Hillary K. Miller. Securing Text and Data Mining Rights for Researchers in Academic Libraries. A Master's Paper for the M.S. in L.S degree. April, 2015. 50 pages. Advisor: Anne Klinefelter

This study describes the results of an online survey of librarians involved in license negotiations at academic institutions in the United States. The survey sought to discover the approaches taken by academic librarians to secure text and data mining rights through licensing of electronic resources.

Headings:

Academic libraries -- Law & legislation

Fair use (Copyright)

License agreements

Text mining (Information retrieval)

SECURING TEXT MINING RIGHTS FOR RESEARCHERS IN ACADEMIC LIBRARIES

by
Hillary K. Miller

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Library Science.

Chapel Hill, North Carolina

April 2015

Approved by

Anne Klinefelter

Table of Contents

INTRODUCTION

Academic libraries’ collections contain an increasing amount of electronic resources in place of the print monographs and serials that were once the focus of their acquisitions. Databases, e-journals, data sets, e-books, and other e-resources take up a larger percentage of libraries’ total collections and budgets each year, and the costs of these e-resources are also rising. In order to provide electronic resources to their academic community, libraries must often subscribe to content, sign a license, and pay an annual subscription fee in order to provide these resources to their patrons. Even when they purchase the content outright, libraries still must often sign a license and pay an annual access fee in addition to the purchase cost. Licenses are negotiated between the libraries and publishers or vendors to set the terms by which the resources can be accessed and used. These licenses vary in detail, but in their basic form, they define authorized users and methods of access, permit or restrict certain uses of the content, state the obligations of the licensor and licensee, and contain a host of other clauses that are common to contracts.

Negotiating acceptable terms for accessing and using the resources ensures that libraries are getting the most value out of their spending, and this value is growing more important as e-resource prices increase. There has been discussion about the licensing of text and data mining rights among those librarians who are responsible for license negotiations at their institutions, mostly through listservs and other informal means of communication. Much of the conversation involves what licensing terms are agreeable or unacceptable, especially in terms of securing rights that are otherwise guaranteed under copyright law. When libraries sign licenses, they agree to let the license supersede what their rights would be under copyright alone. If a license is more restrictive than copyright law, the license is governing. Recent court decisions have affirmed that text and data mining is a fair use of copyrighted materials, but this does not guarantee that libraries will be able to maintain all fair use rights when signing a license. In terms of text and data mining, publishers and vendors could require that libraries give up mining rights in order to access materials, either by restricting mining outright or by placing technical restrictions in the license that prohibit researchers from accessing the large body of texts needed for mining. Libraries are poised to be advocates for fair use rights such as text and data mining because they routinely engage in large-scale negotiations on behalf of their patrons.

Therefore, libraries must carefully consider how to secure this fair use right by negotiating licenses that enable text and data mining research rather than limit it. If libraries choose to explicitly include text and data mining rights in licenses, librarians must understand how text and data mining research is undertaken and their researchers' needs in order to effectively communicate with publishers and vendors. They must negotiate licenses that retain all rights they would otherwise have under fair use and that enable researchers to access content for mining and publish their research results. Finally, librarians must communicate those rights to researchers so that the negotiated rights can be exercised.

Defining Text Mining and Researcher Needs

Text mining is an automated process that extracts meaning and underlying data from text. Because text mining projects can take so many different forms, it is difficult to generalize too much about the process, but an overview of general text mining methods will provide context for the kinds of opportunities and obstacles the research presents. Text mining involves copying text in order to create an index that supports analysis of the text. Basic indexes such as those created for search engines can support word searches by identifying where a word is found in a text. More advanced indexes can show relationships between words, discern the meaning of words, predict the co-occurrence of words, or describe relationships between entities (Clark, 2012). Clark (2012) finds is useful "to distinguish between text mining as the extraction of semantic logic from text, and data mining which is the discovery of new insights" (p. 6). Text mining is the process by which unstructured content like natural language can be transformed into structured information that can be used for data mining. This information becomes data that can be mined to discover new relationships or patterns. Therefore, text mining relies on the underlying information in texts, or non-expressive elements, rather than the fixed language of the author. There are also several types of classic analysis done with text mining, like automated classification of texts, sentiment analysis to reveal author emotion, and bias detection to identify author opinions. Clark identifies four broad reasons to undertake text mining research: content enrichment, systematic literature review, discovery of new insights, and computational linguistics research (p. 7).

Text and data mining is increasingly taking place across a variety of disciplines. Dyas-Correia and Alexopolous (2014) provide an excellent overview of examples of text and data mining research in the fields of economics, business, political science, humanities, law, and medicine (p. 211-212). Another example of text mining research was highlighted by Judge Chin in his 2013 decision for the Authors Guild v. Google, Inc. case (S.D.N.Y): ”

> Using Google Books, for example, researchers can track the frequency of references to the United States as a single entity ("the United States is") versus references to the United States in the plural ("the United States are") and how that usage has changed over time (p. 10).

Jockers, Sag, and Schultz (2012) provide another example of a text mining project in which over three thousand nineteenth century texts were analyzed to reveal stylistic differences between male and female authors. Notably, when the results were visually mapped, it was clear that the works of George Eliot sat firmly among those of the men (p. 30).

In order to perform text mining, researchers must acquire a body of texts for analysis, sometimes from multiple publishers or vendors. Researchers’ access can be limited by technological protection measures (TPMs) that are used to control access to and use of e-resources, and circumventing these TPMs could land them in legal trouble due to copyright legislation like the Digital Millennium Copyright Act, which makes circumventing TPMs illegal. As a result, researchers often have to gain permission to access text from copyright holders. Text mining requires that researchers gain access to usable forms of text, or else they may have to go to great lengths to convert and clean the text. Extensible markup language (XML) is a preferred format for text and data mining, as it is both human and machine readable. However, this is true only if researchers require the text alone for their analysis. Researchers may also wish to analyze other content (like images, figures, or videos) that publishers and vendors are less willing to provide. Some publishers offer application programming interfaces (APIs) to help researchers extract the content they need from the publisher's platform or to perform analysis without removing content from the platform, but a researcher mining content from multiple publishers would have to go through multiple technical requirements to get the entire body of texts they need in addition to getting access permissions and rights to utilize and publish their results from each publisher. These information silos present major difficulties for researchers who wish to text mine content across multiple platforms or publishers.

LIBER's (Ligue des Bibliothèques Européennes de Recherche, or the Association of European Research Libraries) response to Elsevier's text and data mining policy, one of the first model policies released by a major publisher, illustrates many of the needs of researchers by highlighting those needs that are restricted or unmet by Elsevier's policy. LIBER is careful to note that the policy "does *not* mean access to the content on the Elsevier Website that universities subscribe to," (2014, p. 2), but rather that the policy only allows access through Elsevier's API. The API allows access only to text, although researchers, as previously noted, may desire access to other content like images, figures, or videos. Elsevier's policy "explicitly prohibits the use of robots, spiders, crawlers or other automated programs, or algorithms to download content from the website itself, which are the most common ways of performing content mining" (2014, p. 2). This clause, though it prohibits the common method of performing mining, is also one that is common in licensing.

Objections to downloading content straight from the publisher website are often supported by claims that this kind of behavior will overload servers. However, this claim has been addressed and found to be questionable at best. Neylon (2014) describes the experience of the open access platform PLOS and its handling of downloading for content mining undertaken on its platform: “Downloads that result from crawling and content mining contribute a trivial amount to the overall traffic at one of the largest Open Access publisher sites and are irrelevant compared to other sources of traffic” (para. 17). Neylon goes on to state that handling high website traffic is “part of the competent management of any modern website” (para. 19) and claims that “[c]ontent mining, even if it occurred at volumes orders of magnitude above what we see currently, would not be a significant source of issues” (para. 19). As noted by LIBER, extraction would be even smaller as the pool of those with access to subscription material is smaller than the general public that has access to PLOS (2014, p. 3).

LIBER addresses several other aspects of Elsevier’s policy that could potentially restrict researcher needs. It takes issue with the arbitrary limits Elsevier has set, including a maximum of 10,000 articles that may be downloaded per week (p. 2), and the requirement that research outputs can only contain “snippets” of 200 characters or less from original text, far less than is often quoted by researchers in research articles and less even than the number of characters in this sentence (p. 3). LIBER also objects to the control Elsevier attempts to exert over research outputs by requiring that outputs be licensed under a CC-BY-NC Creative Commons license, even though the outputs are very likely to be data and underlying facts that are not copyrightable. This requirement can also present problems for researchers with commercial partners, as the CC-BY-NC license prohibits any commercial use of the content (LIBER, 2014, p. 3).

An additional requirement for researchers to sign a click-through license in order to text mine could create liability issues for researchers. This click-through license is in addition to what has already been agreed upon between Elsevier and the institution and can be changed at any time and without notice by Elsevier (LIBER, 2014, p. 4). Most institutional licenses state that the institutional license is governing and supersedes all click-through licenses that might later be presented to library users. However, the click-through license contains a similar clause, making it unclear which terms researchers must abide by if the terms of the click-through and institutional licenses differ (Elsevier, 2014). LIBER's objections to Elsevier's text mining policy show that researchers needs are diverse and are likely to be inhibited by policies that attempt to exert too much control over the access and use of content.

Increase in Text Mining Research

Text mining is a method of research gaining traction in the academic community, and librarians must respond to this change in order to meet researcher needs. Bergman et al. (2013) cite three reasons for the increase in text mining. First is that the rate at which scholarly literature is increasing has severely outpaced the ability of researchers to keep up with the knowledge in their field. Text mining research offers researchers the ability to analyze large bodies of text far faster than would be possible by humans alone (Bergman et al, 2013, para. 2). For example, "for the last ten years alone, the UK PubMed Central (UKPMC) database lists 312,308 citations with the word 'cancer' in the title – browsing them at the leisurely pace of 85 per day will take you about ten years. And by that time, ten years' worth of new articles on cancer will have appeared" (Reimer, 2012, p. 212). The second reason for growth cited by Bergman et al. is the advancement in the text-mining tools themselves, which have become more accurate and able to be applied in a broader way. Third is the amount of openly accessible text that is being made available in digital format (Bergman et al., 2013, para. 2).

Examining the changes in survey results in the six-year span that the Ithaka S + R Faculty Survey Series was conducted (2006, 2009, and 2012) shows how interest and participation in text mining has grown. The survey series is aimed towards researchers and faculty at United States colleges and universities and examines “changes in faculty member research processes, teaching practices, publishing and scholarly dissemination, the role of the library, and the role of the learned society” (Ithaka S+R, 2013, para. 1). In the 2006 survey, text mining was not mentioned in any part of the survey (Housewright and Schonfeld, 2008), which may indicate that text mining was not widespread enough to be considered worth measuring.

However, in the 2009 survey, under twenty percent of total respondents answered that they “at least occasionally use computational methods such as text-mining and data-mining with electronic collections of academic journal articles, though the far greater use of these methods in the sciences (19%) than in the social sciences (15%) or humanities (8%) could be interpreted to suggest that these methods will likely grow in prevalence” (Schonfeld and Housewright, 2010, p. 8). In the 2012 survey, just over fifty percent of respondents said that they felt scholarly e-books would be more useful if they provided the “[a]bility to perform computational analysis (text mining) over a corpus of electronic monographs” (Schonfeld, 2013, p. 33). Also in 2012, just under twenty percent of respondents in the humanities and the sciences and thirty percent of those in the social sciences said that text mining is very important to their research (Schonfeld, 2013, p. 42).

Ann Okerson spoke to the reported numbers of mining requests that publishers are receiving in her address at the 2013 IFLA Conference:

> The reported numbers of requests for data mining are extremely small. My own instincts say there's something incomplete with how we're collecting those numbers, that explicit demand is quite a bit higher, and potential demand is VERY much higher. I surmise that there are interested researchers who are simply doing their own work in ways that don't get on the radar. Let me just say again that in my view demand will rise much higher very soon. (Okerson, 2013, p. 6)

According to study of publisher practices regarding text mining undertaken by Smit and van der Graaf and commissioned by the Publishing Research Consortium, research mining requests received by publishers are indeed quite small (Smit and van der Graaf, 2011, p. 1). But publishers also believe, like Okerson, that they will increase, and they also believe that current numbers are not reflective of actual text mining practices, due to unauthorized crawling and extracting on their platforms that they cannot account for (Smit and van der Graaf, 2011, p. 1). Therefore, in addition to reported numbers of text mining requests, research may also occur on an unauthorized basis, be done on open access content, or as Okerson says, occur in other ways that are not being reported.

Increase in text and data mining is also apparent from the statements of researchers themselves. Jockers et al. explain the stake that digital humanities researchers had in the outcomes of the Authors Guild lawsuits against Google Books and HathiTrust in their article aptly titled “Don’t Let Copyright Block Data Mining” (2012). These two cases required the courts, in part, to decide if text mining is a fair use. If the court were to rule in favor of the Authors Guild and against the fair use defenses of Google and HathiTrust, Jockers et al. believed it would “set a dangerous precedent — that copyright gives authors and publishers the right to control all, even ‘non-expressive,’ uses of their works that involve copying" (2012, p. 30). This group of researchers and scholars did not and do not want to see the ability to perform text mining research restricted by copyright.

The article describes the reasoning behind the submission of amicus curiae briefs to these two court cases by a group of over one hundred digital humanities and law scholars, including the article's authors, which urged the courts to rule in favor of text mining as fair use. This group has now submitted a total of four amicus briefs, two in support of Google and two in support of HathiTrust (one for each party's case heard in district court, and one for the case heard in the circuit court upon appeal by the Authors Guild). A statement expressed in each of the group's amicus briefs describes the impact that text mining has had in the field of humanities research:

> In short, the possibility of mining huge digital archives and manipulating the data collected in the process has inspired many scholars to reconceptualize the very nature of humanities research. For others, it has played the more modest—but still valuable—role of providing new tools for testing old theories, or suggesting new areas of inquiry. (Brief of Digital Humanities and Law Scholars as Amici Curiae in Support of Defendants-Appellees and Affirmance, Authors Guild v. HathiTrust, 2nd Cir. 2013, pp. 10-11)

Though the humanities have shown lower rates of adoption of text mining research, as seen in the Ithaka S+R survey responses, it is clear from the submission of multiple amicus briefs that researchers in this discipline are seeking out ways to actively support their ability and right to text mine.

Copyright Law in Relation to Text and Data Mining

Two exclusive rights of copyright holders found in the Copyright Act of 1976 apply to text mining. These are the rights to reproduce and prepare derivative works based on the copyrighted work (17 U.S.C. § 106). The central copyright problem with text mining is that it requires researchers to create copies of texts in order to manipulate them into a form that can be analyzed. Additionally, the index created from the copyrighted content would be considered a derivative work.

In addition to barriers to text and data mining caused by exclusive rights of copyright holders found in the Copyright Act, more recent legislation has also created problems for researchers. Hannay considers the possibilities of circumventing technological protection measures (TPMs) to gain access to content in light of the Digital Millennium Copyright Act (DMCA) of 1998. Two Federal Circuit decisions in 2004 and 2005 set a precedent that would prevent the DMCA’s prohibitions on circumventing TPMs from applying in cases where the circumvention does not otherwise infringe on any of the copyright holder’s exclusive rights (Chamberlain Group, Inc. v. Skylink Technologies, Inc., 2004 and Storage Technology Corporation v. Custom Hardware Engineering & Consulting, Inc., 2005).

According to the court’s decisions, “[a] copyright owner alleging a violation of section 1201(a) [of the DMCA] consequently must prove that the circumvention of the technological measure either ‘infringes or facilitates infringing a right protected by the Copyright Act’” (Hannay, 2014, p. 54). However, the Ninth Circuit declined to follow this approach in 2010, stating that it did not believe the language of the DMCA supported such an interpretation (MDY Industries, LLC v. Blizzard Entertainment, Inc., 2010). Therefore, Hannay finds that “the conservative approach for the moment at least would be to avoid circumventing technological protection measures and to opt to seek permission from the owner of the protected works” (p. 54).

The International Federation of Library Associations and Institutions (IFLA) supports the idea that the right to read is the right to mine and believes that copyright restrictions on copying should not apply to text mining in the same way that they would apply to copying for other purposes. IFLA’s Statement on Text and Data Mining explains their position:

> The technical act of copying involved in the process of TDM [text and data mining] falls by accident, not intention, within the complexity of copyright laws – in fact analysis of facts and data has been the basis of learning for millennia. As TDM simply employs computers to “read” material and extract facts one already has the right as a human to read and extract facts from, it is difficult to see how the technical copying by a computer can be used to justify copyright and database laws regulating this activity. (2013, p. 2)

IFLA’s statement reflects the concept of non-expressive uses of copyrighted material. That copyright applies to the particular expression of an author and not to the underlying facts, ideas, or discoveries is considered a “bedrock proposition” of copyright law (Halpern et al., 1999, p. 8). Sag refers to technologies like text mining as “copy-reliant technologies,” and describes their copying of works as being done “in order to process them as grist for the mill, raw materials that feed various algorithms and indices” (Sag, 2009, p. 1608). Sag believes, like IFLA, that the incidental copying of copyrighted works by these “copy-reliant technologies” should not be considered as infringing upon exclusive rights of copyright holders.

Fair Use in Relation to Text and Data Mining

In order to justify text mining of copyrighted works, many researchers and scholars have turned to the concept of fair use. Fair use is a doctrine found in Section 107 of U.S. copyright law that provides certain limitations to the exclusive rights of copyright holders, meaning that those making fair use of a work do not have to request permission from or make payment to the copyright holder. Four factors must be used to determine whether a use of a copyrighted material is a fair use, although others can be used. These factors are found in Section 107 of the Copyright Act of 1976:

> the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; the nature of the copyrighted work; the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and the effect of the use upon the potential market for or value of the copyrighted work (17 U.S.C. § 107).

The first, third, and fourth factors are especially applicable to the consideration of text mining. Because text and data mining taking place in an academic institution has a research and education-driven purpose and because it is most likely noncommercial in nature (academic researchers with commercial partners would need to be weighed differently), the first factor weighs in favor of fair use. In addition, courts have increasingly considered the "transformative" nature of a work in their first factor analysis, or the degree to which a use of a copyrighted work differs from the original purpose of the original creator. Text and data mining has been found to be a particularly transformative use, as the individual purposes of each text do not include the kind of large-scale analysis involved in text mining.

The third factor involves the amount of material taken from a copyrighted work, and both quantity and quality are considered. Traditionally, the larger the amount that is taken, the more likely that criterion will weigh against the researcher in a fair use defense. However, as has been noted above, the copying involved in text mining is incidental to the process and for non-expressive purposes, and therefore, the third factor should not weigh against text mining as a fair use. The fourth factor is how the use of the copyrighted work will affect its value or potential market. Text mining, though it involves the creation of a derivative work, does not create a work that substitutes for the original copyrighted work, and therefore the fourth factor does not weigh against it as a fair use.

There are several characteristics of fair use that are important to note: “Fair use is indefinite, vague, deliberately flexible, deliberately subjective, and intended to apply in different situations” (Smith, 2013, p. 59). Fair use decisions are also case-specific, meaning that there are no specific guidelines or “bright line” rules that determine what categories of use are or are not fair. Each potential fair use must be analyzed on the basis of its own specific characteristics and context, including individual text and data mining research projects. While these qualities allow fair use to apply to emerging technologies like text mining and keep copyright law from becoming overly specific or strict, they also create an air of uncertainty for researchers and librarians who must weigh the benefits of their use of copyrighted works against the risks of being sued for infringement. Although text and data mining, particularly in a noncommercial and research-specific context, has a strong case for fair use, publishers and vendors are not obligated to allow researchers to mine:

> Under the current copyright framework, fair use is properly characterized as a Hohfeldian ‘privilege; rather than as a ‘right.’ That is, the public is free to make fair uses of protected works, but rightsholders owe no affirmative ‘duty’ to make their works available for such uses (Parchomovsky and Goldman, 2007, pp. 1521-2).

As noted above, technological protection measures are one way that text and data mining can be restricted since there is no exception for circumventing these measures in cases of fair use.

In recent years, organizations such as the Association of Research Libraries (ARL) have attempted to guide libraries in making fair use decisions through the creation and adoption of codes of best practices. The Association of Research Libraries' *Codes of Best Practices in Fair Use for Academic and Research Libraries* document was created so that "each institution can undertake its own legal and risk analysis in light of its own specific facts and circumstances" (2012, p. 2) when attempting to determine if a particular use of a copyrighted work is fair. *The Codes* state that "for any particular field of activity, lawyers and judges consider expectations and practice in assessing what is 'fair' within that field" (p.8). Therefore, they were also created to note the norms in libraries which could be used in determining fair use cases in court. *The Codes* provide guidance for libraries who wish to secure rights in licenses for nonconsumptive uses like text and data mining:

> Nonconsumptive uses are an emerging phenomenon at many libraries, and despite their obvious transformative character, there is a risk that the opportunity to make use of these techniques will be lost due to overly restrictive licensing provisions. If librarians agree to licensing restrictions that prohibit such uses, they lose their ability to exercise or permit others to exercise their fair use rights. Librarians should be mindful of this as they negotiate license agreements and should work to preserve their patrons' rights to conduct nonconsumptive research across licensed database materials. (p. 25)

Smith (2013) also commented on this section of *The Codes*, stating that "[e]xercising fair use rights for "non-consumptive research" is efficient, reduces transaction costs, and facilitates an activity that is almost certainly transformative and does not pose a threat to any existing market. In such cases, libraries ought to be willing to facilitate this type of activity" (p. 63).

Although guidelines like those produced by the Association of Research Libraries can be helpful to librarians attempting to determine if a particular use is fair, and although they are sometimes used by courts to determine norms within a field, they do not carry any legal authority. Even guidelines agreed upon between those using copyrighted works and copyright holders, like the Agreement on Guidelines for Classroom Copying in Not-For-Profit Educational Institutions with Respect to Books and Periodicals, do not hold legal authority. As noted in the recent court decision from Cambridge University Press v. Patton, these kinds of guidelines are not legally binding, and more recent court decisions have rejected notions of creating hard rules that do not allow for case-by-case analysis (Cambridge University Press v. Patton, fn. 12, p. 22-23). Because there are no specific guidelines for fair use beyond Section 107 of the Copyright Act, researchers using a copyrighted work must try to ascertain if their use is fair by weighing it against the criteria above and by analyzing existing court decisions. However, there will be no way to definitively know unless they are sued for infringement and the court upholds their use as fair. Objections to uses of copyright material can only be enforced by infringement cases being brought to court and the court deciding against a fair use defense.

Court Decisions with Bearing on Text and Data Mining

The Copyright Act of 1976 is the “sole reference point for the granting and regulation of copyright” (Halpern et al., 1999, p. 1). However, “the United States common-law tradition of judicial interpretation, as well as the language of the Act, has made it subject to extensive judicial interpretation by the federal courts” (Halpern et al., 1999, p. 3). Because copyright falls under federal law alone, cases concerning copyright are heard by federal courts. When a case of copyright infringement is brought before a federal trial court, called a District Court, the judge must decide how to interpret the language of the Copyright Act as it applies to the case before the court. The District Court can also look to decisions made by other courts, but it is only bound by precedent set either by the court of appeals presiding over the geographical area (called a circuit) in which the District Court is located or by the Supreme Court, whose decisions set precedent for all federal courts. If a District Court's decision is appealed, it is sent to the federal court of appeals presiding over its circuit, called a Circuit Court. Similarly, a Circuit Court's decision must be based on the language of the Copyright Act and can take into consideration other court decisions, but the Circuit Court is bound only by precedent set by the Supreme Court. However, courts can look to relevant decisions made by lower courts or made outside of their own circuits for guidance, known as persuasive precedent.

Two recent court cases, Authors Guild v. Google Inc. and Authors Guild v. HathiTrust, and their resulting decisions have upheld text mining as a fair use. Authors Guild v. Google Inc. was the result of a lawsuit filed in the Southern District of New York by the Authors Guild, a professional organization representing over 9,000 published authors, against Google, claiming that its Google Book Search (GBS) project violated the copyright of the authors represented by the Authors Guild. GBS is the result of Google's partnership with libraries in which Google digitizes the libraries' book collections and makes the books searchable to the public in an online database. If the book is out of copyright, GBS displays the full text, and if the book is still in copyright, GBS displays only a small snippet of a few sentences surrounding the search term to provide context for the term's usage. GBS includes links to bookstores and libraries where the books can be acquired.

The Authors Guild filed a copyright infringement suit on behalf of the authors it represents in 2005, and in 2008, the parties reached a settlement that would have allowed Google to commercialize its GBS program, by providing free previews and full text for a fee, if it agreed to make payments to the publishers and authors who held the copyrights to the digitized books. Judge Chin, the judge appointed to the case, rejected this settlement, citing the significant control Google would have to exploit works without copyright owner permission (Authors Guild v. Google Inc., 2011). Judge Chin went on to hear arguments on whether or not Google's use of the books constituted a fair use, and in 2013, he granted summary judgment (a judgment granted without full trial when there is no disputation of facts but rather how the law should be interpreted) for Google. In particular, Judge Chin found that the first factor of fair use, the purpose and character of the use, strongly favored a finding of fair use, due to the particularly transformative uses that GBS offers. These transformative uses include text and data mining:

> Google Books is also transformative in the sense that it has transformed book text into data for purposes of substantive research, including data mining and text mining in new areas, thereby opening up new fields of research. Words in books are being used in a way they have not been used before. Google Books has created something new in the use of book text -- the frequency of words and trends in their usage provide substantive information (Authors Guild v. Google Inc., 2013, pp. 20-21).

The Authors Guild has subsequently filed an appeal with the 2nd Circuit Court, but the outcome is likely to again favor Google, since two of the three 2nd Circuit judges hearing the appeal recently upheld fair use in the Authors Guild lawsuit against HathiTrust and the third is a fair use expert who has shown sympathy to Google's fair use defense (Samuelson, 2014, p. 22).

The Authors Guild's lawsuit against HathiTrust, filed in the Southern District of New York in 2011, made claims similar to those in the lawsuit against Google. However, HathiTrust had a much stronger claim for fair use, since its uses of the books digitized by the GBS project have a purely noncommercial and educational purpose. Google provides copies of the books it digitizes from library collections to its library partners, and close to one hundred of these partners formed HathiTrust in 2008 to join their collections into one repository, the HathiTrust Digital Library (HDL). HathiTrust “brings together the immense collections of partner institutions in digital form, preserving them securely to be accessed and used today, and in future generations” (HathiTrust, n.d.).

The District Court granted summary judgment for HathiTrust in 2012, again affirming the transformative nature of copying texts to support research such as text mining. Judge Baer cited the amicus brief submitted by group of digital humanities scholars as a factor in his decision, noting that it served to “confirm that the underlying rationale of copyright law is enhanced by the HDL” (Authors Guild v. HathiTrust, 2012, p. 21). Judge Baer went on to express his strong belief that the uses that HathiTrust was making of copyrighted works are fair: “I cannot imagine a definition of fair use that would not encompass the transformative uses made by Defendants’ MDP [mass digitization project] and would require that I terminate this invaluable contribution to the progress of science and cultivation of the arts” (p. 22). Again, the Authors Guild appealed the decision of the District Court, but the Circuit Court upheld the decision in HathiTrust’s favor, with the exception of a matter of preservation as fair use, which has been remanded back to the District Court (Authors Guild v. HathiTrust, 2013).

While these two cases and the resulting court decisions are wins for researchers who either want to engage in text mining or are currently doing so, it is important to note that actual precedent has only been set in the Second Circuit in the matter of Authors Guild v. HathiTrust. While courts outside of the Second Circuit can take the district court’s decisions in both cases into consideration, they are not bound by precedent to follow those decisions. Until a fair use case regarding text mining is heard before the

Supreme Court and text mining is affirmed as a fair use, there is no binding precedent outside of the Second Circuit.

Licensing

Although fair use protects many uses that researchers and libraries make of copyrighted works, the norm in library acquisitions of electronic resources is licensing. In order to use bodies of texts that can only be accessed by publishers via licenses, researchers at academic institutions must often rely on their libraries to negotiate text mining rights in licenses. Hannay, having concluded that seeking permission from copyright holders to use their content for mining, notes that licensing still holds problems for researchers and the libraries who often sign licenses for electronic resources.

Although text mining has been affirmed as a fair use, especially in the case of non-commercial research performed at an academic institution, licenses offered by publishers and vendors “may place restrictions on the user that either expressly or inferentially bar data mining without permission,” and as Hannay also notes, [c]opyright law has not been interpreted to preempt or override such contractual restrictions” (p. 54). Even Creative Commons licenses may present obstacles for researchers if the license does not allow derivative works or if the attribution requirement of the license becomes a burden when multiplied by the hundreds or thousands of texts used by a researcher (p. 54).

Lipinski’s *The Librarian’s Legal Companion for Licensing Information Resources and Services* (2013) makes many points relevant to text mining, although it does not specifically address it. Lipinski points out the danger in prohibited use clauses, noting that these clauses can be more restrictive than copyright law and can take away fair use rights that the licensee would otherwise hold (p. 476). Most licenses state that copying is limited to a level far below the amount that text mining requires (often described as an "insubstantial" amount, but not specifically determined). Technical restrictions in licenses, such as bulk downloading from the publisher's website, can also restrict a researcher's ability to access the content they need. Licenses may also include language stating that any rights not explicitly permitted in the license are prohibited, which can affect the licensee's right to any fair use or copyright exemption if it is not explicitly granted in the license (Lipinski, 2013, pp. 485-6). If this language is included and the license does not include explicit permission for text mining, which is not a common licensing term at this time, researchers' ability to text mine will be restricted.

Some licenses contain language to the effect that the terms of the license should not be construed to restrict the licensee's or their authorized users' rights under the fair use exemption found in U.S. copyright law, or that the licensee retains all fair use rights under U.S. copyright law. This is preferable to the inclusion of language that could restrict fair use rights. However, if there is a disagreement over what does and does not constitute fair use, libraries may face problems. Lipinski finds that it is important to enumerate rights that are important or could be contentious, but never to limit the licensee to only those rights that are expressly mentioned (2013, pp. 455-6).

Research Questions

Because text mining is a relatively new area of research and licensing, it is difficult to know how academic libraries in the United States are handling the change. Many things need to be determined. How many libraries have received requests from researchers to negotiate text and data mining rights, how many are attempting to negotiate these rights, and how many have been successful? Are librarians working to understand researcher needs at their institutions, and do they have a plan to notify researchers of the resources that do have mining rights? Are librarians aware that there is a strong argument for text mining as a fair use? What kind of solutions do they find most helpful to overcome barriers to licensing?

LITERATURE REVIEW

No study has been done to understand how text and data mining licensing is being handled in academic libraries in the United States. However, a number of studies, presentations, and reports have looked at relevant text and data mining practices and barriers, and some have offered solutions or guidance for researchers and libraries.

A 2011 study performed by Smit and van der Graaf and commissioned by the Publishing Research Consortium explores mining of journal articles from the perspective of publishers, including their text mining policies, how they handle requests for text mining permissions, their attitudes towards text mining of their content, and their future plans to address the growth of text mining research. The study was conducted through interviews with people who are either involved in mining projects or who are involved in obtaining or granting mining permissions, labeled experts by the authors. The study also surveyed 190 publishers found through mailing lists from Crossref and the International Association of STM Publishers. As a result, the study results may be skewed in the direction of scientific, technical, and medical publishers.

According to the study results, experts predicted an increase in text mining research and new areas of text mining research, not only in the life sciences, but also in social sciences, humanities, business, marketing, and law. Experts predicted this increase will occur because of the availability of larger bodies of digital text, better technologies, and easier access to content. Publishers reported seeing an increase in requests in recent years and expected their numbers to increase further. Most requests currently come from services who abstract and index journal content or from corporate rather than non-profit researchers. As previously noted, many of the publishers reported their belief that unreported mining is occurring, due to crawling and extracting on their platforms that they cannot account for. Publishers largely proceed on a case-by-case basis for mining requests, rather than adopting policies. However, there is an increase in publishers who have model licensing language on standby for new or renewed licenses. The top three agreed-upon solutions to problems presented by text mining among publishers surveyed were standardized content formats that would aid in mining, a common mining platform among publishers, and common rules for access among publishers. The interviewees identified as experts were more likely to be opposed to common platforms or common access rules than were publishers, but more likely to support standardized content formats.

Aimed towards librarians, Ann Okerson's 2013 paper presentation at the ILFA World Library and Information Congress addressed challenges that librarians are facing and ways in which they can overcome these challenges to support researchers engaging in text mining. Challenges include gaining access to resources, gaining permission to engage in text mining, and gaining permission for researchers to publish research outputs. Cross-publisher mining is also a challenge due to non-standard formats of content, multiple platforms, and differing license terms among publishers. Licensing and research support were noted as two areas librarians where librarians should aim their focus. Okerson pointed to model license language provided by library organizations and consortia as a way to guide librarians in negotiating licenses. She also pointed to support roles beyond licensing, including educating researchers, connecting researchers to text mining resources and tools they need, aiding project planning, and providing specialized reference librarians with knowledge of text mining research. Okerson offered the following suggestions for librarians:

> - Libraries can become more aware of campus needs and offer support/expertise
> - Libraries can encourage publishers in licensing and researcher support and offer to codevelop some license principles and services
> - Collaboration is required across publishing, libraries, the research community
> - Librarians can participate in facilitating such activities (p. 6)

Okerson also noted that "librarians do not want to see a future where researchers (and libraries) must depend on costly publisher tools and services, in addition to the large sums we are already paying for e-resources" (p. 6). Therefore, librarians must work to develop text mining expertise among researchers and other librarians so that the expertise does not have to be commercialized.

Dyas-Correia and Alexopoulos (2014) also identify and address problems faced by researchers engaging in text mining that the library can address. Such problems include researchers not being aware of licensing terms between the library and content providers and the possibility that they must go through "multiple layers of permissions" in cases where the copyright holder and the provider of the content are not the same and both must grant text mining permission (p. 213). Dyas-Correia and Alexopoulos suggest that librarians consider adding text mining clauses to licenses, or, if this is not possible, facilitating access for researchers on an individual basis. They also suggest that libraries provide access to text mining software or other products, train text mining specialists and subsequently train researchers, explore the possibilities of mining internal data at universities, and work to communicate license terms to researchers (p. 214).

Kelly and McDonald’s 2012 report on behalf of JISC, an organization that provides digital technology support for higher education and research in the United Kingdom, describes a study that set out to discover the potential, costs, benefits, risks, and barriers surrounding text mining research in the United Kingdom. The study concluded that text mining offers substantial benefits, both economically and for the advancement of knowledge across disciplines. However, the study also concluded that there are significant costs to text mining that are inhibiting the possible benefits. These costs come from “access rights to text-minable materials, transactions costs (participation in text mining), entry (setting up text mining), staff and underlying infrastructure” and are shared by both individual researchers and institutions (p. 49). In addition to the costs, there are also significant risks and barrier to text mining, including “uncertainty regarding the legality of text mining” and “lack of support, infrastructure and technical knowledge” (p. 50). These findings are undoubtedly true for the United States as well.

The report offers some possible solutions to the problems faced by researchers and institutions, including the introduction of a text mining exception to UK copyright law (since UK copyright law does not have an equivalent of the US concept of fair use). Since the publication of this report, this solution has become a reality. As of June 1, 2014, UK copyright law includes an exception for text and data mining for non-commercial purposes, although this does not override licenses already in effect that may prohibit mining and does not address the technical barriers still in place (Mounce, 2014). Though no additions or edits have been made to the United States Copyright Act, the court decisions in the Authors Guild v. Google, Inc. and Authors Guild v. HathiTrust cases have provided support for text mining as a fair use exception to copyright law. However, it remains to be seen if these court decisions have greatly increased the uncertainty surrounding the legal aspects of text mining for librarians at academic and research institutions in the US.

METHODOLOGY

The purpose of this study was to discover the approaches of academic librarians in securing text and data mining rights through licensing. The survey included three preliminary questions that required survey participants to self-identify themselves as the correct survey target audience, but did not require them to self-identify with their name or position title:

1) Do you work in an academic library?

2) Is this library located in the United States?

3) Are you involved in licensing negotiations for electronic resources?

Survey participants were also asked to provide the name of their institution, or, if they preferred not to give it, to provide information about their institution, including whether it is public or private, the institution's size, and whether or not the library is a member of the Association of Research Libraries. The institution's size was based the Size and Setting System established by the Carnegie Foundation for the Advancement of Teaching, and the following choices were given to survey participants:

Very small two-year: FTE enrollment of fewer than 500 students
Small two-year: FTE enrollment of 500–1,999 students
Medium two-year: FTE enrollment of 2,000–4,999 students
Large two-year: FTE enrollment of 5,000–9,999 students
Very large two-year: FTE enrollment of at least 10,000 students
Very small four-year: FTE enrollment of fewer than 1,000 degree-seeking students
Small four-year : FTE enrollment of 1,000–2,999 degree-seeking students
Medium four-year: FTE enrollment of 3,000–9,999 degree-seeking students
Large four-year : FTE enrollment of at least 10,000 degree-seeking students

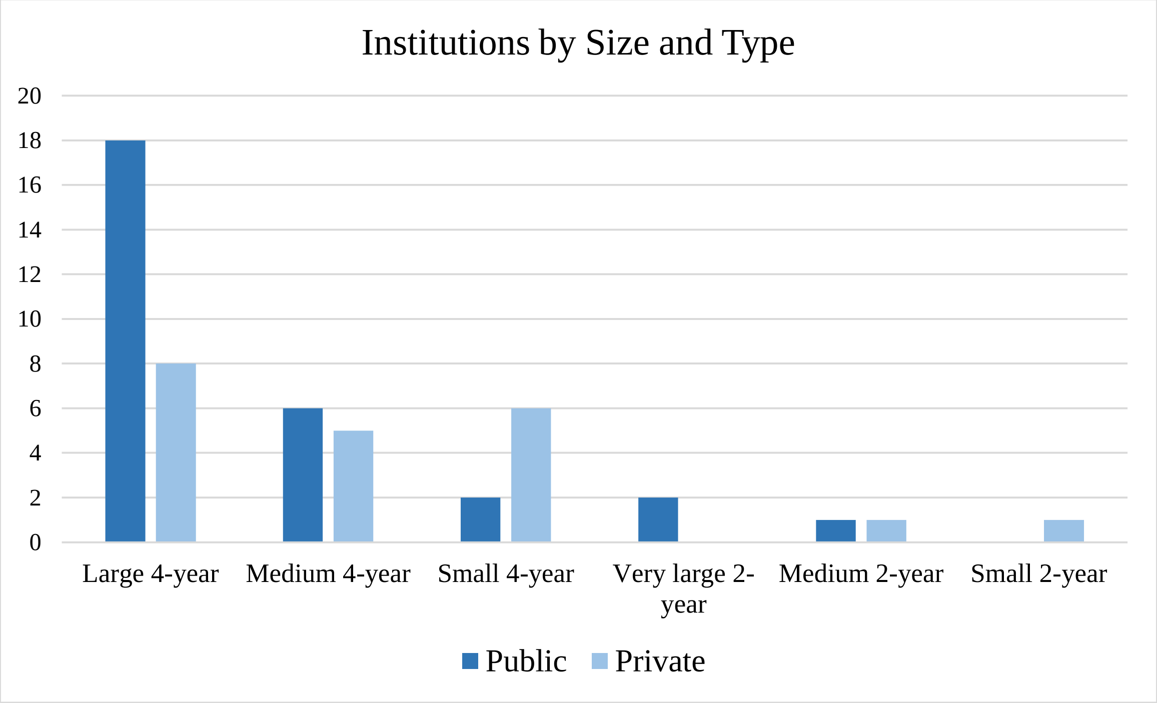
The survey included twelve additional data collections questions that sought to determine how librarians are approaching securing text and data mining rights for their researchers through licensing.

The survey was constructed using Qualtrics software, and a link to the survey was sent to five listservs that are commonly used by academic librarians involved in licensing negotiations and electronic resources acquisitions: ERIL-L (Electronic Resources in Libraries List), LIBLICENSE, SERIALIST (Serials in Libraries Discussion Forum), the ALCTS Electronic Resources Interest Group, and the ACRL Technical Services Interest Group. Sampling from librarians at ARL institutions was considered and decided against for several reasons. First, there is no uniform job title for those who handle licensing, so identifying the correct participant at each institution would have been difficult and time-consuming. Additionally, many institutions have multiple libraries who may each have its own individual acquisitions process and its own librarian who handles license negotiation. A third consideration was the fact that some consortia or library systems may license collectively for all their libraries, adding to the difficulty of identifying the correct participant.

RESULTS

Of the 89 participants who began the survey, 50 participants both qualified for and completed the survey (56%). 10 participants (11%) did not meet the qualifications (working in an academic library in the United States and involvement with electronic resource license negotiation), and the remaining 29 participants (33%) did not complete the survey. Of the respondents, 29 (58%) worked in private institutions, and 21 (42%) worked in public institutions. 24 (48%) of respondents worked in member libraries of the Association of Research Libraries. Figure 1 displays the respondents' institutions by their size and their public or private status.

Figure 1 (Institutions by Size and Type)



Question 1 asked respondents, “Does your library have model or preferred license language that guides your negotiation for electronic resources?” 30 respondents (60%) answered “yes,” and 20 (40%) answered “no.”

Figure 2 (Question 1)

Answer	Number of Responses	Percent
Yes	30	60%
No	20	40%

Question 2 was only displayed to those who answered “yes” to Question 1 and asked respondents, “Has your library added language concerning text and data mining to this model or preferred language?” 12 respondents (40%) answered “yes,” and 18 respondents (60%) answered “no.”

Figure 3 (Question 2)

Answer	Number of Responses	Percent
Yes	12	40%
No	18	60%

Question 3 asked respondents, “When looking for information on licensing text and data mining rights, which resources have you searched?” and asked them to select all of the answers that applied. 14 respondents (28%) selected “I have not looked for this type of information before.” Figure 4 displays the selections of the 36 remaining respondents (72%) who had searched for this type of information. Those who selected “Other” were asked to specify their response. Their responses included model licenses, webinars, and communication with publishers.

Figure 4 (Question 3)

Answer	Number of Responses	Percent
Books	7	19%
Journal articles	22	61%
Websites	31	86%
Listservs	30	83%
Librarian with expertise	22	61%
Other	10	28%

Question 4 asked respondents, “When looking for information on licensing text and data mining rights, which resources have you found helpful?,” and was not displayed to those who answered “I have not looked for this kind of information before” to Question 3. Those who selected “Other” were again asked to specify their response. Their responses included model licenses, webinars, communicating with vendors, and conferences.

Figure 5 (Question 4)

Answer	Number of Responses	Percent
Books	2	6%
Journal articles	11	31%
Websites	25	69%
Listservs	29	81%
Librarian with expertise	22	61%
Other	10	28%

Question 5 asked respondents, “Has your library attempted to negotiate text and data mining rights in an electronic resource license?” This question, as well as questions 6, 7, 9, and 10 included the option to answer “I do not know,” to provide an option for those respondents who might have been part of a team of multiple librarians who negotiate licenses and therefore could not answer for the practices and experiences of their entire institution, or for those respondents who might not have been sure of the practices and experiences of their institutions in the specific context of licensing text and data mining rights. 23 respondents (46%) answered "yes," 24 respondents (48%) answered "no," and 3 respondents (6%) answered "I do not know."

Figure 6 (Question 5)

Answer	Number of Responses	Percent
Yes	23	46%
No	24	48%
I do not know	3	6%

Question 6 asked respondents, "Has your library successfully negotiated text and data mining rights in any licenses?" Of the 23 respondents who answered that their institutions had attempted to negotiate text and data mining rights in an electronic resource license, 19 respondents (82%) answered "yes," 2 respondents (9%) answered "no," and 2 respondents (9%) answered "I do not know."

Figure 7 (Question 6)

Answer	Number of Responses	Percent
Yes	19	82%
No	2	9%
I do not know	2	9%

Question 7 asked respondents, "Has your library received any requests from faculty or other library users to negotiate text and data mining rights for an electronic resource?" 19 respondents (38%) answered "yes," 28 respondents (56%) answered "no," and 3 respondents (6%) answered "I do not know."

Figure 8 (Question 7)

Answer	Number of Responses	Percent
Yes	19	38%
No	28	56%
I do not know	3	6%

Question 8 asked respondents, “Have you consulted with any relevant librarians or researchers at your institution to determine their needs or desires for text and data mining?” 24 respondents (48%) answered “yes,” and 26 respondents (52%) answered “no.”

Figure 10 (Question 8)

Answer	Number of Responses	Percent
Yes	24	48%
No	26	52%

Question 9 asked respondents, “Does your library have a plan in place to notify relevant librarians and researchers of resources that do have text and data mining rights?” 8 respondents (16%) answered “yes,” 36 respondents (72%) answered “no,” and 6 respondents (6%) answered “I do not know.”

Figure 11 (Question 9)

Answer	Number of Responses	Percent
Yes	8	16%
No	36	72%
I do not know	6	12%

Question 10 asked respondents, “Does your library attempt to include fair use clauses in your electronic resource licenses if they are not already present?” 36

respondents (72%) answered “yes,” 13 respondents (26%) answered “no,” and 1 respondent (2%) answered “I do not know.”

Figure 12 (Question 10)

Answer	Number of Responses	Percent
Yes	36	72%
No	13	26%
I do not know	1	2%

Question 11 asked respondents, “In your opinion, is text and data mining a fair use of copyrighted material?”

Figure 13 (Question 11)

Answer	Number of Responses	Percent
Yes	27	54%
No	5	10%
I do not know	18	36%

Question 12 provided a list of proposed methods to simplify the process of securing text and data mining rights for researchers and asked respondents to rate each of the proposed methods by how helpful they thought each would be. The following methods were included:

1. Clarity about the copyright status of derivative works generated by text and data mining
2. Centralized text and data mining platform from which to access content from multiple publishers and vendors
3. Centralized platform to handle permission requests for multiple publishers and vendors
4. Commonly agreed licensing terms for text and data mining that is non-commercial and has a clear research focus
5. Publishers handle requests for text and data mining on individual researcher basis

Respondents were given the options of rating each proposed method as not at all helpful, slightly helpful, moderately helpful, and very helpful. Additionally, respondents could choose not to rate the method if they felt they did not know enough about the method to answer. Figures 14, 15, 16, 17, and 18 present each of the proposed solutions and their perceived helpfulness by number and percentage of respondents selecting each possible response.

Figure 14 (Clarity about the copyright status of derivative works generated by text and data mining)

I do not know enough about this method to answer.		Not at all helpful		Slightly helpful		Moderately helpful		Very helpful	
#	%	#	%	#	%	#	%	#	%
9	18%	0	0%	6	12%	16	32%	19	38%

Figure 15 (Centralized text and data mining platform from which to access content from multiple publishers and vendors)

I do not know enough about this method to answer.		Not at all helpful		Slightly helpful		Moderately helpful		Very helpful	
#	%	#	%	#	%	#	%	#	%
10	20%	1	2%	10	20%	14	28%	15	30%

Figure 16 (Centralized platform to handle mining permission requests for multiple publishers and vendors)

I do not know enough about this method to answer.		Not at all helpful		Slightly helpful		Moderately helpful		Very helpful	
#	%	#	%	#	%	#	%	#	%
9	18%	2	4%	9	18%	15	30%	15	30%

Figure 17 (Commonly agreed licensing terms for text and data mining that is non-commercial and has a clear research focus)

I do not know enough about this method to answer.		Not at all helpful		Slightly helpful		Moderately helpful		Very helpful	
#	%	#	%	#	%	#	%	#	%
2	4%	0	0%	1	2%	11	22%	36	72%

Figure 18 (Publishers handle requests for text and data mining on individual researcher basis)

I do not know enough about this method to answer.		Not at all helpful		Slightly helpful		Moderately helpful		Very helpful	
#	%	#	%	#	%	#	%	#	%
4	8%	20	40%	17	34%	5	10%	4	8%

CONCLUSIONS

About half of the survey respondents reported that their institutions had attempted to negotiate text and data mining rights, but only about a quarter of total respondents had added text and data mining language to their library's preferred licensing language (12 out of the 30 respondents who reported that their library has preferred license language). This suggests that libraries are attempting to negotiate more on a case-by-case basis, and fewer are attempting to include text and data mining licensing as a matter of course in all licensing negotiations.

Of those 19 respondents whose library had received a request to negotiate text and data mining rights, 3 respondents (16%) reported that the library had not attempted to negotiate the rights, while 1 respondent (5%) was unsure if the library had attempted to negotiate the rights. Conversely, of those 28 respondents whose library had not received a request to negotiate text and data mining rights, 7 respondents (25%) reported that the library had attempted to negotiate text and data mining rights in a license. Therefore, 6% of all survey respondents reported that their library had received a request but had not attempted to license, while 14% of all survey respondents reported that their library had attempted to negotiate text and data mining rights in a license without receiving a specific request to do so from faculty or other library users.

A comparison of the resources respondents used to find information with those resources they found to be helpful reveals where helpful information can be found and where it is lacking. The survey, as expected, supported the fact that there is not much traditional literature on the subject of licensing text and data mining rights, particularly in the form of books and journal articles. Only 2 respondents found anything of help in a book; most recent books published on licensing electronic resources do not yet include sections on text and data mining. Websites were the resource that the largest number of respondents had searched for information. However, while 86% of respondents searched websites, only 69% of respondents found websites helpful. Listservs were the next most searched resource (83%), followed by librarians with expertise (61%). Respondents found these resources the most helpful. Only one respondent who sought information from listservs did not find them helpful, and the numbers of respondents who sought help from librarians with expertise and those who found them helpful were the same. Those who answered "other" also found the resources they searched helpful. The specified responses that were searched were all specified again as those that were found to be helpful: model licenses, webinars, and open communication with publishers and vendors. Because librarians with expertise were most consistently found to be helpful, building librarian expertise in this area, both in understanding text and data mining research and in licensing the needed rights, is a priority. Perhaps most importantly, the conversations that occur on listservs and in one-on-one communications must be gathered and summarized into more formal publications so that the guidance they offer can reach the widest possible audience.

A large majority of respondents (72%) reported that their libraries attempt to include fair use clauses in licenses if such clauses are not already present, which is important for retaining a host of fair use rights, including text and data mining.

Respondents were more divided on if they believed text and data mining is a fair use of copyrighted material. While a slight majority of respondents (54%) answered “yes,” 36% of respondents answered “I do not know.” These numbers suggest that a large number of librarians are not confident enough to make a declaration of text and data mining fair use, which is indicative of an uncertainty that can lead to less confident negotiations.

Respondents overwhelming reported that they did not want publishers to handle requests from researchers on a case-by-case basis, thereby leaving researchers to handle negotiations for themselves. Instead, respondents showed a strong preference for developing commonly agreed upon licensing terms: 72% believed it would be very helpful, and another 22% believed it would be moderately helpful. Overall, respondents found it to be more helpful than bringing clarity of the copyright status of text and data mining: only 70% believed it would be either moderately or very helpful.

Clarity about the copyright implications of text and data mining will help give libraries a stronger position from which to negotiate rights in licenses, if they choose to explicitly license rather than rely on fair use. Explicit licensing may also be necessary to prevent technical provisions in licenses from restricting researchers’ ability to mine. However, given the strong case for non-commercial and research-oriented text and data mining as a fair use, mining should not be treated as a right of copyright holders that is granted to licensees but rather as a right that is already possessed by way of fair use. Therefore, if publishers or vendors do not want to allow licensees to exercise that right, libraries should receive some kind of compensation for giving up that right.

Future studies might seek to find out what is driving so many libraries to negotiate text and data mining rights in their licenses. Most of the survey respondents who had attempted to secure these rights had also received specific requests to do so from researchers, but 14% reported that they had attempted to negotiate without a specific request to do so. It is unclear at present what prompted these libraries to negotiate without specific requests. Were they responding to the likelihood that researchers at their institution would be interested in text and data mining, and if so, what indications did they have of researcher interest? Alternatively, were they merely responding to a general trend towards text and data mining as a more common licensing term? Whatever the reasons, it seems that many libraries are proactively pursuing text and data mining rights through licensing.

In order to encourage researchers to engage in text and data mining research, libraries must communicate existing mining rights to relevant liaison librarians and researchers. Only 16% of respondents reported that their library had a plan in place to notify relevant librarians and researchers about mining rights in licenses. If even the most basic language permitting text and data mining is included in a license, encouraging researchers to exercise these rights will provide more use cases and promote greater understanding of researcher needs and how libraries, publishers, and vendors can work together to meet those needs.

BIBLIOGRAPHY

Association of Research Libraries. (2012). Code of Best Practices for Fair Use in Academic and Research Libraries. Retrieved from: http://www.arl.org/storage/documents/publications/code-of-best-practices-fair-use.pdf

Authors Guild v. Google Inc., 770 F.Supp.2d 666 (S.D.N.Y. 2011).

Authors Guild v. Google Inc., 954 F. Supp. 2d 282 (S.D.N.Y. 2013).

Authors Guild v. HathiTrust, 755 F.3d 87, 92-93 (2nd Cir. 2014).

Authors Guild v. HathiTrust, 902 F. Supp. 2d 445 (S.D.N.Y. 2012).

Bergman, C. M., Hunter, L. E., & Rzhetsky, A. (2013, April 17). Announcing the PLOS Mining Collection. Retrieved from http://blogs.plos.org/everyone/2013/04/17/announcing-the-plos-text-mining-collection/

Brief of Digital Humanities and Law Scholars as Amici Curiae in Support of Defendants-Appellees and Affirmance, Authors Guild v. HathiTrust, (2nd Cir. 2013) (no. 12-4547).

Cambridge University Press v. Patton, 769 F.3d 1232, 1267 (11th Cir. 2014).

Chamberlain Group, Inc. v. Skylink Technologies, Inc., 381 F.3d 1178 (Fed. Cir. 2004).

Clark, Jonathan. (2013). Text Mining and Scholarly Publishing, a report for the Publishing Research Consortium. Loosdrecht, The Netherlands & London: 5-6.

Dyas-Correia, Sharon and Michelle Alexopoulos. (2014). Text and Data Mining: Searching for Buried Treasures. Serials Review, 40(3), 210-216. doi: 10.1080/00987913.2014.950041

Elsevier. (9 July 2014). Text and Data Mining Service Agreement. Retrieved from http://dev.elsevier.com/tdm_service_agreement.html

Halpern, Sheldon W., Craig Allen Nard, and Kenneth L. Port. (1999). Fundamentals of United States Intellectual Property Law: Copyright, Patent, and Trademark. The Hague, The Netherlands: Kluwer Law International.

Hannay, W. M. (2014). Legally Speaking -- Of Mindfields and Minefields: Legal Issues in Text and Data Mining. Against The Grain, 26(1), 52-55

HathiTrust. (n.d.). Welcome to the Shared Digital Future. HathiTrust.org. Retrieved from http://www.hathitrust.org/about

Housewright, R., & Schonfeld, R. (2008). Ithaka's 2006 Studies of Key Stakeholders in the Digital Transformation in Higher Education.

International Federation of Library Associations. (2013). IFLA Statement on Text and Data Mining. Retrieved from http://www.ifla.org/files/assets/clm/statements/iflastatement_on_text_and_data_mining.pdf

Ithaka S+R. (2013, April 1). Faculty Survey Series. sr.ithaka.org. Retrieved from http://www.sr.ithaka.org/research-publications/faculty-survey-series

Jockers, M. L., Sag, M., & Schultz, J. (2012). Don't let copyright block data mining: Matthew L. Jockers, Matthew Sag and Jason Schultz explain why humanities scholars have pitched in to the Authors Guild v. Google lawsuit. Nature, 490(7418), 29-30.

Kelly, Ursula and Diane McDonald. (2012.) Value and Benefits of Text Mining. JISC. Retrieved from http://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining

LIBER. (2014). Response to Elsevier's text and data mining policy: a LIBER discussion paper. Retrieved from http://libereurope.eu/blog/2014/03/28/liber-responds-to-elseviers-text-and-data-mining-policy/

Lipinski, Tomas A. (2013). The Librarian's Legal Companion for Licensing Information Resources and Services. Chicago, IL: American Library Association.

MDY Industries, LLC v. Blizzard Entertainment, Inc., 629 F.3d 928, 950 (9th Cir. 2010).

Mounce, Ross. (2014). The right to read is the right to mine: Text and data mining copyright exceptions introduced in the UK. Retrieved from http://blogs.lse.ac.uk/impactofsocialsciences/2014/06/04/the-right-to-read-is-the-right-to-mine-tdm/

Neylon, Cameron. (2014). Best Practice in Enabling Content Mining. Plos.org. Retrieved from http://blogs.plos.org/opens/2014/03/09/best-practice-enabling-content-mining/

Okerson, Ann. 2013. 'Text and data mining: a librarian overview', paper presented at IFLA World Library and Information Congress, 17-23 August 2013, Singapore. Retrieved from http://library.ifla.org/252/1/165-okerson-en.pdf

Parchomovsky, Gideon and Kevin A. Goldman. (Oct. 2007). Fair Use Harbors. Virginia Law Review, 93(6), 1483-1532.

Reimer, T. (2012). Key Issue Text mining, copyright and the benefits and barriers to innovation. Insights: The UKSG Journal, 25(2), 212-215. doi: 10.1629/2048-7754.25.2.212

Sag, M. (2009). Copyright and Copy-Reliant Technology. Northwestern University Law Review, 103(4), 1607-1682.

Samuelson, P. (2014). Mass digitization as fair use. Communications of the ACM, 57(3), 20–22. doi:10.1145/2566965

Schonfeld, R. (2013). Ithaka S + R US Faculty Survey 2012. doi:10.3886/ICPSR34651

Schonfeld, R., & Housewright, R. (2010). Ithaka S + R Faculty Survey 2009: Key Strategic Insights for Libraries, Publishers, and Societies.

Smit, Eefke and Maurits van der Graaf, (2011). Journal Article Mining, a research study into practices, policies, plans.....and promises. Commissioned by the Publishing Research Consortium. Retrieved from http://www.publishingresearch.org.uk/documents/PRCSmitJAMreport2.30June13.pdf

Smith, Kevin and Susan Davis. (2013). Copyright in a Digital Age: Conflict, Risk, Reward. The Serials Librarian: From the Printed Page to the Digital Age, 64(1-4), 57-66. doi: 10.1080/0361526X.2013.759875

Storage Technology Corporation v. Custom Hardware Engineering & Consulting, Inc., 421 F.3d 1307 (Fed. Cir. 2005).

The United States Copyright Act, 17 U.S.C. §§ 101 - 810 (1976).

APPENDIX A

Survey on Licensing of Text and Data Mining Rights

Preliminary Questions

1. Do you currently work in an academic library?
Yes
No (survey will end)

2. Is this library located in the United States?
Yes
No (survey will end)

3. Are you involved in licensing electronic resources?
Yes
No (survey will end)

4. What is the name of your institution? If you would prefer not to answer, you may leave the text box blank and move to the next set of questions, where you will be asked to select the attributes that describe your institution. (Questions 5, 6, and 7 are only shown if Question 4 is left blank.)

5. Is your institution public or private?
Public
Private

6. Please select the size range of your institution:
Very small two-year: FTE enrollment of fewer than 500 students
Small two-year: FTE enrollment of 500–1,999 students
Medium two-year: FTE enrollment of 2,000–4,999 students
Large two-year: FTE enrollment of 5,000–9,999 students
Very large two-year: FTE enrollment of at least 10,000 students
Very small four-year: FTE enrollment of fewer than 1,000 degree-seeking students
Small four-year : FTE enrollment of 1,000–2,999 degree-seeking students
Medium four-year: FTE enrollment of 3,000–9,999 degree-seeking students
Large four-year : FTE enrollment of at least 10,000 degree-seeking students

7. Is your library a member of the Association of Research Libraries?
Yes
No

Data Collection Questions

1. Does your library have model or preferred license language that guides your negotiation for electronic resources?
Yes
No (survey will skip to Question 3)
I do not know

2. Has your library added language concerning text and data mining to this model or preferred language?
Yes
No
I do not know

3. When looking for information on licensing text and data mining rights, which resources have you searched? (select all that apply)
Books
Journal articles
Websites
Listservs
Librarian with expertise
Other (please specify)
I have not looked for this type of information before (survey will skip to Question 5)

4. When looking for information on licensing text and data mining rights, which resources have you found helpful? (select all that apply)
Books
Journal articles
Websites
Listservs
Librarian with expertise
Other (please specify)

5. Has your library attempted to negotiate text and data mining rights in an electronic resource license?
Yes
No
I do not know

6. Has your library successfully negotiated text and data mining rights in any licenses?
Yes
No
I do not know

7. Has your library received any requests from faculty or other library users to negotiate text and data mining rights for an electronic resource?

Yes
No
I do not know

8. Have you consulted with any relevant librarians or researchers at your institution to determine their needs or desires for text and data mining?
Yes
No

9. Does your library have a plan in place to notify relevant librarians and researchers of resources that do have text and data mining rights?
Yes
No
I do not know

10. Does your library attempt to include fair use clauses in your licenses if they are not already present?
Yes
No
I do not know

11. In your opinion, is text and data mining a fair use of copyrighted material?
Yes
No
I do not know

12. The following is a list of proposed methods to simplify the process of securing text and data mining rights for researchers. Please rate each of the proposed solution by how helpful you think it would be:

- Clarity about the copyright status of derivative works generated by text and data mining
- Centralized text and data mining platform from which to access content from multiple publishers and vendors
- Centralized platform to handle permission requests for multiple publishers and vendors
- Commonly agreed licensing terms for text and data mining that is non-commercial and has a clear research focus
- Publishers handle requests for text and data mining on individual researcher basis

1: I don’t know enough about this method to answer
2: not at all helpful
3: slightly helpful
4: moderately helpful
5: very helpful