Identifying Information

Name:	Sag, Matthew
School:	Loyola University Chicago School of Law

Paper Information

Title:	The legal landscape for text mining and machine learning, fair use and beyond
Abstract:	Individually and collectively, copyrighted works have the potential to generate information that goes far beyond what their individual authors expressed or intended. Various methods of computational and statistical analysis of text—usually referred to as text data mining ("TDM") or just text mining—have the potential to unlock that information. However, because almost every use of TDM involves making copies of the text to be mined, the legality of that copying has become a fraught issue. One of the most fundamental questions for copyright law in the Internet age is whether the protection of the author's original expression should stand as an obstacle to the generation of insights about that expression. How this question is answered will have a profound influence on the future of research across the sciences and the humanities, and for the development of the next generation of information technology: machine learning and artificial intelligence. The recent Authors Guild cases (Authors Guild v. Google and Authors Guild v. HathiTrust) provided a partial answer to this question. Now that the dust has settled on the Authors Guild cases, this Article aims to take stock of the legal context for TDM research in the United States. The Authors Guild cases held unambiguously that reproducing copyrighted works as one step in the process of knowledge discovery through text data mining was transformative, and thus ultimately a fair use of those works under United States law. This Article explains why that ruling must be correct as a matter of copyright's most fundamental principles and why the precedent established in the Authors Guild cases is likely to remain settled law in the United States. This Article sets out a four-stage model of the lifecycle of text data mining research and uses this model to identify and explain the relevant legal issues beyond the core holdings of the Authors Guild cases in relation to TDM as a non-expressive use. It is essential to take this broader view into consideration because neither the HathiTrust case, nor the Google Books case addressed issues arising under contract law, the Computer Fraud and Abuse Act, the Digital Millennium Copyright Act, or cross-border copyright issues. Furthermore, although Google Books addressed the display of snippets of text as part of the communication of search results, and both Authors Guild cases addressed security issues that might bear upon the fair use claim, those holdings were a product of the particular factual circumstances of those cases and can only be extended cautiously to other contexts.