PUBLIC SUBMISSION

Class 7(a) and 7(b)_Reply_Matthew Sag

Posted by the **U.S. Copyright Office** on Mar 11, 2021

View More Comments 159 (/document/COLC-2020-0010-0075/comment)

Share

Comment

Short-Form Reply Comment
Class 7(a) and 7(b)—Text and Data Mining
Submitter: Matthew Sag

Association of American Publishers relies in part on, and misconstrues, my article, The New Legal Landscape for Text Mining and Machine Learning, 66 J. Copyright Soc'y of the U.S.A. 291 (2019). I write to correct the record:

1. The fair use status of TDM is well-established.

AAP relies in part on my article to argue that the fair use status of text data mining (TDM) is reasonably open to doubt. It is not. As I explained, that TDM is fair use is clear from the unambiguous holdings of the two differently constituted panels in HathiTrust and Google Books. Id. 317–19. I wrote, "In all of the Authors Guild decisions, the relevant court held that library digitization to enable TDM research and full text search was transformative; each decision further held that in light of this transformativeness, such uses were fair." Id. at 319. This is not merely a Second Circuit holding. It is also clear from the 2009 Fourth Circuit iParadigms case. That case does not refer to TDM as such, but that is the basic technology that makes plagiarism detection possible and that was challenged as copyright infringement in that case.

Even the most assertive copyright plaintiffs now appear to be unwilling to challenge the central holdings of HathiTrust and Google Books. Both cases were so clear that, in its appeal in Fox News v. TVEyes, Fox did not even try to object to the basic TDM function of the service. What it objected to was the fact that the service would facilitate the public performance of very large chunks of the underlying copyrighted works. Id. at 331. As I explained, nothing in TVEyes disturbs the conclusion that it is now abundantly clear that text mining and other non-expressive uses of in-copyright works qualify as fair use. Id. at 335.

2. Computational analysis is fair use even when combined with "traditional textual analysis."

AAP also relies in part on my article to argue that computational analysis is no longer fair use when

combined with “traditional textual analysis.” Leaving aside whether this is what the exemption seeks to enable, AAP’s argument is a selective reading of my scholarship.

Computational analysis encompasses externally generated observations about a work or a set of works. In an example I discuss in my article, researchers used advanced TDM methods built on computational models of linguistic phenomena to observe trends in the language discussing male- and female-identified characters across 100,000 novels. Within each novel the software they developed made statistical observations about which text was associated with which characters; it also made observations about the gender implications of that text. These are examples of externally generated observations abstracted from the original works. Id. at 296–99. In aggregate, these observations showed that the proportion of female-identified character space declined steadily from the nineteenth century through the early 1960s. They also showed that gender divisions between fictional characters have become less sharply pronounced over the past 170 years.

This project would still be computational research, and still fair use, even if the researchers analyzed snippets of the text of the original works to evaluate their computational analysis. As I explained, “[m]ost researchers using text mining tools will need to compare their metadata to selections of the actual text from time to time in order to evaluate the reliability of an algorithm or some other aspect of their methodology.” Id. at 321. While such expressive uses must be evaluated to confirm they are unlikely to pose any risk of expressive substitution, “limited expressive uses for purposes such as presenting search results in context or verifying the accuracy of results fit easily within the traditional transformative use paradigm.” Id. Any citation of my work to reach the contrary conclusion is mistaken.

3. Security concerns should not stand in the way of granting the exemption.

AAP relies on a statement I made about the risks of providing researchers with direct access to text to suggest that the proposed exemption should be denied, but in fact the article does not oppose direct access to text for researchers. Instead, as noted above, I took the view that researchers should take reasonable precautions to safeguard the collections they create and I noted that "[s]mall digital archives consisting entirely of works that are not in print or that do not have a well-established commercial market should not require the same level of security as a massive repository like the HathiTrust or Google Books. Id. at 355 (emphasis added).

I agree that taking reasonably appropriate security measures is important. But I believe that researchers are capable of adequately securing the sorts of collections they are likely to create under the proposed exemption.

Comment ID

COLC-2020-0010-0198

Tracking Number

km2-3ho2-4oo3

Comment Details | **Submitter Info**

Submitter Name

Matthew Sag