# COPYRIGHT AND COMPUTATIONAL ANALYSIS OF EXPRESSIVE WORKS

MICHAEL W. CARROLL*

*[Note to IPSC participants. This is a précis of a larger project that will likely include versions aimed at non-legal audiences. Your thoughts about scoping this piece would be most welcome. ]*

What is the role of copyright law in the world of "big data"? The term is often used as shorthand for both the size of datasets and to statistical and other forms of computational analysis of these data. Privacy and consumer protection tend to be the fields of law invoked most frequently as sources of regulation of the collection and analysis of these data. But, when data are comprised of original works of authorship, copyright also has a role to play. Within the copyright context, the collection and computational analysis of original works of authorship has emerged as a legal or policy issue concerning "text mining" of research articles, "digital humanities" research or the "non consumptive uses" of texts in the Google Books corpus.

At bottom, however, these terms all refer to the same basic activities of data gathering and analysis. Although, copyright comes to the fore in relation to information resources published through traditional channels, much of the data collected and analyzed as "big data" by Facebook or Google, for example, also involve analysis of copyrighted works of authorship. Therefore, although this project will focus on the copyright analysis of "content mining" in scientific research, the essential copyright analysis applies equally to computational analysis of original works of authorship in these other contexts as well.

Mining scientific publications and data for insights and discoveries holds out great promise for promoting the progress of science and useful arts. Scientific publishers also see it as a potential new line of business. Consequently, the copyright licenses they offer to research libraries for use of databases of publications frequently prohibit or severely limit users' abilities to download textual or graphic data in bulk for computational analysis. At the same time, these publishers are developing proprietary software tools to provide computational analysis of the corpus of their publications.

Responding to pressures from European policymakers who seek a

* Professor of Law and Director, Program on Information Justice and Intellectual Property, American University Washington College of Law.

competitive edge through better content mining,[1] these publishers are in an active debate with open access advocates about whether content mining requires a license under copyright law, and if it does whether the proper policy response is to facilitate licensing or to increase limitations and exceptions to enable content mining without the need for licensing. The International Association of Scientific, Technical & Medical Publishers ("STM") argue that licensing is necessary and offer "model" licenses that the association claims meets researchers' needs. http://www.stm-assoc.org/text-and-data-mining-stm-statement-sample-licence/. The United Kingdom government has instead pursued copyright reform.[2]

Policymakers in the United States are largely unaware of, or indifferent to, the competitive advantage that the U.S. enjoys to foster computational analysis because most forms of content mining can be done without a license. This point is important because many researchers are unaware of the freedom to mine that they enjoy in the United States and because many research librarians who sign publisher licenses that disallow content mining do not realize that they are signing away a user's right rather than acknowledging a publisher's right.

### Factual Background

*This section will provide a more detailed description of the state of the art in content mining in scientific research. These are still early days. The potential for speeding the pace of discovery is significant, but many observers remain skeptical. These skeptics often overlook the ways in which "big data" computational analysis in the commercial sector already is demonstrating the power of pattern analysis in a range of ways.*

*The section will also foreshadow the legal discussion by showing how mining can be done more or less efficiently depending upon the legal constraints the researcher must work under.*

### Legal Discussion

*This section will make the case for the competitive advantage that U.S. copyright law gives to researchers. Two provisions in particular do this work. First, the author's exclusive right to "reproduce the work in copies" is limited because a "copy" must be "fixed," which means that it must be capable of being perceived, reproduced or*

---

[1] http://blogs.nature.com/news/2014/04/european-commission-report-urges-legal-reform-to-help-scientists-text-mine-research-papers.html.

[2] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/315014/copyright-guidance-research.pdf.

*communicated for a period of more than transitory duration. Accepting* Cablevision's *interpretation of this right, my analysis shows how content mining does not even exercise the copyright owner's reproduction right when the process involves making temporary copies that last in memory for a second or less and that result in durable outputs that do not copy expression from articles and other publications. This discussion explores the legislative history of the fixation provision in the course of supporting* Cablevision's *reading of the text.*

*Second, researchers are likely to also want to keep a durable copy of the original works of authorship that they mined as a reference copy to validate their research outputs. So long as they keep these copies private or within the team of researchers with whom they work, making such copies is a fair use. Matthew Sag has already made much of this case, and my discussion will generally agree with his analysis and adapt it to this context. The Second Circuit's recent* HathiTrust *opinion bolsters this analysis, and the court's opinion in Google Books may well do so even more.*

*Finally, this discussion will close with comparative notes about the enabling power of copyright's limits, and also call attention to ways in which U.S. policy and practice should be realigned to make more extensive use of the freedom to mine granted by U.S. law.*