Fair use rights to conduct text and data mining and use artificial intelligence tools are essential for UC research and teaching

By Rachael Samberg, Tim Vollmer and Samantha Teremi / March 12, 2024

The UC Libraries strive to preserve fair use rights when licensing electronic resources—including the fair use rights to conduct computational research and incorporate artificial intelligence (AI) tools in academic studies and scholarship.

Academic scholars like those on our campuses use licensed content for computational research, sometimes referred to as text and data mining, or TDM. As the electronic resource licensing landscape has evolved, there has been a concerning rise in publishers' attempts to restrict fair uses, particularly for TDM and any use of AI tools in the process. Fair use restrictions on computational research and AI usage have deleterious effect: they curtail freedom of inquiry, exacerbate bias in research questions and methodologies, and amplify the views of an unrepresentative set of creators given the limited types of materials otherwise available with which to conduct research studies. Such restrictions can also disadvantage UC researchers relative to colleagues in more than forty other countries where publishers are prohibited from using contracts to nullify exceptions to copyright for research.

In this blog post, we explain more about why and how the UC Libraries work to deter fair use restrictions in license agreement negotiations, and to protect UC scholars' ability to make curiosity-driven and cutting-edge discoveries that further the pursuit of knowledge.

Fair Use is essential for research

The University of California produces over 8% of all scholarly publishing in the U.S. and performs 9% of all U.S. academic research and development. UC scholars' extraordinary research—into which nearly five billion dollars are invested each year—betters the world, and advances knowledge in ways that directly improve health, education, technology, literature, art, and quality of life.

Little of this research and publishing would be possible, however, without scholars' reliance on the fair use exception to copyright. Copyright grants exclusive rights to the creators of original expression as an incentive to advance societal knowledge and progress. Were there not exceptions to this exclusivity, no scholar would be able to make use of existing knowledge resources that are needed to create new knowledge goods. All international copyright treaties

permit, and all countries have, exceptions from copyright to support various purposes—including the flexible "fair use" exception for education and research in the United States. Copyright exceptions like fair use matter for research: they result in a higher production of new works of scholarship, and drive emerging research methodologies as a means of extracting information and advancing knowledge. Indeed, in recognition of fair use's paramount importance to the research and teaching enterprise, the University adopted a policy in 2015 pursuant to which it will defend any UC employee or student exercising fair use rights in an informed and reasonable manner, and to the fullest extent of the law.

Emerging contracts threaten fair use

Unfortunately, scholars' and students' reliance on fair use is jeopardized through publishers' imposition of terms in content license agreements to nullify fair use rights. In the U.S., the prospect of "contractual override" means that, although fair use is statutorily prescribed—and supports the wide range of speech contemplated by the First Amendment—private parties may contract to derogate the fair use exception. As such, academic libraries like those within the UC are forced to pay significant sums each year to try to preserve fair use rights for campus scholars within the content license agreements that libraries sign. Often, publishers require libraries to pay additional fees for the right to conduct lawful activities on top of the cost of licensing the content itself. When such costs are beyond institutional reach, the publisher or vendor may then offer similar contractual terms directly to research teams or individuals, who may feel obliged to agree in order to get access to the content they need. Vendors may charge tens or even hundreds of thousands of dollars for this type of access.

This "pay-to-play" landscape of charging institutions for the opportunity to rely on existing statutory rights is particularly detrimental for computational research methodologies like text and data mining, which require the use of "massive datasets with works from many publishers, including copyright owners that cannot be identified or are unwilling to grant licenses." TDM and other computational research methodologies allow researchers to identify and analyze patterns, trends, and relationships across volumes of data that would otherwise be impossible to sift through. Not all TDM research methodologies necessitate usage of AI tools or models. For instance, sometimes TDM can be performed by developing algorithms to detect the frequency of certain words within a corpus, or to parse sentiments based on the proximity of various words to each other. In other cases, though, scholars must employ machine learning techniques to train AI models before the models can make a variety of assessments. And over the past decade, UC scholars have relied on TDM, including TDM with AI modeling, to analyze issues like: changes in gender significance in fiction; slave narratives and perceptions of religion; social and economic impact of Hurricane Katrina on Black New Orleans; the spread of conspiracy theories; and the representation of race, gender, and place in film and television.

The UC Libraries invest more than $60 million each year licensing systemwide electronic content needed by scholars for these and other studies. (Indeed, the $60 million figure represents license agreements made at the UC systemwide and multi-campus levels. But each individual campus also licenses electronic resources, adding millions more in total expenditures.) Our libraries secure campus access to a broad range of digital resources including books, scientific journals, databases, multimedia resources, and other materials. In doing so, the UC Libraries must negotiate licensing terms that ensure scholars can make both lawful and comprehensive use of the materials the libraries have procured. Increasingly, however, publishers and vendors are presenting libraries with content license agreements that attempt to preclude, or charge additional and unsupportable fees for, fair uses like training AI tools in the course of conducting TDM.

These publisher efforts contravene settled law, as every court case to have addressed fair use in the context of computational research has confirmed that the reproduction of copyrighted works to create and mine a collection of copyright-protected works is indeed fair (see *Authors Guild v. HathiTrust*, *Authors Guild v. Google*, and *A.V. ex rel. Vanderhye v. iParadigms*). These cases further hold that making derived data, results, abstractions, metadata, or analysis from the copyright-protected corpus available to the public is also fair use, as long as the research methodologies or data distribution processes do not re-express the underlying works to the public in a way that could supplant the market for the originals. And for the same reasons that the TDM processes constitute fair use of copyrighted works in these contexts, the training of AI tools to do that text and data mining is also fair use: the AI is being trained for predictive and analytical purposes, thus similarly transformative (and further bolstered by being undertaken in nonprofit scholarly or educational context); and training AI does not reproduce or communicate the underlying copyrighted works to the public, precluding market supplantation.

While AI training is no different from other TDM methodologies in terms of fair use, there is a distinction to make between the inputs for AI *training* and *generative* AI's *outputs*. The overall fair use of generative AI outputs cannot always be predicted in advance. Yet, training inputs should not be conflated with outputs, and training AI models by using copyright-protected inputs falls squarely within what courts have already determined in TDM cases to be a transformative fair use.

Restrictive agreements disadvantage UC research

If the UC Libraries are unable to protect these fair uses, UC scholars will be at the mercy of publishers aggregating and controlling what may be done with the scholarly record. Further, UC scholars' pursuit of knowledge will be disproportionately stymied relative to academic colleagues in other global regions, given that a large proportion of other countries preclude contractual override of research exceptions.

Indeed, in more than forty countries—including all those within the European Union (EU)—publishers are prohibited from using contracts to abrogate exceptions to copyright in non-profit scholarly and educational contexts. Article 3 of the EU's Directive on Copyright in the Digital Single Market preserves the right for scholars within research organizations and cultural heritage institutions (like those researchers at UC) to conduct TDM for scientific research, and further proscribes publishers from invalidating this exception by license agreements (see Article 7). Moreover, under AI regulations recently adopted by the European Parliament, copyright owners may not opt out of having their works used in conjunction with artificial intelligence tools in TDM research—meaning copyrighted works must remain available for scientific research that is reliant on AI training, and publishers cannot override these AI training rights through contract. Publishers are thus obligated to—and do—preserve fair use-equivalent research exceptions for TDM and AI within the EU, and can do so in the United States, too.

Nevertheless, certain publisher concerns about the usage of AI are understandable. Some publishers have expressed desires to curb AI usage because of apprehension regarding: i. the security of their licensed products, and the fear that researchers will leak or release content behind their paywall; and ii. AI being used to create a competing product that could substitute for the original licensed product and undermine their share of the market. As we have explained elsewhere, while these concerns may be reasonable, they reflect longstanding fears over users' potential generalized misuse of licensed materials for which publishers already can—and do—impose robust and effective contractual restrictions that preclude such outcomes. Additionally banning the use of AI for acts that are *already* forestalled is unnecessary from a business perspective, and harmful to research methodologies that have long been in use.

In all events, adaptable licensing language can address publishers' concerns by reiterating that the licensed products may be used with AI tools only to the extent that doing so would not: i. create a competing or commercial product or service for use by third parties; ii. unreasonably disrupt the functionality of the subscribed products; or iii. reproduce or redistribute the subscribed products for third parties. In addition, license agreements can require commercially reasonable security measures (as also required in the EU) to extinguish the risk of content dissemination beyond permitted uses. In sum, these licensing terms can replicate the research rights that are unequivocally reserved for scholars elsewhere.

As standards around AI continue to develop, we hope to see express contractual allowance for AI training become the norm in academic licensing. While a different legislative and regulatory approach may be appropriate in the commercial context, academic research licenses should preserve fair uses—including the right to undertake TDM and incorporate AI, and without additional costs being passed on to subscribing institutions or individual users—as a fundamental element of ensuring a diverse and innovative scholarly record.