IDCC18 | *Research Paper*

Data Mining Research with In-copyright and Use-limited Text Datasets: Preliminary Findings from a Systematic Literature Review and Stakeholder Interviews

Megan Senseney
School of Information Sciences
University of Illinois at Urbana-Champaign

Eleanor Dickson
University Library
University of Illinois at Urbana-Champaign

Beth Namachchivaya
Library
University of Waterloo

Bertram Ludäscher
School of Information Sciences
University of Illinois at Urbana-Champaign

Abstract

Text data mining and analysis has emerged as a viable research method for scholars, following the growth of mass digitization, digital publishing, and scholarly interest in data re-use. Yet the texts that comprise datasets for analysis are frequently protected by copyright or other intellectual property rights that limit their access and use. This paper discusses the role of libraries at the intersection of data mining and intellectual property, asserting that academic libraries are vital partners in enabling scholars to effectively incorporate text data mining into their research. We report on activities leading up to an IMLS-funded National Forum of stakeholders and discuss preliminary findings from a systematic literature review, as well as initial results of interviews with forum stakeholders. Emerging themes suggest the need for a multi-pronged distributed approach that includes a public campaign for building awareness and advocacy, development of best practice guides for library support services and training, and international efforts toward data standardization and copyright harmonization.

Submitted 21 January 2018

Correspondence should be addressed to Megan Senseney, 501 E. Daniel St., MC-493, Champaign, IL, 61820. Email: mfsense2@illinois.edu

The 13th International Digital Curation Conference takes place on 19–22 February 2018 in Barcelona, Spain URL: http://www.dcc.ac.uk/events/idcc18/

Introduction

Text data mining (TDM) is rapidly gaining traction among scholars, but the level of interest in this research method exceeds the current level of usage (JISC, 2012). While tools for mining and computational analysis of text datasets have proliferated, the texts themselves are frequently protected by copyright or other intellectual property (IP) rights, or subject to license agreements that limit their access and use. These IP and licensing considerations can complicate a researcher's efforts to access the dataset, incorporate it into analytical research, and communicate the output and related methods transparently to a broader audience.

Libraries already partner with scholars to provide a range of support and management services, and they are well situated to expand support services for TDM that range from gaining access to text data sets to providing training on best practices for workflows and sharing reproducible outcomes. Yet most services in libraries are currently limited to ad hoc access negotiation (Miller, 2015).[1] The full range of issues relating to TDM with use-limited text datasets is poorly understood, and the library community has yet to develop service models for supporting the many facets of text data mining (Orcutt, 2015; Schwarcz, 2017). This paper discusses the potential role of libraries in data mining with in-copyright and use-limited text datasets and reports on activities leading up to an IMLS-funded National Forum of stakeholders that includes librarians, content providers, legal experts, and scholars with active text data mining projects. It includes preliminary findings from a systematic literature review as well as initial results of interviews with forum stakeholders.

Background

Text Data Mining

The terminological imprecision across the literature on text data mining promotes misunderstanding across communities of practice and introduces potential legal risks (Colonna, 2013). Text mining, text data mining, content mining, and computational text analysis are often used interchangeably and described as either a field of inquiry (as in Bergman, Hunter, & Rzhetsky, 2013) or an analytical approach (as in Reilly, 2012). For our purposes, we have elected to use the term *text data mining* (TDM) to refer to computational processes for applying structure to unstructured electronic texts and employing statistical methods to discover new information and reveal patterns in the processed data.

Among the scholars actively engaged in TDM, corpus building and data gathering strategies tend toward the opportunistic. This leads to an overreliance on open access scholarship, works in the public domain, or data provided through a single access

[1] This is not to overlook trailblazing library initiatives that support both the pedagogical and the technical aspects of text data mining and analysis. One notable exemplar is the "Digging Deeper, Reaching Further" initiative out of the University of Illinois, which has developed and disseminated a "train the trainer" curriculum for library and information professionals on text mining and digital scholarship methods (https://teach.htrc.illinois.edu/).

point.[2] Where scholars aren't aware of access restrictions, this same opportunism manifests in the use of technical procedures like web scraping that may violate licensing agreements. The library literature on text data mining often includes anecdotes in which the author first learns of TDM activities on campus when a vendor shuts down access to a database in response to unauthorized use (Dyas-Correia & Alexopoulos, 2014; Williams, et al., 2014; Orcutt, 2015). The legal experts who were interviewed in this study often emphasized that they believe the application of the fair use principles has well-established precedent for TDM. In practice, scholars may find working with in-copyright data too daunting, either because negotiating direct access to the necessary data is too cumbersome a process or because black box solutions for non-consumptive research – discussed in more detail below – add an additional layer of complexity to an already complicated process. Yet working with convenience samples based on the data that are available rather than the data that best support a research question or hypothesis runs the risk of drawing biased, poor, or even dangerous, conclusions from the research results. In effect, the results of data mining are only as good as the quality of the data and its fitness for use.

Use-Limited Data

The phrase *use-limited data*, which we employ throughout this paper, also runs the risk of being misunderstood or variously interpreted. For our purposes, we are interested in textual data where use and access are limited, or potentially limited, due to copyright, licensing, and other contractual terms. During proposal development and throughout the first months of our project, the team sought to distinguish data subject to intellectual property restrictions from data that are restricted due to the ethical and privacy concerns surrounding human subjects. At the time, the team had emphasized difficulties related to acquisition, and early project documents opted to describe these data as "limited-access." Over the course of the literature review and stakeholder interviews, however, the team noted that some form of access ultimately occurs in cases where projects are not abandoned entirely, and scholars working within this framework are occasionally granted unlimited access. The more restrictive facet of research with these data is how they may be used, which encompasses a spectrum of activities ranging from modes of access to redistribution for validation and re-use.

In terms of copyright limitations, original works fall into one of three possible categories: works in the public domain, orphan works, and copyrighted works. Texts in the public domain may have 1) exceeded their copyright period, 2) been released into the public domain by their creator, or 3) been created under conditions such that they are born into the public domain (e.g., government documents). Because these works fall outside the protections of copyright, many scholars presume that they are unrestricted for text data mining purposes. Within the United States, however, contracts may supersede questions of copyright, and it is common to enter into contractual agreements as a condition for accessing texts that have been digitized, organized, or otherwise maintained by third parties. For example, scholars who seek access to public domain texts through the HathiTrust are required to sign a Google Agreement before the data in question will be released. Similarly, access to public domain works may be licensed to intermediaries, such as university libraries, by content providers like Gale

[2] For an illustrative example, refer to Ryan Cordell's anecdote about conducting research with text data from historic U.S. newspapers that excludes representation from the entire state of Massachusetts due to data silos and difficult licensing negotiations (Rathemacher, 2013).

and ProQuest. The terms and conditions are subject to negotiation and may impact use and redistribution. Where licensing agreements are silent on the question of text data mining, researchers are often unsure what activities are allowable. In the case of orphan works, the copyright status of a given text is undetermined, so a researcher's ability to use and reproduce that text is unclear, regardless of the means of access. The legal complexities become even more daunting when considering these issues in an international context.

Among the works in copyright that we consider within the framework of our study, we include all but those that are openly licensed for use and redistribution via schemes such as Creative Commons. For this project we consider digital copies of texts that are owned outright (i.e., scanned directly from print or purchased), digital copies of texts that have been lawfully purchased but include technical protection measures (TPM), texts that sit behind a paywall for which access has been licensed, or in-copyright texts that are freely available on the open web but subject to terms of use.

There are a variety of strategies for provisioning access to these texts. Some content providers have preferred to send physical media by mail while others provide controlled, web-based access. Accepted modes of systematic access for gathering large amounts of data from these sources may vary. In the case of licensed databases and data on the open web, an API may be provided as the preferred means of systematically accessing data for use by machine, and a robots exclusion protocol may specifically disallow scraping content in part or whole. Complicating the question of access on the open web is the fact that terms and conditions may not be enforceable if they are too inconspicuous (e.g., an unobtrusive link to a separate page in a small, light colored font in the footer of a website).

To avoid data security and corpus scale concerns that prohibit distributing text datasets, another strategy for supporting text data mining is to bring the algorithms to the data in such a way that the researcher never has unlimited access to the texts. The most exemplary use case for this model is the HathiTrust Research Center's Data Capsule for non-consumptive research (Zeng, Ruan, Crowell, Prakash, & Plale, 2014). In the 2010 *Amended Settlement Agreement* between the Author's Guild and Google Inc., the term *non-consumptive research* was defined as "research in which computational analysis is performed on one or more Books, but not research in which a researcher reads or displays substantial portions of a Book to understand the intellectual content presented within the Book". Instead of transferring data, the researcher logs in to a secure virtual environment, conducts analysis, and exports results in derived formats that conform to the Non-Consumptive User Research Policy (Dickson et al., 2017). This strategy might be considered analogous to the virtual data enclaves that are used for research with highly sensitive data.[3] Approaches such as the Data Capsule might also provide a starting point for further research towards socio-technical solutions (e.g., to reconcile the competing requirements of limited access on one hand and the transparency and reproducibility of analysis on the other).

A National Forum

Resolving the logistical difficulties of text data mining with use-limited data requires a socio-technical perspective that draws on expertise from a range of stakeholders. Our goal is to guide academic libraries in the development of services that support

[3] While most data enclaves are physical, site-based centers, ICPSR also provides a virtual equivalent for select data (https://www.icpsr.umich.edu/icpsrweb/content/ICPSR/access/restricted/enclave.html).

researchers throughout the TDM process and provides specific recommendations for dealing with texts protected by intellectual property rights. This includes access provision but may extend to guides for advocacy; technical training; strategies for documentation and communication about a researcher's data, methods, and workflows; and best practices for communicating and distributing results in cases where data sharing is desirable but unfeasible or where the research method falls outside disciplinary conventions. This is a first step toward operationalizing legal precedent and information policy into a shared set of procedures and practices, which we believe will contribute to closing the gap between interest and uptake.

To guide our recommendations, the research team is convening a one-and-a-half-day national forum that brings together legal experts, content providers, librarians, researchers, and representatives of key scholarly and professional societies. While achieving full consensus across a diverse group of constituents is unlikely in such a brief time-frame, our goal is to develop a shared understanding of the challenges perceived by each community and establish a common policy, research, and development agenda for libraries and other stakeholders to address these concerns.

Method

The National Forum is preceded by a two-part research initiative: a systematic literature review and semi-structured interviews with participating stakeholders. These activities will drive the development of an initial discussion paper and assist participants in drafting a forum statement and an analysis of Strengths, Weaknesses, Opportunities, and Threats (SWOT).[4] Outcomes from preliminary research will be combined with the forum statements and SWOT analyses to shape the final agenda for the forum. The discussion paper and an annotated bibliography will be shared on the project's website in March 2018, and a final white paper will be published through the Association of College and Research Libraries following the forum.

Literature Review

In fall 2017, we performed a targeted literature review of scholarship on issues related to mining texts that are under copyright, subject to licensing agreements, or otherwise restricted due to intellectual property rights in relation to data mining. The review was limited to works in English from 2000-2017 with an emphasis on research in the United States. Disciplinary coverage includes Law, Library and Information Science, Computer Science, Linguistics, eScience, Digital Humanities, and Computational Social Science. We included any materials that focused on providing library services, developing computational workflows, and addressing issues related to data sharing. Our initial database search returned 103 results across seven categories, with

[4] The purpose of a SWOT analysis is to think strategically and systematically about existing and potential advantages and risks surrounding a certain topic. For this project, participants were asked to think of the object of analysis as "the research enterprise of conducting text mining with text data that is in copyright, licensed, or otherwise protected by intellectual property rights" and to foreground the interests and concerns of their particular stakeholder communities (e.g., researchers, librarians, legal experts, content providers, and professional societies). Participants were also encouraged to consider these issues at multiple levels of granularity from the personal to the organizational to the societal. For more on the origins of and recent extensions to the use of SWOT for strategic planning, see Helms and Nixon, 2010.

the majority of articles discovered in library and information science (42%) or law (27%). Citation chaining has since expanded the body of literature to 150 discrete items.

Stakeholder Interviews

Potential stakeholders were identified through the literature review and subsequent snowball sampling. The final set of 25 forum participants includes representatives of professional societies (CNI, RDA, ACRL, ARL); researchers from across the sciences, digital humanities, and computational social sciences; university-affiliated legal experts specializing in intellectual property and copyright; librarians engaged with research data, licensing, and the development of data service models; and content providers and brokers (Elsevier, Gale, Crossref). As of January 2018, the project team has conducted 23 out of 25 semi-structured interviews with forum participants. Upon completion of all interviews, the project team will conduct a conventional qualitative content analysis of the transcribed interviews to identify key topics and establish cross-cutting themes identified by participants from across different stakeholder communities (Hsieh & Shannon, 2005). Preliminary findings are drawn from initial notes taken by interviewers directly after each session.

Preliminary Findings

Literature Review

As the *technological cost* (Surden, 2013) of digitization and data processing has diminished, courts and creators have begun to address the legality of text data mining, and much of the legal discussion of text mining as it relates to copyright has focused on it as a research method made possible by mass digitization projects. Concepts such as "non-consumptive" and "non-expressive use" emerged from cases where U.S. courts ruled in favor of text- and data-mining uses of digital libraries. Nevertheless, the legal literature pays scant attention to the mechanics and processes of text mining, which is a broad research method encompassing multiple data-analysis techniques, and which intersects with a number of related concepts, including information retrieval, artificial intelligence, and digital humanities. This lack of specificity in the literature exacerbates the blurred boundaries of fair use, which risk-averse universities may be reticent to push (Elkin-Koren and Fishman-Afori, 2017).

The library and information science literature frequently cites uncertainty related to fair use (Miller, 2015), and much of the literature defaults to focusing on TDM licensing negotiations with established commercial vendors (Lowey and Blixrud, 2012; Lammey, 2014; Miller 2015). At present, there is little analysis on the information needs of scholars conducting text data mining or developing models to support the TDM process with use-limited texts beyond the point of acquisition. Yet, the terms of licensed content may impact how scholars use data, document their processes, and communicate their results. The goal of the national forum is to identify and begin to address gaps in the literature to establish a more comprehensive view of text data mining. The forum will ground discussion in scholars' actual practices and information needs. This practice orientation will aid librarians and content providers in reconciling their services with users' requirements while also striving to establish a common framework for assessing and mitigating risks associated with TDM.

Stakeholder Interviews

In one-on-one telephone interviews that explored their perspectives on the current state of using IP-protected text data for TDM, participants frequently characterized the situation as uncertain and intractable. Despite the fact that fair use for text data mining has established legal precedent in the United States, scholars lack clarity on how to proceed in the absence of bright-line rules. This confusion is further complicated by the terms under which textual data (regardless of copyright status) are licensed for use where contractual agreements disallow uses that may otherwise be perceived as fair. Participants observed that for researchers, this may lead to relying on more convenient but incomplete data sets, resulting in conclusions that are biased, poor and possibly dangerous. One researcher speculated that in the absence of more standardized strategies for supporting TDM with use-limited texts, institutions are also likely to continue bearing the burden of research practices that violate contractual agreements. Two participants also addressed the question of critical mass in terms of evaluating current and future investments: one researcher raised concern that funders would lose interest in continuing to support research and development in this area if uptake remains low, and a content provider expressed concern that there is not enough use to warrant the dedicated support and development required for streamlining TDM services.

After completing the SWOT analysis portion of the interview, participants were asked to brainstorm potential strategies for addressing the threats and weaknesses that they identified. Their responses fell into one of four categories: legal uncertainties and legal boundaries, policy and advocacy, training and uptake, and standardization and access workflows.

Legal Uncertainty and Legal Boundaries

While the legal experts repeatedly asserted the fair use argument for text data mining with in-copyright works, they also acknowledged that these cases do not provide guidance on how to proceed in practice. A few participants cited remaining questions about what could be done with their text data after they have created their corpus for analysis, particularly with regard to communicating their results and sharing derivative datasets. One participant advocated that testing the limits of fair use was an opportunity for librarians and other stakeholders to bring clarity to the process. Others were more risk averse, and participants more commonly considered the possibility of legal action a threat rather than an opportunity.

Even in cases where participants felt confident about their standing with regard to in-copyright texts, navigating licenses and other contractual terms and conditions proved more difficult. Several participants acknowledge that the lengthy negotiation process hampers productivity, and one researcher expressed frustration about the chains of communication required to begin the process at all. Several participants also discussed the multiplication of effort that occurs when gaining access to many, discrete text data sets from different sources. One participant discussed prior success with implementing a standard model licensing clause for TDM but cautioned that even where a license may exist for data mining, the researcher might not have the necessary infrastructure, tools, or technical skills in place to act on that license. Another participant appreciated the clarity that negotiated licenses and terms of service can provide for researchers, but expressed concern about the degree to which these terms are obscured or entirely decoupled from the data when using common access mechanisms (the example provided was the use of RSS feeds for gathering news content). Obscurity can also impact whether terms and conditions are unenforceable when accessing data on

the open web, and one participant noted that this creates difficulties in helping scholars evaluate terms of service while still impressing on them the larger licensing landscape and its potential implications when those licenses are binding and enforceable.

A common legal theme across multiple stakeholder groups was the reality that research occurs across multiple institutional and jurisdictional boundaries. International, multi-institutional collaborations are the most affected by research limitations due to intellectual property constraints. Fair use for text data mining only applies in the United States, and researchers in other countries must navigate both copyright and *sui generis* laws pertaining to the databases that contain text data. Participants thinking in an international context were more likely to discuss paths toward formal copyright exceptions for TDM, advocating that the right to read is the right to mine. In the absence of clear copyright exceptions, licensing adds another layer of complexity to collaborative TDM. Licenses with content providers are frequently negotiated at an institutional or consortial level and sharing data outside these boundaries is often specifically disallowed, creating institutional divides among the "haves" and "have nots".

Policy and Advocacy

Consistent with the goals of a SWOT analysis, participants frequently framed their reflections on threats and weaknesses in terms of risk. Participants tended to discuss risk management at the organizational level, with one researcher noting that aversion to legal risk in a university setting introduces intellectual and economic risks in terms of opportunity costs. For content providers granting data access, the risks discussed shifted toward questions of data security and assessing the business case for investing resources toward the development of TDM services. Speaking to both concerns, a legal expert recommended assessing the relationship between market value and security in managing risks.

When asked to reflect on strategies for addressing threats and weaknesses, two participants recommended deeper, continued cross-stakeholder exchanges that extend beyond the level of engagement supported by a national forum grant, suggesting that the root of the difficulty lies in the competing interests and lack of shared goals among multiple stakeholder groups. Several participants recommended moving beyond the personal and institutional levels toward awareness building, advocacy, and policy development at the level of professional societies and government agencies. Recommendations for advocacy included copyright exemptions and open access policies.

Training and Support

Participants regularly discussed the skills and competencies required for conducting text data mining as well as the knowledge necessary for understanding and evaluating intellectual property assertions. One participant spoke at length about data literacy, which runs the spectrum from data access procedures to TDM workflows to transparently communicating the results of analysis. Several participants suggested that guiding scholars through the TDM process was an appropriate and desirable contribution for librarians, which would expand data services from brokering access toward more soup-to-nuts engagement. Without detracting from library-based TDM services entirely, one participant advised against placing too much emphasis on introductory level training and re-skilling initiatives due to the expectation that systems will become easier to use over time and prioritizing less skilled researchers may be a

detriment to high-end research. Among the participants who are currently engaged in some level of TDM service provision, one confided that there were already more requests for text mining support than a single person can manage for the entire institution and another described one institution's current process of assessing current practices with the goal of scaling collaborative services from ad hoc consulting to a more systematic model.

Other topics within the scope of TDM service provision and training explored how operating within a use-limited framework adds layers of complexity to scholarly communication. One researcher cited the importance of working within disciplinary norms where transparency and reproducibility have become important criteria for evaluating scholarship. Within the humanities, another participant discussed reproducibility and data sharing, but this participant situated the problem within a set of disciplines that do not have norms to comfortably allow for communicating the data and methods associated with computational research.

Standardization and Access Workflows

A final major theme to emerge from stakeholder interviews was that of interoperability and standardization. Participants were concerned about the lack of basic shared terminology across disciplinary and professional boundaries, ad hoc procedures for transferring data, uneven data quality, and idiosyncratic use of data formats among content providers. This concern was shared among content providers, researchers, and librarians. Several participants cited the need to create datasets that integrate text data from multiple content providers. Among those who discussed this aspect of TDM, there was also a general concern about the effect of data silos on research and how the absence of standards exacerbated that effect. One participant recommended convening a standards body similar to W3C for text data mining with in-copyright and limited access texts.

Discussion

Without forging a path from legal precedent to practical implementation, our greatest risk is that key stakeholders will lose patience with the enterprise. Funders and content providers may ultimately perceive the obstacles as so intractable that they shift their focus and investments to more accessible content. Researchers who have grown weary of trying to work through the proper channels may adopt extra-legal strategies to obtain content, if they don't abandon their research projects altogether. Together, these observations indicate a need to think critically about institutional risk assessment and risk management strategies, with the goal of balancing legal concerns with economic and intellectual opportunity costs. Beyond the institutional level, there is clear need for a multi-pronged approach that articulates and distributes action across stakeholder groups. Large institutional consortia, professional and scholarly societies, and lobbyists are well positioned to formalize more uniform agreements and data transfer practices with content providers; establish best practices and disciplinary norms; and advocate for legislation that enacts a copyright exception or codifies more open policies for text data mining, a move that may be in the interests of academia and industry alike. Ultimately, this effort exists within the context of global research partnerships, and while European stakeholders in this area frequently cite the United States as leading the way on text data

mining in research, the United States would do well to engage with the European Union on multi-national copyright harmonization efforts.

Library and information professionals in particular have an opportunity and a professional imperative to remain at the forefront of TDM by brokering access, developing collaborative partnerships, and building service models to support researchers through the logistical challenges posed by text data mining with use-limited text datasets. Within the constellation of identified stakeholder groups, libraries are perhaps best situated to make a positive impact across all four of the key thematic areas identified by stakeholder groups. With a history of advocacy for fair use, libraries can – and should – adopt policies and services that enable researchers to exercise fair use to the greatest extent possible. As a professional community that serves the information needs of scholars across disciplinary and institutional boundaries, libraries will also be instrumental in public campaigns to build awareness and advocate for fair policies, in much the same way that libraries have lobbied for open access to federally funded scholarly output. Information professionals already serve in prominent positions on W3C working groups, and their expertise in metadata, data formats, and data transfer protocols is essential to the development of international standards for interoperability and data interchange.

At the local level, incorporating TDM into library-based digital scholarship services would alleviate the pain of trying to identify a centralized point person, which was cited by one of the researchers in our participant interviews. It would also serve to consolidate knowledge and lessons learned from past experiences across multiple units and disciplines, and it could aid in establishing preferred procedures at the local level. Drawing from interviews with library-based participants who are already experimenting with TDM services, we recommend prototyping a service model that could be adopted and implemented by college and research libraries across the country. Early prototypes should develop sample workflows and training modules for guiding scholars through the TDM process, covering topics such as evaluating consumptive vs. non-consumptive approaches to TDM, assessing data quality, acquiring and integrating datasets as needed, processing and analysing data, documenting research workflows, packaging results in shareable formats to support reproducibility, and developing strategies for communicating results. Building upon project outcomes and to guide future hiring and professional development, future work will include articulating the necessary knowledge, experience, and technical competencies of information professionals poised to deliver new services and providing information professionals with strategies for access negotiations and advocacy for use.

Conclusion

Text data mining and analysis methods hold strong potential to enable transformative and significant scholarly inquiry. Libraries are well positioned to facilitate this research as part of digital scholarship and research data services. No single agency or institution can develop the policy and best practices framework for libraries to facilitate access to text datasets for research data mining, but a forum of key stakeholders can serve to catalyze, organize, coordinate, and synthesize the conversation into a cohesive agenda that will serve as a foundation for research and practice in libraries, and across the scholarly community.

Acknowledgments

The project described in this paper has been generously funded by the Institute of Museum and Library Services award LG-73-17-0070-17. The authors would also like to gratefully acknowledge the many contributions of the national forum stakeholder participants and our local advisory committee.

References

Amended Settlement Agreement: Authors Guild, Inc., et al., v. Google Inc. (2009).

Authors Guild, Inc., et al., v. Google Inc. (2015).

Authors Guild, Inc., et al., v. HathiTrust (2014).

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

Colonna, L. (2013). A Taxonomy and Classification of Data Mining. *SMU Science & Technology Law Review*, *16*, 309.

Dickson, E. F., Tracy, D. G., McIntyre, S., Glushko, B., McDonald, R. H., Butler, B., & Downie, J. S. (2017, August). Creating a Policy Framework for Analytic Access to In-Copyright Works for Non-Consumptive Research. Poster presented at Digital Humanities 2017, Montreal, Canada.

Dyas-Correia, S., & Alexopoulos, M. (2014). Text and Data Mining: Searching for Buried Treasures. *Serials Review*, *40*(3), 210–216. https://doi.org/10.1080/00987913.2014.950041

Elkin-Koren, N., & Fischman-Afori, O. (2017). Rulifying Fair Use. *Arizona Law Review, 59*, 161.

Miller, H.K. (2015). Securing Text and Data Mining Rights for Researchers in Academic Libraries [master's thesis]. University of North Carolina at Chapel Hill, Chapel Hill, North Carolina. Retrieved from https://cdr.lib.unc.edu/record/uuid:704c0c1e-e103-4242-85d7-d3abf5b25835

Helms, M. M., & Nixon, J. (2010). Exploring SWOT analysis - where are we now? *Journal of Strategy and Management, 3*(3), 215-251.

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative health research*, *15*(9), 1277-1288.

JISC. (2012). The Value and Benefit of Text Mining to UK Further and Higher Education. *Digital Infrastructure.* Retrieved from http://bit.ly/jisc-textm

Lammey, R. (2014). CrossRef's Text and Data Mining Services. *Learned Publishing, 27*(4), 245–250. https://doi.org/10.1087/20140402

Lowry, C. B. & Blixrud, J. C. (2012). E-Book Licensing and Research Libraries -- Negotiating Principles and Price in an Emerging Market. *Research Library Issues*, (280), 11–19.

Orcutt, D. (2015). Library Support for Text and Data Mining. *Online Searcher, 39*(3), 27–30.

Rathemacher, A. J. (2013). Developing Issues in Licensing: Text Mining, MOOCs, and More. *Serials Review*, *39*(3), 205–210. https://doi.org/10.1080/00987913.2013.10766397

Reilly, B. F. (2012). CRL reports: When machines do research, part 2: Text-mining and libraries. *The Charleston Advisor*, *14*(2), 75–76. Retrieved from http://charleston.publisher.ingentaconnect.com/content/charleston/chadv/2012/00000014/00000002/art00022

Schwarcz, A. (2017, October 20). Text and Data Mining: A New Service for Libraries? [blog post]. Retrieved from https://epthinktank.eu/2017/10/20/text-and-data-mining-a-new-service-for-libraries/

Surden, H. (2013). Technological Cost as Law in Intellectual Property. Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2383529

Williams, L. A., Fox, L. M., Roeder, C., & Hunter, L. (2014). Negotiating a Text Mining License for Faculty Researchers. *Information Technology and Libraries (Online); Chicago*, *33*(3), 5–21.

Zeng, J., Ruan, G., Crowell, A., Prakash, A., & Plale, B. (2014). Cloud Computing Data Capsules for Non-consumptiveuse of Texts. In *Proceedings of the 5th ACM Workshop on Scientific Cloud Computing* (pp. 9–16). New York, NY, USA: ACM. https://doi.org/10.1145/2608029.2608031