd in part*, 883 F.3d 169 (2d Cir. 2018).

would likely find that this factor favors copying large numbers of journal articles that are relevant to text and data mining.

c. Copying to Conduct and Validate Research Does Not Affect the Markets that Matter

The fourth fair use factor, "the effect of the use upon the potential market for or value of the copyrighted work,"[280] focuses on the economic impact of the use. The principal focus of this inquiry is on the use's potential to substitute for the copyright owner's work or its derivative in the marketplace.[281] Courts also recognize that impacts on a copyright owner's market for licenses of the works can be part of the markets that matter.[282]

Not all impacts on licensing carry equal weight in fair use analysis. If they did, a circularity problem would arise because a copyright owner would simply have to offer, or be willing to offer, a license for the use to be deemed unfair. To avoid this result, courts have imposed important limits on when a use that impacts actual or potential licensing will tip the fourth factor in the copyright owner's favor.

First, "[o]nly an impact on potential licensing revenues for traditional, reasonable, or likely to be developed markets should be legally cognizable."[283] This limit applies most frequently where the use is not transformative. Even when a use involves copying enough expressive content that one might expect to need a license, if the copyright owner has no plausible economic reason to develop and support a licensing scheme to monetize such a use, then the fourth factor generally favors the user. This is one of the lessons of the evolution from *Williams & Wilkins* to *Texaco*.[284]

In *Williams & Wilkins*, the absence of a plausible licensing market for photocopying individual articles led the court to find that the fourth

[280] 17 U.S.C. § 107 (2019).

[281] *Google Books*, 804 F.3d at 223 ("The fourth fair use factor . . . focuses on whether the copy brings to the marketplace a competing substitute for the original, or its derivative, so as to deprive the rights holder of significant revenues because of the likelihood that potential purchasers may opt to acquire the copy in preference to the original.").

[282] *See id.*; Bill Graham Archives v. Dorling Kindersley Ltd., 448 F.3d 605, 614 (2d Cir. 2006).

[283] Am. Geophysical Union v. Texaco, Inc., 60 F.3d 913, 930 (2d Cir. 1994).

[284] *See* Rebecca Tushnet, *Copy This Essay: How Fair Use Doctrine Harms Free Speech and How Copying Serves It*, 114 YALE L.J. 535, 555-56 (2004).

factor favored the user.[285] Once journal publishers had developed the Copyright Clearance Center's photocopying license, the court accepted this as reasonable market in *Texaco* in part because Texaco was well able to budget for, and to afford, such a license.[286] It is critical to recall that the court limited its holding to copying for which such a license was available to avoid the problems posed by CCC's patchwork coverage highlighted by Judge Jacobs' dissent.[287] Moreover, a reasonable market requires that it be reasonable for both parties to engage in licensing. The *Texaco* court explicitly indicated that its fair use analysis might be different had the user been a university-based researcher engaged in non-commercial research.[288]

Second, where the use is transformative under the first factor, courts discount claims of harm to actual or potential licensing markets because the transformative use is not a competing use.[289]

Third, in *Google Books*, Judge Leval identified and explicated an important additional limit on the markets that matter. Even when the copyright owner has developed a licensing scheme, the license must be for the potentially substitutional use of the work's expressive content for an impact on that market to count under the fourth factor.[290]

"Licensing" often is used ambiguously. Formally, a provision is a copyright license only when it grants permission for an act that would otherwise infringe one or more exclusive right under copyright.[291] A copyright license may also be a contract, but it need not be one to be effective.[292] Under the fourth fair use factor, the effect on the copyright owner's licensing market focuses on the type of license that is necessary to avoid infringement.

---

[285] *See* Williams & Wilkins Co. v. United States, 487 F.2d 1345, 1356-57 (Ct. Cl. 1973), *aff'd by an equally divided Court*, 420 U.S. 376 (1975).

[286] *See Texaco, Inc.*, 60 F.3d at 930.

[287] *See supra* notes 104–107 and accompanying text.

[288] *See Texaco, Inc.*, 60 F.3d at 916.

[289] *See* Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 591-92 (1994) ("The market for potential derivative uses includes only those that creators of original works would in general develop or license others to develop."); Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 99 (2d Cir. 2014) ("In other words, under Factor Four, any economic 'harm' caused by transformative uses does not count because such uses, by definition, do not serve as substitutes for the original work."); Bill Graham Archives v. Dorling Kindersley Ltd., 448 F.3d 605, 614 (2d Cir. 2006).

[290] *See* Authors Guild, Inc. v. Google, Inc. (*Google Books*), 804 F.3d 202, 226 (2d Cir. 2015).

[291] *See* CRAIG JOYCE ET AL., COPYRIGHT LAW 323 (10th ed. 2016).

[292] *See, e.g.*, Philpot v. Media Research Ctr. Inc., 279 F. Supp. 3d 708, 713 (E.D. Va. 2018) (stating that a meeting of the minds is not required for a license to be sufficient).

Conversely, some terms of use associated with copyrighted works are merely contractual covenants. For example, copyright owners do not have an exclusive right to control access to their works, with one exception.[293] Therefore, one who sneaks into a movie theater to watch a public performance of a motion picture violates no exclusive right of the copyright owner but does violate the theater's contractual requirement that only ticketholders be allowed in the theater. To the extent that one construes the movie ticket as a license to view the motion picture, a use that impacts these types of terms or agreements carry much less weight under the fourth fair use factor.

Similarly, the copyright owner's exclusive rights do not include a right to provide factual information that is extracted from a copyrighted work or factual information about a copyrighted work, such as metadata or related information in search results. For this reason, Judge Leval reasoned, even if the plaintiffs had a plausible market for licensing copying to produce search results, the impact of the Google Books service on the market for providing information about copyrighted works would not tilt the fourth factor against this transformative use.[294] Rejecting the plaintiffs' reliance on cases involving retransmissions of music or provision of musical ringtones, Judge Leval wrote:

> In the cases cited, however, the purpose of the challenged secondary uses was not the dissemination of information *about* the original works, which falls outside the protection of the copyright, but was rather the re-transmission, or re-dissemination, of their expressive content. Those precedents do not support the proposition Plaintiffs assert — namely that the availability of licenses for providing unprotected information about a copyrighted work, or supplying unprotected services related to it, gives the copyright holder the right to exclude others from providing such information or services.[295]

When applying this reasoning to the text and data mining context, one must first recognize that even though some publishers offer text and data mining "licenses" that restrict how journal articles may be used in exchange for access to these articles, access is not a right protected by copyright. These terms are merely contractual and are not copyright licenses to the extent that they apply only to access. These access

[293] *See* 17 U.S.C. § 1201(a) (2019) (making unlawful the circumvention of a technological protection measure that effectively controls access to a work protected under Title 17 of the U.S. Code).

[294] *See Google Books*, 804 F.3d at 226.

[295] *Id.*

licenses are analogous to the unpaid search licenses — such as Amazon's Search Inside the Book service — relied upon by the *Google Books* plaintiffs. Since the computational phase of text and data mining does not exercise any exclusive right, and saving reference data is necessary for the transformative purpose of scientific validation, these access licenses are similarly regulating unprotected services.

Under the fourth fair use factor, the relevant issue is whether a researcher's making and keeping a dataset of journal articles for computational research, for future reference, and for providing access to this dataset for reproducibility purposes or to enable follow-on computational research would impact a publisher's market for text and data mining licenses, and if there is an impact, whether that weighs against fair use. On both counts it seems unlikely that a publisher would be able to persuade a court that the fourth factor weighs in its favor.

First, the impact of an individual researcher's conduct would likely be minimal because it would be unlikely for other researchers to turn to their colleague rather than to the publisher when undertaking a new line of computational research. Even when the aggregate effect of many researchers keeping datasets of journal articles is considered, the impact still would likely be minimal because these datasets would vary greatly in the portion of the publisher's catalog that are included in any one dataset. A researcher seeking to initiate a new line of computational research would still be more likely to seek copies of journal articles from the publisher to get exactly the articles needed rather than to canvas the collections of other researchers in the hopes of assembling an up-to-date dataset that meets the researcher's needs.

Second, even in cases in which one researcher avoids the publisher's text and data mining license by obtaining access to journal articles from another researcher, this impact would still have little effect on the analysis because the impact on licenses for unprotected services is not the focus of analysis under the fourth fair use factor. The publisher does not have an exclusive right to provide access, and, as the analysis above demonstrates, making transitory copies to extract unprotected information from these journal articles also does not exercise an exclusive right under copyright.

As a result, a court would be likely to conclude that a researcher who keeps copies of journal articles used for text and data mining and who provides these copies to other researchers who seek to reproduce or extend the original researcher's computational research is making a fair use of these articles so long as these copies are shared for these purposes.

d. Data Security Is Relevant and Favors This Use

Publishers challenging the fairness of keeping reproducibility copies would also likely argue that permitting such as use in the aggregate increases the risk of infringement by third parties who get access to these copies through hacking. The *HathiTrust* and *Google Books* courts recognized that even if a secondary use does not directly harm the economic value of the plaintiff's in an unreasonable way under the fourth factor, the court could still consider the risk that a secondary use could allow a third party to act in a way that would greatly diminish or destroy the economic value of the copyright in the work.[296] Judge Leval announced a rule of reason for this consideration. A use that is otherwise fair may become unfair if the secondary user does not take reasonable action to limit or mitigate the risk of third-party diminution of the copyright's value.[297]

This consideration does not amount to an independent, unenumerated factor. Instead, this consideration is part of the character of the use under the first factor. As with similar proportionality rules, the reasonableness of a secondary user's security precautions should be judged in relation to the probability of third-party harm and the foreseeable magnitude of its impact. It should take objective unreasonableness on the part of the secondary user for this consideration to tip the first factor against the use. Any imaginative advocate can conjure up hypothetical bad acts by third parties. Without a specific, credible threat, and objective unreasonableness, society should not be deprived of the benefits of uses that are otherwise fair.

Nonetheless, one would expect journal publishers to lean hard upon this consideration, arguing that individual researchers lack the ability to provide data security comparable to that of the libraries in *HathiTrust* or technology companies like Google. However, researchers usually operate within an institutional context and could rely upon libraries or cloud service providers to use the same kinds of security that were used in those cases to secure their reference copies of journal articles. Researchers storing such reference copies should take reasonable

[296] *See id.* at 228; Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 96-97 (2d Cir. 2014).

[297] *See Google Books*, 804 F.3d at 227 ("If, in the course of making an arguable fair use of a copyrighted work, a secondary user unreasonably exposed the rights holder to destruction of the value of the copyright resulting from the public's opportunity to employ the secondary use as a substitute for purchase of the original (even though this was not the intent of the secondary user), this might well furnish a substantial rebuttal to the secondary user's claim of fair use.").

precautions to impede hacking of such copies, and they have the means to do so.

Moreover, to the extent that risks to data security are a relevant consideration, this analysis should also take account of the larger environment. Researchers should not be held to a higher security standard than is reasonable in light of the fact that the publishers themselves have been unable to maintain the security of their journal data. Publishers' own security protocols have been breached or overcome, and infringing copies of a significant portion of the scientific literature is now readily available on the internet through Sci-Hub.

2. Copying from an Infringing Source Necessary for TDM Research Is Still a Fair Use

The final question this Article addresses is whether the above analysis concerning the computational phase of text and data mining or the maintenance of reference copies changes if a researcher takes advantage of the ready access to the scientific literature that Sci-Hub offers to computationally analyze all or subsets of this literature. In short, it would still be legal to perform text and data mining even if access is from an infringing source such as Sci-Hub.

a. Sci-Hub

Sci-Hub is a website that contains infringing copies of a substantial portion of the published scientific literature. Responding to restrictions imposed by both the costs and terms of access, Alexandra Elbakyan created Sci-Hub, which is an infringing collection of a substantial portion of the published literature. Data shows that many readers who use Sci-Hub have access to these publications through institutional subscriptions, but they find the convenience of Sci-Hub appealing.[298] For purposes of this Article, the relevant question is whether researchers in the United States may take advantage of the access Sci-Hub provides to conduct computational research. This Article answers in the affirmative.

[298] *See* Daniel S. Himmelstein et al., *Research: Sci-Hub Provides Access to Nearly All Scholarly Literature*, ELIFE SCI. (Feb. 9, 2018), https://elifesciences.org/articles/32822 [https://perma.cc/BYW5-SWQR] (noting "a large contingent of scientists supporting Sci-Hub's mission").

Alexandra Elbakyan founded Sci-Hub in 2011 when she was a graduate student and budding neuroscientist in Kazakhstan.[299] From her perspective, Sci-Hub is the world's first open-access research library. From the publishers' perspective, Sci-Hub is a case of large-scale copyright infringement.[300] It is commonly referred to as the Napster of academic publishing.[301]

Sci-Hub features a large collection of scholarly articles, book chapters, monographs, and conference proceedings.[302] The majority of articles on Sci-Hub come from journals published by Elsevier, Springer, the American Chemical Society, Sage Publications, and JSTOR, among others.[303] Commentators note that Sci-Hub's collection is likely more than 64.5 million articles.[304] This collection continues to grow. If a researcher requests a paper not currently on Sci-Hub, it obtains a copy through unclear means to add to the collection, allowing it to continually grow.[305]

Elbakyan has not been fully transparent as to how she has amassed such a large collection, but she claims that she uses legitimate online credentials, i.e., user IDs and passwords, to acquire access to the bulk of her articles.[306] Elbakyan also alleges that articles and online credentials are voluntarily donated by academics.[307] Sci-Hub collects data on what articles generate the most downloads and which countries and cities download the most articles.[308] For the most part, graduate students around the world,[309] particularly in China, India, Iran, Russia, and the United States, use Sci-Hub even if their university pays for academic subscriptions because Sci-Hub is easier to use and contains almost every article a researcher could want.

---

299 John Bohannon, *Who's Downloading Pirated Papers? Everyone*, SCIENCE (Apr. 28, 2016, 2:00 PM), http://www.sciencemag.org/news/2016/04/whos-downloading-pirated-papers-everyone [https://perma.cc/C2RP-KS2U].

300 *See id.*

301 *See id.*

302 *See id.*

303 *See id.*

304 *See* Rebecca Flowers, *Cloudflare Terminates Service to 'The Pirate Bay of Science'*, VICE (Feb. 9, 2018 7:30 AM), https://www.vice.com/en_us/article/59kgv5/cloudflare-terminates-service-to-the-pirate-bay-of-science [https://perma.cc/9CK3-EUGU].

305 *See* Bohannon, *supra* note 299.

306 *See id.*

307 *See id.*

308 *See id.*

309 *See id.* ("The download requests came from 3 million unique IP addresses, which provides a lower bound. But the true number is much higher because thousands of people on a university campus can share the same IP address.").

While Sci-Hub is a boon to researchers, it also is infringing copyright. Elbakyan defends the creation of Sci-Hub as an act of civil disobedience to respond to the problem that she and her peers commonly encountered — the high cost of academic and scientific articles from prestigious publishers.[310] While some journals are now open access[311] — meaning that their content is freely available upon publication under an open license that permits republication with attribution — the majority of the scientific literature is still published in subscription-based journals that rely on restricting access through so-called "paywalls" as the means of requiring a subscription or a purchase as the means of access.[312] Elbakyan argues that "[j]ournal paywalls are an example of something that works in the reverse direction, making communication less open and efficient."[313] Since journal articles are used for communication in science, Elbakyan contends that "the word 'communication' implies common ownership," which is why she created Sci-Hub.[314]

A number of publishers have sued in the United States to limit or eliminate access to Sci-Hub. In October 2015, Elsevier sued Elbakyan for copyright infringement and obtained a preliminary injunction.[315] In June 2017, the court issued a default judgment in the amount of $15 million. Although the sci-hub.org web domain was seized in November 2015, the servers that power Sci-Hub are based in Russia, beyond the influence of the U.S. legal system.[316] Barely skipping a beat, the site popped back up on a different domain. In addition to copyright

[310] *See* John Bohannon, *The Frustrated Science Student Behind Sci-Hub*, SCIENCE (Apr. 28, 2016, 2:00 PM), http://www.sciencemag.org/news/2016/04/frustrated-science-student-behind-sci-hub?IntCmp=scihub-1-11 [https://perma.cc/D4HF-RKSN].

[311] Disclosure: I am a member of the Board of Directors of the Public Library of Science (PLOS), which is the one of the first open access journal publishers. All of PLOS's journal content is available for download in machine-readable form to enable TDM research at https://www.plos.org/text-and-data-mining.

[312] *See, e.g.*, *Paywalls: Are They Effective?*, IO TECHNOLOGIES BLOG, https://iotechnologies.com/blog/monetization-paywalls? (last visited Sep. 8, 2019) [https://perma.cc/NJZ6-992R] (describing different types of paywall restrictions used by publishers); *see also* Jason Schmitt, *Paywall: The Business of Scholarship*, PAYWALL THE MOVIE (2018), https://paywallthemovie.com/ [https://perma.cc/ZVP8-3GMV] (documentary film about impacts of paywall access to scholarly journals and open access, which includes an interview with Elbakyan).

[313] Bohannon, *supra* note 310.

[314] *See* Flowers, *supra* note 304.

[315] *See id.*

[316] *See id.*

infringement, Elbakyan has also been charged "with illegal hacking under the U.S. Computer Fraud and Abuse Act."[317]

b. Text and Data Mining Sci-Hub Is Lawful

This Article argues that a researcher can legally download all or a portion of the Sci-Hub collection solely for TDM research. This conclusion is based on the same assumptions made above, that copies are made only for computational research and that the durable outputs of any text and data mining analysis would be factual data and would not contain enough of the original expression in the analyzed articles to be copies that count. Reference copies would be kept and shared only for reproducibility purposes or for further computational research and would not be otherwise made available.

The principal argument that could be advanced against such a researcher are that they lose their fair use rights because they are not acting in good faith. Whether good faith is, or should be, a legally cognizable factor in the fair use analysis is a contested issue that cannot be fully discussed within the scope of this Article. A proponent of a good faith inquiry in the present context would likely characterize fair use as an "equitable rule of reason" and would cite for support *Harper & Row, Publishers, Inc. v. Nation Enterprises*,[318] *Atari Games Corp. v. Nintendo of America Inc.*,[319] *Los Angeles News Service v. KCAL-TV Channel 9*,[320] and *NXIVM Corp. v. The Ross Institute*,[321] while recognizing that the Court treated the issue ambiguously in a footnote in *Campbell v. Acuff-Rose Music, Inc.*[322] However, these courts have provided little reasoning to support their position other than to gesture at the judge-made origins of fair use as inviting in all equitable considerations.

[317] Bohannon, *supra* note 299 (describing a civil lawsuit against Elbakyan for violations of the U.S. Computer Fraud and Abuse Act).

[318] Harper & Row, Publishers, Inc. v. Nation Enters., 471 U.S. 539, 562 (1985) ("Also relevant to the character of the use is the propriety of the defendant's conduct. Fair use presupposes good faith and fair dealing.") (internal citation omitted) (internal quotations omitted).

[319] 975 F.2d 832, 843 (Fed. Cir. 1992) ("To invoke the fair use exception, an individual must possess an authorized copy of a literary work.").

[320] 108 F.3d 1119, 1122 (9th Cir. 1997) ("'[T]he propriety of the defendant's conduct' is relevant to the character of the use at least to the extent that it may knowingly have exploited a purloined work for free that could have been obtained for a fee.").

[321] 364 F.3d 471, 475-78, 482 (2d Cir. 2004) (reading *Harper & Row* to "direct[ ] courts to consider a defendant's bad faith in applying the first statutory factor" but then holding the use to be fair after such consideration).

[322] 510 U.S. 569, 585 n.18 (1994).

The courts and commentators who have given the matter more thought have concluded that fair use is a *legal* doctrine that balances the private interests of the copyright owner against the public's interest in receiving the benefits of a secondary use. As such, the subjective motives of the secondary user are irrelevant. Justice Brennan voiced this view in his dissent in *Harper & Row*,[323] and the argument was given additional force by Judge Leval's influential article[324] that provided the Court with its current formulation of the first fair use factor.

My views on this topic are closely in accord with those of Simon Frankel and Matt Kellogg.[325] They identify three contexts in which courts have considered the user's intent as part of the fair use analysis: (1) when the plaintiff challenges the propriety of the user's *access* to the work(s) in suit; (2) when the user did not seek permission to use or was denied such permission; and (3) when the plaintiff alleges a harm from the use that is distinct from economic harm, such as the user's failure to give attribution.[326] With respect to the issue of the user's means of access, they argue that this is an inappropriate consideration because it muddles the line between copyright and other areas of law that govern authorized access, such as trade secret, contract, and, I would add, the Computer Fraud and Abuse Act.[327] Moreover, considering the user's means of access could undermine the very purpose of fair use, which is to provide the public with the benefits of certain secondary uses. Treating an otherwise fair use as unfair because it was made from an infringing source would lead a court to deny the public access to the products of secondary uses that fair use is designed to encourage. In sum, courts should dispense with the good faith inquiry in fair use analysis because:

> It is rooted in a misunderstanding of the "equitable" nature of fair use. It is inconsistent with a traditional analysis of fair use,

---

[323] *Harper & Row*, 471 U.S. at 594 (Brennan, J., dissenting) ("If the Copyright Act were held not to prohibit the use, then the copyright owner would have had no basis in law for objecting.").

[324] Pierre N. Leval, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1126 (1990) (arguing that using a good faith inquiry in fair use analysis "produces anomalies that conflict with the goals of copyright and adds to the confusion surrounding the doctrine"); *see also* Pierre N. Leval, Campbell *as Fair Use Blueprint?*, 90 WASH. L. REV. 597, 612-13 (2015) ("The public's access to important knowledge should not be barred because of bad behavior by the purveyor of the knowledge. A copier's bad faith has no logical bearing on the scope of the original author's copyright.").

[325] *See* Simon J. Frankel & Matt Kellogg, *Bad Faith and Fair Use*, 60 J. COPYRIGHT SOC'Y U.S. 1 (2012).

[326] *See id.* at 23-24.

[327] *Id.* at 24; *see also* 18 U.S.C. § 1030 (2019).

> which centers on the works and not their makers. It tends to confuse fair use with other areas of law like contracts, torts, and criminal law and to introduce new considerations like moral rights without careful inspection. It makes fair use more costly and less predictable for both defendants and plaintiffs and raises concerns about copyright's built-in First Amendment protections. And, perhaps most important, it does not further and often frustrates the basic goal of the fair use doctrine, and of copyright generally, to increase public access to new, socially valuable works.[328]

Even if good faith were relevant, courts have found that knowing use of an infringing source is not bad faith when the user acts in the reasonable belief that their use is a fair use.[329] The implicit theory that the good faith inquiry should consider how a secondary user obtained a copy of the work attempts to import a "fruit of the poisonous tree" doctrine that limits a user's rights with respect to a work of authorship even when the source copy of that work is the result of infringement.[330] This reasoning is entirely circular. Using an infringing copy of a work to make a fair use is only bad faith if one assumes the conclusion that fair use treats using such a copy as bad faith.

In certain specific cases, Congress has limited some user's rights in this manner. For example, such a limit applies to a teacher's public performance of a motion picture in the course of face-to-face teaching if the teacher knew that the performance was from an infringing copy,[331]

---

328 Frankel & Kellogg, *supra* note 325, at 36.

329 *See*, *e.g.*, NXIVM Corp. v. Ross Inst., 364 F.3d 471, 478-79, 482 (2d Cir. 2004) (considering that defendant's "access to the manuscript was unauthorized or was derived from a violation of law" but nonetheless concluding that the use was fair).

330 *See* Mark A. Lemley, *The Fruit of the Poisonous Tree in IP Law*, 103 IOWA L. REV. 245, 248 (2017). The Federal Circuit misread *Harper & Row, Publishers, Inc. v. Nation Enter.*, 471 U.S. 539 (1985), to impose such a requirement. *See* Atari Games Corp. v. Nintendo of Am. Inc., 975 F.2d 832, 834 (Fed. Cir. 1992) ("To invoke the fair use exception, an individual must possess an authorized copy of a literary work."). Two years later the Supreme Court made clear that this limited view of the fair use inquiry is error. *See* Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 585 n.18 (1994) ("If the use is otherwise fair, then no permission need be sought or granted. Thus, being denied permission to use a work does not weigh against a finding of fair use.").

331 *See* 17 U.S.C. § 110(1) (2019) (providing that the following is not an infringement: "performance or display of a work by instructors or pupils in the course of face-to-face teaching activities of a nonprofit educational institution, in a classroom or similar place devoted to instruction, unless, in the case of a motion picture or other audiovisual work, the performance, or the display of individual images, is given by means of a copy that was not lawfully made under this title, and that the person responsible for the performance knew or had reason to believe was not lawfully made").

and the first sale doctrine applies to copies "lawfully made under [Title 17 of the U.S. Code]."[332]

But fair use is not, and should not be interpreted to be, limited by the lawfulness of the *copy* from which the use is made because fair use focuses on fairness of the use of the *work of authorship* regardless of how that work has been embodied. If the use is otherwise fair, then it is definitionally the kind of use the Copyright Act seeks to promote, or at least allow. A limit on fair uses based on the lawfulness of the copy from which the use is made would undermine the balance that fair use provides by depriving society of the benefits of the secondary use as a punishment for a prior infringing act for which liability would already lie.

The case law already recognizes this important point. For example, in the image search cases, the respective records demonstrated that some of the images that each search engine had reproduced and was publicly displaying as thumbnail images was from infringing sources.[333] The Ninth Circuit held in each case that the use was nevertheless fair because the reproductions and displays were made for the transformative purpose of providing a search service.[334]

Moreover, the close relationship between fair use and the fundamental right to freedom of expression supports this result. Fair use is a "built-in First Amendment accommodation []."[335] While the Court has indicated that a speaker's First Amendment interest in using another's original expression to make her point weighs less heavily than the interest in expressing oneself directly,[336] fair use provides the "free

---

[332] *Id.* § 109(a); *see* Kirtsaeng v. John Wiley & Sons, Inc., 568 U.S. 519, 525 (2013) (holding that the first sale limit applies to sales of lawfully made copies outside of the United States that are later imported into the country).

[333] *See, e.g.*, Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1157 (9th Cir. 2007) ("Some website publishers republish Perfect 10's images on the Internet without authorization. Once this occurs, Google's search engine may automatically index the webpages containing these images and provide thumbnail versions of images in response to user inquiries.").

[334] *See, e.g.*, *id.* at 1168:

> In this case, Google has put Perfect 10's thumbnail images (along with millions of other thumbnail images) to a use fundamentally different than the use intended by Perfect 10. In doing so, Google has provided a significant benefit to the public. Weighing this significant transformative use against the unproven use of Google's thumbnails for cell phone downloads, and considering the other fair use factors, all in light of the purpose of copyright, we conclude that Google's use of Perfect 10's thumbnails is a fair use.

[335] Eldred v. Ashcroft, 537 U.S. 186, 219 (2003).

[336] *See id.* at 221.

speech safeguard" necessary to provide for such uses when appropriate.[337]

The Court's reasoning demonstrates that the speech interests of a researcher engaged in text and data mining carry additional weight because this secondary use of copyrighted works is necessary for making one's own speech in terms of the results of computational research rather than simply quoting the speech of another.[338] Transformative uses generally provide a public benefit because the secondary use adds something to the original work.

This analysis demonstrates how fair use provides the United States with a competitive edge over the European Union in innovation policy. The flexibility of the fair use analysis and its focus on outputs rather than inputs, makes central the public benefits of certain secondary uses, including uses that promote innovation. In contrast, the European Union's Directive on Copyright in the Digital Single Market ("DSM Directive") greatly limits the scope of its text and data mining exceptions by focusing on how users obtain copies of the materials to be mined.[339] Article 3 of the DSM Directive requires member states to provide a copyright exception for "reproductions and extractions" made by "research organisations and cultural heritage institutions" of works "*to which they have lawful access*."[340] This exception applies only to those with a status of a "research organization" or a "cultural heritage institution" and they must have "lawful access," which means they either already own copies of the materials to be mined or they must obtain copies of those works pursuant to a contract with publishers. Article 7 purports to limit any TDM-related use restrictions that a publisher may impose in such a contract.[341] But, one of the lurking issues in the DSM Directive is whether access is still "lawful" under Article 3 if such access is conditioned on contractual use restrictions that "research organization" has violated.

Article 4 provides a more general TDM exception for the rest of the public, but this exception applies only for TDM on "lawfully accessible works." The contractual override in Article 7 does not apply to this more general exception, and a user may take advantage of this only if

337 Golan v. Holder, 565 U.S. 302, 329 (2012).

338 *Cf.* Bartnicki v. Vopper, 532 U.S. 514, 535 (2001) (holding that the First Amendment protects publication of an illegally intercepted private cell phone conversation when the subject of conversation is a matter of public concern).

339 *See generally* DSM Directive, *supra* note 6.

340 *Id.* art. 3, ¶ 1, at 113 (emphasis added).

341 *See id.* art. 7, ¶ 1, at 114.

the rightsholder has not expressly reserved its rights to control TDM uses, making this a highly contingent exception.[342]

B. Most Copies for Computation Are Transitory

Researchers concerned only about whether copyright law applies to the temporary copies made during the computational analysis step of TDM research can program their machines to run algorithms to make only copies that do not count for copyright purposes. The copies that count have to last for "more than a transitory duration," and in many settings text and data mining processing can be done in less than one second, which should be unequivocally transitory. The *Cablevision* court's caution in stating that this durational limit on copyright had to be interpreted in a context-specific manner is understandable insofar as changes in digital technologies may present cases in which the court would seek to distinguish these buffer copies. At the same time, even if a copy lasting only 1.2 seconds had more economic significance, it is difficult to imagine which facts should appropriately be relevant to determining when that same amount of time would not be "transitory."

The pertinent legislative history of the 1976 Act expresses an intent to insulate all transient copies from liability. In the 1966 House Report, the judiciary committee explained why the fixation requirement had been added to the bill. With respect to the transitory duration language, the committee wrote:

> The discussions on [treating live broadcasts as fixed], as well as questions raised in connection with computer uses, further emphasized the need for a clear definition of 'fixation' that would *exclude from the concept* purely evanescent or transient reproductions such as those projected briefly on a screen, shown electronically on a television or other cathode ray tube, or *captured momentarily in the 'memory' of a computer*.[343]

However, the Second and Fourth Circuits have both reserved some space for judicial discretion in the application of the "transitory duration" requirement for actionable copying.[344] This introduces some uncertainty into the law of temporary copies. While this author has argued that flexible standards in copyright law create space for courts

[342] *See id.* art. 4, ¶ 3, at 114.

[343] H.R. REP. NO. 89-2237, at 45 (1966) (emphasis added).

[344] *See supra* notes 105–108 and accompanying text.

to avoid the costs of a one-size-fits-all approach,[345] it is not clear that treating transitory duration as one of these points of flexibility adds much to the judicial toolkit while potentially sacrificing some certainty for users and copyright owners alike.

Aaron Perzanowski argues in favor of broad flexibility in the application of the transitory duration requirement, relying on a range of contextual information such as media in which a work is embodied; whether the copies are automated by-products of, or are necessary for the operation of, a machine; and whether temporary copies serve as the functional equivalent of more durable copies.[346] Lydia Loren also supports a focus on the market impact of temporary copies in the application of the transitory duration requirement because of the notice role that fixation plays.[347] More radically, Christina Mulligan proposes eliminating the reproduction right altogether to reclaim analog-digital parity in private uses of copyrighted works.[348] Other scholars, however, would keep the reproduction right but do away with the transitory duration requirement altogether because it limits copyrightability for conceptual art, landscape design, and certain other forms of creative expression.[349]

---

[345] *See* Michael W. Carroll, *One for All: The Problem of Uniformity Cost in Intellectual Property Law*, 55 AM. U. L. REV. 845, 899-900 (2006) (arguing that courts can use doctrines that define copyright's scope flexibly to avoid uniformity cost).

[346] Perzanowski, *supra* note 163, at 1107-08.

[347] *See* Lydia Pallas Loren, *Fixation as Notice in Copyright Law*, 96 B.U. L. REV. 939, 964-65 (2016) ("[I]n the infringement context, instantiations that are too evanescent to interfere with the market for tangible manifestations of the copyrighted work are not the concern of the reproduction and distribution rights." (citation omitted)); *see also* Wendy J. Gordon, *An Inquiry into the Merits of Copyright: The Challenges of Consistency, Consent, and Encouragement Theory*, 41 STAN. L. REV. 1343, 1383 (1989); Laura A. Heymann, *How to Write a Life: Some Thoughts on Fixation and the Copyright/Privacy Divide*, 51 WM. & MARY L. REV. 825, 857-59, 872 (2009) (arguing that hinging copyrightability on fixation separates the author from her work and empowers the audience to chart the work's future); Douglas Lichtman, *Copyright as a Rule of Evidence*, 52 DUKE L.J. 683, 730-34 (2003) (arguing that the fixation requirement for copyrightability serves important evidentiary function).

[348] *See* Christina Mulligan, *Copyright Without Copying*, 27 CORNELL J.L. & PUB. POL'Y 469, 470 (2017).

[349] *See, e.g.*, Megan Carpenter & Steven Hetcher, *Function Over Form: Bringing the Fixation Requirement into the Modern Era*, 82 FORDHAM L. REV. 2221 (2014) (arguing that certain forms of contemporary art are unfairly prejudiced by transitory duration requirement); Zahr K. Said, *Copyright's Illogical Exclusion of Conceptual Art*, 39 COLUM. J.L. & ARTS 335, 337 ("This Essay will argue that copyright illogically excludes conceptual art from protection on the basis of fixation, given that well-settled case law has interpreted the fixation requirement to reach works that contain certain kinds of change so long as they are sufficiently repetitive to be deemed permanent."); *see also* Gregory S. Donat, Note, *Fixing Fixation: A Copyright with Teeth for Improvisational*

While law reform on this topic is unlikely, further judicial elaboration of how to apply the transitory duration requirement is already underway. Assuming that some form of discretion is current law, I would favor a variation on the Fourth Circuit's quantitative/qualitative approach. If a copy lasts for a sufficiently short period — up to at least a few seconds — no further inquiry is needed and the copy is not fixed. Copies that last for longer periods should be the subject of a qualitative inquiry that focuses on whether they are transient or intermediate steps in a process and whether they can function as market substitutes for more durable embodiments. On this analysis, *Tickets.com* erroneously reached the fair use issue because the copies made during the web crawler's mining operation were transient, intermediate steps in a process to extract uncopyrightable information.

This qualitative analysis would have its limits. The reproduction right would still apply to a well-established range of intermediate copies. Certainly, the uses of Disney's animations of Pinocchio in advertisements for a film under development would still be copies that count.[350] Similarly, the qualitative step for transitory duration would leave undisturbed the application of the reproduction right in the reverse engineering cases[351] and any similar intermediate copying cases for which fair use is the better mode for assessing competing interests. It is conceivable that certain forms of reverse engineering might occur so quickly as to not cross the fixation threshold in the future, but there is no reason to shift the focus from fair use at this time.

Fortunately for the TDM researcher, neither law reform nor adoption of any scholarly proposals concerning fixation is necessary to conclude that TDM mining is lawful under current law. Even if context should govern which temporary copies are non-transitory, nothing about TDM

Performers, 97 COLUM. L. REV. 1363 (1997); Carrie Ryan Gallia, Note, *To Fix or Not to Fix: Copyright's Fixation Requirement and the Rights of Theatrical Collaborators*, 92 MINN. L. REV. 231, 240 (2007).

350 Walt Disney Prods. v. Filmation Assocs., 628 F. Supp. 871, 875-76 (C.D. Cal. 1986) (rejecting argument that copies of Disney's version of Pinocchio used in advertisements for a planned, non-infringing, "New Adventures of Pinocchio" were transitory steps toward a completed film).

351 *See* Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1518-19 (9th Cir. 1992); *see also* Sony Comput. Entm't., Inc. v. Connectix Corp., 203 F.3d 596, 602-10 (9th Cir. 2000); Bateman v. Mnemonics, Inc., 79 F.3d 1532, 1539 n.18 (11th Cir. 1996); Atari Games Corp. v. Nintendo of Am. Inc., 975 F.2d 832, 843 (Fed. Cir. 1992); Mitel, Inc. v. Iqtel, Inc., 896 F. Supp. 1050, 1056-57 (D. Colo. 1995), *aff'd on other grounds*, 124 F.3d 1366 (10th Cir. 1997); *cf.* DSC Commc'ns Corp. v. DGI Techs., Inc., 81 F.3d 597, 601 (5th Cir. 1996) (holding that defendant could likely show copyright misuse by alleging infringement for intermediate copying of plaintiff's operating system as a step in developing a competing microprocessor chip).

provides a reasoned basis for applying the fixation standard differently than it was to Cablevision's buffer copies. Those copies were made as part of a service sold to consumers,[352] whereas copies made to mine non-copyrightable information are made as part of a research process that increases the store of human knowledge. Context only reinforces the need to treat mining copies as transitory.

This Article further argues that fixation has greater salience for the scope of copyright law in the United States and internationally than has been recognized to date. First, *Cablevision*'s correct interpretation of the Copyright Act — which is consistent with a careful reading of *MAI* — makes the internet safe to make computational copies while text and data mining.[353] Second, the decision blunts attempts by the United States to impose international obligations on its trading partners to regulate "temporary" copies, which, it is argued, include all buffer copies made by internet users and service providers.[354]

This analysis also points out the limits of copyright law as a means to stop data scraping on the internet. When a competitor makes temporary copies of web pages for the purposes of extracting uncopyrightable factual information, such as price data, other sources of law such as the Computer Fraud and Abuse Act[355] and a common law action for trespass to chattels will have to do the work.[356]

For the researcher conducting research through TDM, the above analysis demonstrates that in most cases the temporary copies of journal articles made during the course of the analysis do not count for copyright purposes because of their transitory nature. If the durable results of this computational analysis are primarily factual data drawn from the source articles and do not contain enough original expression from the underlying articles to count, then the operation of TDM tools is outside the scope of U.S. copyright law.

---

[352] *See* Cartoon Network LP v. CSC Holdings, Inc. (*Cablevision*), 536 F.3d 121, 125 (2d Cir. 2008).

[353] *See supra* Parts I.B–I.D (describing text and data mining process, including the short duration during computational analysis).

[354] *See, e.g.*, Sherwin Siy, *Does the TPP (Still) Make Buffer Copies Illegal?*, PUB. KNOWLEDGE BLOG (Nov. 17, 2014), https://www.publicknowledge.org/blog/does-the-tpp-still-make-buffer-copies-illegal/ [https://perma.cc/JPE3-YDMZ] (providing a leaked draft of the Trans-Pacific Partnership Agreement, which stated that "'fixation' means the embodiment of sounds, or of the representations thereof, from which they can be perceived, reproduced, or communicated through a device.").

[355] *See* 18 U.S.C. § 1030 (2019).

[356] *E.g.*, Intel Corp. v. Hamidi, 71 P.3d 296, 306 (Cal. 2003) (holding that data scraping is actionable as trespass to chattels if the volume of requests impairs the functioning of the server).

CONCLUSION

This Article argues that U.S. copyright law provides a competitive advantage in the global race for innovation because it permits researchers to conduct computational analysis — text and data mining — on any materials to which they have access. The law in the European Union lacks this flexibility, which is why European lawmakers are in the process of adopting a specific exception to permit text and data mining. The United States is squandering its competitive advantage because researchers cannot get access to full-text scholarly journal articles without agreeing to license agreements that limit their use of copyright law's flexibilities. The European proposal would override any contractual restrictions on researchers' rights under the new exception.

Two features of U.S. copyright law make text and data mining lawful in the United States: the limit on the copies that count in Section 106(a) and the way that transformative uses limit the markets that matter in the analysis of fair use under Section 107. Taken together, U.S. researchers may lawfully conduct computational research on any scientific articles or data to which they have access so long as the durable outputs of this research do not incorporate more original expression than is permissible. Moreover, these researchers may store copies of full-text articles for the purpose of enabling others to reproduce their research results or to extend their computational analysis. These copies cannot, however, be used as a substitute source of the articles for human readers. This analysis holds even if the researcher obtains access to scientific journal articles from an infringing source, such as Sci-Hub, because the computational copies made from such a source will not count or will otherwise be covered by fair use, and keeping reproducibility copies will not affect the markets that matter under U.S. copyright law.

For the first audience, competition over text and data mining rights illustrates three large points about this moment in copyright's evolution.

First, the change in European law that requires member states to adopt limitations on copyright to permit computational research recognizes a broader point — the breadth of copyright's reach will stifle innovation and undermine other values in the absence of new exceptions and limitations on the exclusive rights currently required by international treaties, such as the Trade-Related Aspects of Intellectual

Property Rights ("TRIPS") Agreement[357] and the World Intellectual Property Organization ("WIPO") Copyright Treaty.[358] This point is evidenced by the adoption of WIPO's Marrakesh Treaty,[359] which requires member states to adopt provisions permitting copying and distribution of copies of works without permission in formats designed for people with print disabilities. It is further evidenced by the new proposed European directive, which mixes greater expansion of copyright owners' rights with the TDM proposal to scale them back.

Second, the fact that general structural limits in U.S. copyright law enable text and data mining without the need for special legislation is evidence that these flexible user's rights create a legal framework that promotes innovation. Other, more tailored, limitations, such as limitations on internet service providers, also serve this goal.[360]

Third, the strong version of freedom of contract practiced in the United States reduces the competitive edge that innovative users would otherwise enjoy because copyright owners are able to supplement their control over uses of their works by contract.[361]

The spirit of scientific and scholarly inquiry that copyright law is designed to promote is alive and well. Those who support scientific progress and scholarly publishing should work to ensure that copyright law continues to support researchers who seek and find new discoveries by developing and using computational tools that process and extract non-copyrightable information from textual and data sources.

---

[357] Agreement on Trade-Related Aspects of Intellectual Property Rights, Apr. 15, 1994, Marrakesh Agreement Establishing the World Trade Organization, Annex 1C, 1869 U.N.T.S. 299, 33 I.L.M. 1197 (1994).

[358] World Intellectual Property Organization Copyright Treaty, Dec. 20, 1996, 2186 U.N.T.S. 121, 36 I.L.M. 65 (1997).

[359] World Intellectual Property Organization, Marrakesh Treaty to Facilitate Access to Published Works For Persons Who Are Blind, Visually Impaired Person, or Otherwise Print Disabled, June 27, 2013, WIPO Pub. No. 218 (E), https://www.wipo.int/edocs/pubdocs/en/wipo_pub_218.pdf.

[360] *See, e.g.*, Michael W. Carroll, *Pinterest and Copyright's Safe Harbors for Internet Providers*, 68 U. MIAMI L. REV. 421 (2014) (arguing that the limits in 17 U.S.C. § 512 enable wealth creation and innovation for companies such as Pinterest).

[361] *See* Bowers v. Baystate Techs., Inc., 320 F.3d 1317, 1323-24 (Fed. Cir. 2003) (citing authority for the proposition that contracts that limit user rights are not preempted).