JLSC

ISSN 2162-3309 | JLSC is published by the Iowa State University Digital Press | http://jlsc-pub.org

Volume 11, 1 (2023)

Licensing Challenges Associated With Text and Data Mining: How Do We Get Our Patrons What They Need?

Peter McCracken & Emma Raub

McCracken, P. & Raub, E. (2023). Licensing Challenges Associated With Text and Data Mining: How Do We Get Our Patrons What They Need?. *Journal of Librarianship and Scholarly Communication*, 11(1), eP15530. https://doi.org/10.31274/jlsc.15530

This article underwent fully anonymous peer review in accordance with JLSC's peer review policy.

ISSN 2162-3309 10.31274/jlsc.15530

THEORY ARTICLE

Licensing Challenges Associated With Text and Data Mining: How Do We Get Our Patrons What They Need?

Peter McCracken
Cornell University Library (ORCiD: 0000-0002-0145-4253)
Emma Raub
Cornell University Library (ORCiD: 0000-0002-8280-260X)

ABSTRACT

Today's researchers expect to be able to complete text and data mining (TDM) work on many types of textual data. But they are often blocked more by contractual limitations on what data they can use, and how they can use it, than they are by what data may be available to them. This article lays out the different types of TDM processes currently in use, the issues that may block researchers from being able to do the work they would like, and some possible solutions.

Received: 07/22/2022 Accepted: 09/02/2022

IMPLICATIONS FOR PRACTICE

1. Readers will gain an understanding of the significant challenges associated with licensing content for text and data mining (TDM) and how it is similar to the early days of licensing e-journals.
2. This article provides an overview of several different ways in which today's content providers offer access to data for TDM work, ranging from simple delivery of the entire data set through File Transfer Protocol or external media to complex and expensive contractual negotiations that greatly limit researchers' ability to use data for TDM.
3. The article introduces a very recent change in the Digital Millennium Copyright Act, which allows for greater access to data via TDM. Some content providers are apparently still not aware of this change.
4. The authors present a workable option that many vendors could use to provide data sets that are effective for TDM work but do not replace subscription-based full-text access. This approach is already in use by one corpus vendor and could be used by others as well.

INTRODUCTION

Text and data mining, often abbreviated as TDM, allows researchers to gain insights by analyzing large sets of relatively standardized data, generally in a programmatic fashion. By looking at such large collections of data, researchers in many different areas of study can uncover meaningful information that would not have been available to them otherwise. A researcher performing TDM might choose to analyze, for example, a digital newspaper archive spanning a period of decades—something essentially impossible, or at least monstrously time-consuming, through "human" reading. Unless the digital content owner (for example, the vendor or publisher) provides a text analytics platform—as we discuss in our section "Controlled access to subscribed content"—TDM requires a degree of coding skills and statistical knowledge to automatically "scrape" content and data from the selected material. The researcher might then look for patterns, trends, or relationships between words, such as asking when the words "climate change" begin appearing in relation to natural disasters in local newspapers. Materials to be analyzed range from website content (such as publicly available Twitter posts) to ancient manuscripts to scientific papers.

Although TDM work is not especially new, multiple challenges have kept it difficult to complete. While the availability of digitized content has grown immensely in the past decade—and programming tools like R, Python, and Jupyter notebooks offer enhanced methods of

analyzing these data sets—major licensing challenges remain. Vendors and content providers pursue many different paths to maintain control over the data that are being mined. Often, it seems that the text of the license that would oversee the TDM work is more important than the texts that researchers will actually be studying; if librarians are unable to come to an agreement with vendors about the content in the licenses overseeing that work, then the researchers will not be able to access the content in the first place. Or librarians may find themselves pushed to agree to licensing that they would never otherwise accept, after faculty or graduate students have identified and are pursuing access to content that will work for their research but are not aware of significant limitations that the content provider's licensing would impose on them.

As this paper's authors worked to establish TDM access for a collection of faculty and graduate students from a range of disciplines, departments, and even universities, we discovered significant and remarkable limitations in how much we were able to achieve with different content providers. Given the limited understanding of TDM opportunities among faculty researchers within our institution, we felt that an overview of how the market currently exists—and how little it has changed for the better in recent years—would be of particular interest to other librarians who are dealing with similar challenges. These challenges impact librarians outside of e-resources. Any librarian interacting with researchers to support their data collection may encounter TDM questions and obstacles: students approach reference and instruction librarians with academic projects that cannot be addressed by basic database searches; selectors and faculty liaisons must make decisions about purchasing TDM access to content that may be limited to particular researchers or projects; and when researchers unwittingly run afoul of vendors' TDM licensing restrictions through site scraping or excessive downloading, library staff must intervene to get resources up and running again for the entire community. Site scraping, for example, is often completed with an external tool that programmatically reviews and collects ("scrapes") data from websites. Although the broad collection of the data in that manner is probably not legal from a copyright point of view, it is difficult for a vendor to limit or prevent—so when they see rapid and continuous requests for content, they are quick to block access, because they assume that some form of scraping is occurring.

In working to make content legally accessible to patrons, we were surprised to discover a broad range of vendor approaches to TDM, with these approaches guided by a similarly broad range of views on what the vendor could offer, or hoped to offer. There is a very clear correlation between the level of control that a vendor aims to maintain over the TDM content they offer and the flexibility they extend to their customers and the customers' patrons.

We hope that these insights will be useful in understanding why certain publishers are so willing to share the content they offer and why others are so intransigent in limiting patron

access to—and the patrons' scholarly insights of—information drawn from data-driven analysis of large sets of textual data.

LITERATURE REVIEW

In a 2015 master's paper written at the University of North Carolina at Chapel Hill, Hillary K. Miller (2015) provides a significant overview of the status of licensing rights for TDM at that point in time; this review makes it quickly clear that not much has changed in the ensuing seven years. In this literature review, Miller (2015) notes Ann Okerson's 2013 message 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. 10). The future that understandably concerned Okerson and others has, we find, very much come to pass.[1]

Miller (2015) surveyed academic librarians on their experience with licensing for TDM. Among an admittedly small number of respondents, 60% of academic librarians reported that they do have "model or preferred license language that guides your negotiation for electronic resources" (p. 32). About half of the respondents reported that they had attempted to negotiate TDM rights into their licenses, and among those who had tried, over 80% said they had been successful in doing so.

In the same year, Darby Orcutt (2015) argued strongly that TDM should be a basic extension of the services that libraries provide to our patrons. Orcutt had negotiated an initial agreement with Gale/Cengage to provide TDM access to patrons at North Carolina State University, and he (and Gale) based future agreements with other content providers and libraries on their initial license. Orcutt wrote, "Librarians should expect that content mining rights and access come without (much) additional charge. Access to data sets should not be a profit center for vendors" (2015, p. 30). This aspiration has varied greatly across the marketplace; some vendors in fact do not charge for providing content, some charge a great deal, and some focus on charging for the portal that they offer, though it is the only available path to get to their content.

More recently, Courtney et al. (2020) provided a brief but very valuable overview of copyright law as it relates to TDM work, particularly with regard to fair use. The act of preparing textual data for TDM analysis—that is, the scanning and digitizing of the text—was found to be

[1] Dr. Prathik Roy recently presented a webinar about opportunities for TDM within resources provided by his employer, Springer Nature Publishing Group. Dr. Roy's job title is "Group Product Manager for Data Monetization."

transformative by a district court and was affirmed by the US 2nd Circuit Court. But as the authors point out, TDM that could be allowed under fair use permissions can be blocked or forbidden by contractual language that the content provider refuses to modify. These authors, with others, created a 4-day institute (Building LLDTM, 2020) that followed publication of this article and led to publication of a valuable monograph as well (Samberg & Vollmer, 2021).

In the May 2022 issue of *American Libraries*, Carrie Smith gathered feedback from several librarians about three major TDM platforms, from Gale, ProQuest, and LexisNexis (Smith, 2022). Smith's comments (2022) focus on the features and abilities of the three platforms (all mentioned later) but without analysis around licensing concerns in working with these vendors.

SPECIFIC EXAMPLES OF TDM ACCESS AND LIMITATIONS

Overview: Four types of TDM access

We have identified four specific practices by which vendors enable TDM access for libraries and will explain these four practices in more detail in the following subsections. Some provide access to very large data sets through web interfaces that have been designed for TDM work (Option 1 that follows). Some offer essentially free access to collections that libraries have already purchased but offer no platform for doing that work; these vendors generally deliver the data and stop there (Option 2 that follows). Vendors who are reselling a third-party's content may offer some access to that data, but often with significant contractual and technical limitations (Option 3 that follows). Finally, some vendors have created platforms dedicated to TDM work; they hope to be the place where librarians, professors, and students learn how to perform text mining (Option 4 that follows). The last two options may often be combined, because the vendor-specific platform provides a controlled space in which the vendor can provide access to the third-party content, without losing their own control over continued access to the content.

Option 1: Text-based resources optimized for TDM

Among the first of these four groups are a few vendors who provide services specifically built for TDM work and analysis. Google's Ngram Viewer (Google Books Ngram Viewer, n.d.) can be used by some in this space, and Google's Dataset Search (n.d.) can guide users to data sets in repositories across the web. For researchers affiliated with a HathiTrust partner institution, the HathiTrust Digital Library (n.d.) provides access to millions of monographs. (Individuals who are not affiliated with a HathiTrust partner institution do not have

equivalent TDM access.) CORE (CORE, n.d.) claims to host the world's largest collection of open access research papers and makes that collection accessible for text mining through application programming interfaces (APIs) and other tools; access is free to some researchers and fee-based, for others. Although this paper primarily focuses on licensed data sets from traditional library vendors, it is important to note that popular social media platforms also serve as commonly requested data sets. Twitter, for example, offers multiple APIs that one can use to access and download tweets[2].

English-Corpora.org provides a remarkable collection of multiple full-text data sets based on numerous sources, from the official record of the British Parliament to transcripts of US soap operas. Some collections are freely available for use, whereas others are available for purchase or subscription. When downloaded, there are no limitations on how they can be used. Interestingly, however, most of these collections remain covered by copyright that is not held by English-Corpora.org. The vendor manages that limitation by removing 5% of all content. By removing the last 10 of every 200 words, the vendor has created a collection that essentially has no resale value but is still fully valid for linguistic analysis and research (Corpusdata.org, n.d.).

Option 2: Delivering a complete data set

Vendors who control the copyright to either the content or its presentation (if the underlying content is free of copyright restrictions, due to age or US government publication or public copyright license designations, such as CC-BY) may willingly offer to provide the entire set of data via external media if the library has already purchased perpetual access to the collection. Among content providers that we analyzed, most who provided perpetual access did not seriously object to helping the library arrange TDM access for patrons. The library had, after all, already paid for the product; although they likely pay a regular annual platform maintenance fee, and may pay a small fee to get the data delivered in a format that better meets the library's needs (and avoids significant network traffic for the vendor), the vendor has limited opportunities for raising additional revenue from the library through this content.

Accessible Archives, for instance, charges a small fee to deliver a DVD or hard drive that contains content that a library has already purchased, when the library identifies content for TDM research. By the same token, Accessible Archives does not offer TDM access for content to which a library subscribes. When a library subscribes to a product, they will lose access if they

[2] Emory Libraries offers an excellent LibGuide to performing TDM using Twitter's content. See "Using Twitter for Research" (Emory Libraries; https://guides.libraries.emory.edu/main/text-data-mining/twitter [Retrieved August 18, 2022]).

quit paying the subscription rate, so it does not make sense for the vendor to deliver the full content of a database to which a library only subscribes. When a library has purchased perpetual access, however, offering a full content file makes sense. Of course, the patron will not know which resource has been purchased and which is only leased. Electronic resources or TDM-focused librarians must keep track so that they know which databases are eligible for TDM use, and which might require conversion from a subscription to a purchase before TDM use is possible. Adam Matthew Digital offers TDM access similar to Accessible Archives, although Adam Matthew tends to only offer perpetual purchase of content, so they do not need to differentiate between customers who subscribe to their content and those who purchase outright. After the customer completes a one-page form, Adam Matthew will deliver all metadata, raw text files, and finalized content to the customer for TDM use. The librarian remains the critical connection point between the patron and the vendor, managing necessary forms, signatures, and downloads, and ensuring that the delivered data meet the patron's needs as closely as possible. At present, if the content is delivered by File Transfer Protocol, there is generally no fee; a small fee will apply if the content is delivered on a hard drive.

Policy Commons offers full TDM access to the specific collections that a library acquires, and after a representative for the library reviews and signs a single-page agreement, the librarian can download the entire collection and then make it available to the patron, who can text mine that collection with whichever tools that they choose. There is no additional fee, beyond the original access cost, for a subscribing library to use this service. In each example aforementioned, these vendors offer just the data, which libraries can then mine as they see fit; these vendors do not provide online platforms where libraries can perform TDM analysis.

Large aggregators like Gale/Cengage and ProQuest offer content for sale and also through subscription. Given the variations in the products they offer, there is no surprise that they also offer variations on what and how customers can perform TDM work. Generally speaking, if a library has purchased permanent access to content from these vendors, the library can also acquire the underlying data files, upon which text mining can be performed. These vendors will generally provide *access to* subscribed content, although they will not provide the underlying data files.

Gale and ProQuest both offer large sets of historical data that can be mined, after the library has purchased permanent access to the collections. Generally, the underlying content is out of copyright, although the aggregators' modifications of the data, and how it is presented, likely create copyrightable content that is protected from broader distribution. But it is also available to these vendors to use and resell, without their needing to work with a third-party that holds the copyright. These vendors can therefore sell perpetual access to libraries, and also provide the data for TDM use. For this type of content, they can act in the same way that vendors such

as Accessible Archives and Policy Commons do and deliver the full content directly to the library, where patrons can then work with it in whatever environment they choose.

Option 3: Controlled access to subscribed content

For recently published content that is covered by copyright claims, ProQuest, Gale, and others may be able to provide *access to* content that can be data mined, but the copyright holder will rarely allow them to provide a full data set directly to the library. As a result, they are much more likely to require that the content analysis be done in their own application or interface. These vendors will generally create a portal that manages how individuals can use and access their data. To make them more appealing, these portals are commonly offered as tools not just for accessing the content provider's unique content but as a place where students can learn how to perform TDM.

In some cases, a vendor may allow data mining, but they forbid the researcher from downloading the resulting data set. Researchers may be able to preserve some analytical data about the full text, but not the text itself. In addition to ProQuest and Gale, other vendors in this area include Elsevier, LexisNexis, JSTOR, and NewsBank. Given their challenges in negotiating copyright with a primary publisher, and their need to control the data mining process, vendors in this situation often provide the most challenging instances in which to work.

Option 4: Fee-based vendor portals

Gale, ProQuest, JSTOR, and LexisNexis have each created their own online, fee-based platforms, which they hope libraries will use for most of their TDM work. Each space is a revenue-generating platform in which patrons will learn how to perform text analysis and also a space through which the vendor can limit and control access to its proprietary content. Like other platforms, Gale's Digital Scholar Lab and ProQuest's TDM Studio each use Jupyter notebooks and offer Python or R programming languages. Each vendor offers the ability for one to bring external data (for example, data provided on a hard drive or DVD from a vendor like Accessible Archives) into the platform, so it can be mined using the platform's tools; importantly, however, they generally do not allow a library patron to export their proprietary data and analyze it in a different platform. If different platforms offer different tools for analysis, it is possible that a researcher may not be able to perfectly replicate their analysis across different data sets. When using multiple platforms for a single project, librarians and researchers should ensure that their analysis can be sufficiently replicated across those platforms.

JSTOR's TDM platform, called Constellate (Constellate, n.d.), is perhaps more assertive than others about creating a space to learn how to perform text mining. Like the products

mentioned earlier, Constellate gives users access to Jupyter notebooks and the R programming language, but they also offer access to a broad collection of content. This includes, of course, most of the resources in JSTOR, but also some (though not all) of the resources in Portico, plus open access collections and collections from other content providers, as well.

LexisNexis offers several TDM options, such as bulk delivery APIs, search and retrieve APIs, and Nexis Data Lab (LexisNexis, n.d.). Their bulk delivery options deliver large volumes of full-text, unfiltered documents for historical analysis and predictive modeling, while their search and retrieve APIs offer access to data sets with search and retrieve functionality and post-search filters. These services require additional license agreements and costs. With Nexis Data Lab, the institution purchases a "bucket" of IDs so that users can access the Lab and perform data mining. The user receives an ID, runs a search, creates a workspace in the Lab to store it, then uploads up to 100,000 documents, which can be manipulated as needed using Jupyter and Python tools. Patrons can save, download, and reproduce their work, but LexisNexis requires that full-text content remain in the workspace, for copyright compliance. As with the other platforms, the user can also import data from other sources.

Control of access to the vendor's content is clearly an important part of these platforms. Through proprietary platforms, each vendor can block researchers from downloading full data sets and conducting data analysis. These limits may prevent researchers from being able to reproduce their work or share their data sets with others. If a funder or publisher will require that a researcher show reproducibility of their work, it is critical that the researcher investigate and understand what the content provider or platform will allow prior to beginning their data analysis.

As an extreme example, NewsBank allows limited TDM work on some of the data that they offer, but only within very tight parameters. As an aggregator, NewsBank does not hold copyright to the content they offer and states that the high costs associated with their TDM solution are related to fees that they must pay to copyright holders. NewsBank has not built a platform like TDM Studio, Digital Scholar Lab, or Constellate. Instead, they will only allow TDM work to be done within a bespoke online environment that they call a "walled garden." The cost to create this garden, in both time and money, is significant, and it must be repeated for each project that would use these resources. When negotiating with NewsBank, we found their licensing team unwilling to modify their contract to meet current TDM law, or to accept any changes that might have made the agreement more workable for researchers and the library.

In contrast, science-focused publishers such as Elsevier and Clarivate usually provide quick and simple access to their data via APIs, which researchers can use to perform TDM

work, synthesis reviews, and advanced citation analysis[3]. For both Clarivate's Web of Science and Elsevier's ScienceDirect, researchers who want to use the respective APIs must visit a dedicated website where they create an account that is associated with a specific institution's entitlements. They then request an API key. Approval and delivery of the API key may be immediate, or it may take several days. In both cases, the focus is generally on projects such as integrative data analysis, in which multiple independent data sets are combined and analyzed together, rather than on text mining. API access is standardized and extensive.

While Project MUSE does not have a platform for doing TDM work, or for doing more than keyword searching, agreements have already been made with the publishers whose content MUSE is hosting, and they are open and flexible in offering up their data. For subscribers, they offer a variety of enhanced TDM options, including agreements to scrape their site and pull content into the researcher's system. MUSE treats paywall content and open access content the same for TDM. If the institution has purchased the data, then Project MUSE can define and provide targeted, full-text content for researchers' use. TDM use falls under the main MUSE contract: limitations on use are the same as copyright compliance, and researchers perpetually own their gathered data. Content that the institution has not purchased would need to be considered by the MUSE sales team and may require a new license and additional costs.

DISCUSSION AND CHALLENGES FOR RESEARCHERS

When researchers develop papers and reports, they seek materials that are the most relevant to their theses, which likely come from multiple vendors. While acceptance of TDM by vendors is growing, one of the current problems with vendor-permitted TDM is an uncooperative, siloed approach. Rather than broadly permitting TDM of their aggregated materials as a reasonable part of fair use in an era ruled by computers and programming languages, vendors frequently allow TDM only within controlled environments. As noted earlier, multiple vendors provide bespoke platforms in which TDM can be performed on their materials and they will permit uploading of other vendors' materials onto their platform; however, they will not permit their materials to be systematically downloaded and then uploaded into another vendor's site. If each vendor only allows TDM to be performed on their materials within their own platform, then the seeming generosity of allowing another vendor's data into their site is moot, and researchers are left either limiting their TDM to a single vendor or performing it across multiple platforms—a potentially expensive prospect when vendors always charge extra for use of their TDM environments.

Furthermore, the vendor-provided TDM environments often come with user restraints. Researchers may not be able to save the results of their mining but rather extracts of their

[3] Clarivate's API access is distinctly different from the TDM options from ProQuest, which Clarivate recently purchased. TDM options for the two companies are not yet integrated or standardized.

results, sometimes with word limitations, like 50 consecutive words. Because these spaces are produced and controlled by the vendor, they are also freely accessible to the vendor, and we have reviewed TDM agreements that explicitly state that the vendor may check in on the researcher's work at any time to ensure compliance with their terms. When the research project terminates, or if the library ceases to subscribe to that vendor or the vendor's TDM platform, then the space containing the researcher's data ceases to exist or be accessible. The researcher could be forever cut off from the work they need to demonstrate their results. Moreover, many fields demand reproducibility of results by outside academics, and a closed TDM environment makes that impossible.

Although librarians have been arguing for TDM rights for years (Orcutt, 2015), these rights have been slow in coming. Librarians and data managers must continue to push for the standardization of TDM usage rights in a way that allows faculty and students to use the content as effectively and as easily as possible. In addition to ensuring that publishers allow for TDM access, librarians might consider asking TDM-hesitant vendors to apply the English-Corpora approach, in which 10 out of every 200 words are programmatically deleted; vendors could then safely offer complete access to their content without running a risk of delivering the entire full-text collection that they are still licensing. As English-Corpora has found, a data set that is missing 5% of its words is a completely viable tool for data analysis but essentially useless for full-text reading (Corpusdata.org, n.d.). Vendors would not feel required to build a TDM portal in which they control access to data. Patrons would be able to work on a local data set, with the tools that work best for them and that could be used on multiple data sets at one time, and patrons would be able to establish reproducibility and provide examples of the data set they used.

Whether or not vendors are willing to take this approach, it is critical that librarians ensure that vendors hear the needs of our patrons and push back against expensive and onerous TDM limits that quite literally block the development and growth of human knowledge.

DIGITAL MILLENNIUM COPYRIGHT ACT EXEMPTIONS REGARDING TDM

One potential new development in the licensing of products for TDM is the October 2021 regulatory changes by the Librarian of Congress. The Digital Millennium Copyright Act, or DMCA, prohibits users of a copyrighted work from circumventing a "technological measure that effectively controls access to [the] work." However, Congress permits certain non-infringing "fair uses" of copyrighted works to protect freedom of expression and promote further creation. In the DMCA, Congress directed the Register of Copyrights—the head of the Copyright Office within the Library of Congress—to "monitor developments in the marketplace for copyrighted

materials," and Congress authorized the Librarian, upon the Register's recommendation, to grant selective exemptions to the DMCA's anti-circumvention rule.[4]

The most recent exemption to the anti-circumvention rule excludes activities necessary to circumvent technological measures of content in order to conduct TDM, within certain parameters. The activities must be undertaken by researchers, or students or staff at the direction of a researcher, affiliated with non-profit institutions of higher education to deploy TDM techniques on literary works for the purpose of scholarship and teaching. The content must be lawfully acquired or licensed by the institution without a time limitation on access. Researchers' access and viewing of the content is solely for the purpose of verification of their research findings, and the institution must use security measures to prevent further downstream dissemination or downloading of the content.[5 6]

Many universities and their researchers practicing TDM meet the criteria set forth by this exemption. If no license is in place expressly prohibiting TDM, it certainly seems possible to interpret the Librarian of Congress's rulings as permitting TDM on literary works acquired by an institution. If nothing else, a library might consider using the exemption as a point of negotiation for the permission of TDM in a vendor's contract. The exemption pushes us to see TDM as part of fair use, employed for the purpose of further creation, and not as an act of technological theft.

CONCLUSION

As TDM projects become important research tools at institutions of all sizes, libraries and content providers need to develop solutions that make TDM easy to license, manage, and implement. TDM platforms that are designed for teaching students how to implement TDM are valuable tools that will have long-term benefits. Librarians should review the available platforms and, when applicable and appropriate, subscribe to and implement them for all patrons at their institution. TDM portals that are intended solely to control access to data sets should be removed, and librarians and vendors should push to implement a 5% removal solution—as English-Corpora.org has done—so that researchers can quickly obtain access to that content while the publishers do not fear losing copyright control of those data sets.

[4] For a good summary of exemptions under the DMCA, see *Medical Imaging & Technology Alliance v. The Library of Congress*. No.1:22-cv-499. The United States District Court for the District of Columbia. February 25, 2022. Sections 4–7.

[5] Set forth in Section 1201 Rulemaking: Eighth Triennial Proceeding to Determine Exemptions to the Prohibition on Circumvention. Recommendation of the Register of Copyrights. October 2021. See Proposed Class 7 (a) and 7(b): Motion Pictures and Literary Works—Text and Data Mining.

[6] As codified in 37 CFR Part 201.40(b) and 37 CFR Part 201.40(c). See *Federal Register*, Vol. 86, No. 206. October 28, 2021. Rules and Regulations.

In some ways, TDM licensing still seems like a Wild West, where too many vendors are taking too many approaches, causing too many librarians to have to figure out far too many different licensing options. It is not unlike the early days of e-journals. And like e-journals today, we need to get to a point where content providers share generally agreed-upon approaches to licensing and delivery of TDM resources.

REFERENCES

Building LLDTM. (2020). *Legal literacies for text data mining*. https://buildinglltdm.org

Center for Research Libraries (2014). *Liblicense: Licensing digital content*. http://liblicense.crl.edu/licensing-information/model-license/

Constellate. (n.d.). https://constellate.org

CORE. (n.d.). https://core.ac.uk

Corpusdata.org. (n.d.). *Full-text corpus data [limitations]*. https://www.corpusdata.org/limitations.asp

Courtney, K., Samberg, R., & Vollmer, T. (2020). Big data gets big help: Law and policy literacies for text data mining. *C&RL News*, *81*(4), 193–196. https://doi.org/10.5860/crln.81.4.193

Google Books Ngram Viewer. (n.d.). Google. https://books.google.com/ngrams

Google Dataset Search. (n.d.). Google. https://datasetsearch.research.google.com/

HathiTrust Digital Library. (n.d.) HathiTrust. https://hathitrust.org

LexisNexis. (n.d.). *Scholarly research with Nexis Data Lab*. https://www.lexisnexis.com/en-us/professional/academic/nexis-data-lab.page

Miller, H. K. (2015). *Securing text mining rights for researchers in academic libraries* [master's thesis, University of North Carolina at Chapel Hill]. https://doi.org/10.17615/v7p8-bv53

Okerson, A. (2013). Text and data mining: A librarian overview. IFLA World Library and Information Congress, August 17–23, 2013, Singapore. https://library.ifla.org/252/1/165-okerson-en.pdf

Orcutt, D. (2015). Library support for text and data mining. *Online Searcher*, *39*(3), 27–30.

Samberg, R. & Vollmer, T. (Eds.). (2021). *Building legal literacies for text data mining*. University of California, Berkeley.

Smith, C. (2022, May 2). Digging deeper: Text and data mining platforms for research libraries. American Libraries. https://americanlibrariesmagazine.org/2022/05/02/digging-deeper/

APPENDIX: MODEL LICENSING LANGUAGE

The Liblicense Model License, dating from 2014, proposes the following language regarding text and data mining. Given today's expanded focus on enhancing revenue streams, it's difficult to say how many vendors would be willing to use this wording.

Text and Data Mining. Authorized Users may use the Licensed Materials to perform and engage in text and/or data mining activities for academic research, scholarship, and other educational purposes, utilize and share the results of text and/or data mining in their scholarly work, and make the results available for use by others, so long as the purpose is not to create a product for use by third parties that would substitute for the Licensed Materials. Licensor will cooperate with Licensee and Authorized Users as reasonably necessary in making the Licensed Materials available in a manner and form most useful to the Authorized User. If Licensee or Authorized Users request the Licensor to deliver or otherwise prepare copies of the Licensed Materials for text and data mining purposes, any fees charged by Licensor shall be solely for preparing and delivering such copies on a time and materials basis (Center for Research Libraries, 2014).