Synthesis of cross-stakeholder perspectives on text data mining with use-limited data: Setting the stage for an IMLS National Forum[1]

Eleanor Dickson (Investigator), Megan Senseney (Co-PI), Beth Namachchivaya (Co-PI), Bertram Ludäscher (PI)

Abstract

This paper foregrounds discussion at “Text Mining with Limited-Access Data,” a National Forum funded by the Institute for Museum and Library Services. The forum will convene twenty-five stakeholders for a day-and-a-half long meeting to explore the current landscape for text data mining using texts that are under copyright or protected by licenses, including terms of use licenses, that limit access and use. The stakeholders include librarians, researchers, legal experts, content providers, and representatives of professional organizations. Prior to the forum, project team members interviewed stakeholders to learn more about their perspectives on research that employs use-limited text data, and to guide forum participants in drafting written statements and SWOT (Strength, Weakness, Opportunity, and Threat) analyses. This paper reports on the progress of the National Forum project, as well as synthesizes participant responses to the interviews and in their forum statements. It intends to establish a shared vocabulary and key points of consideration for the forum.

Introduction

Despite considerable attention given to data mining as a research method, and a rising flood of digitized and born-digital text, academic interest in text mining exceeds its current level of usage (JISC, 2012, Publishing Research Consortium, 2016). Some publishers and content providers have responded to the call to make their text content available as data by seeking to incorporate text mining into their platforms and service models.[2] Yet researchers who hope to incorporate text mining often find that despite the wealth of existing digital text, their use of the data is limited by intellectual property restrictions, licenses, or unfavorable access protocols. This situation has created a conundrum where, even when content-provider-created text mining services exist, researchers continue to point to the difficulties they encounter accessing and making use of text data as a key limiting factor to their text data mining (TDM) research (Green, 2017). Intellectual property and licensing restrictions can encourage an overreliance on certain open datasets, stymie researchers’ ability to incorporate use-restricted text data into their preferred computational research workflows, and limit communication of scholarly results and related methods transparently to a broad audience.

As the technological cost (Surden, 2013) of digitization and data-intensive research have shrunk, the law has begun to grapple with the issues surrounding digital collections, full-text indexing, and data mining. In the United States, where lawsuits over mass digitization efforts such as Google Books and

[1] This discussion paper, and the National Forum, scheduled to take place April 5-6, 2018 in Chicago, Illinois, are funded by the Institute of Museum and LIbrary Services (IMLS),IMLS LG-73-17-0070-17.

[2] Examples include the Elsevier Developer Portal, JSTOR’s Data for Research, Gale Cengage’s Gale Artemis: Primary Sources, and the HathiTrust Research Center.

HathiTrust have led to court decisions affirming the fair use of text for data mining, considerable uncertainty remains about the legal limits of permissible text mining. This lack of clarity is caused in part by terminological imprecision in the literature on text data mining, which promotes misunderstanding across communities of practice and introduces potential legal risks (Colonna, 2013). Legal uncertainty is further complicated by divergent national and supranational copyright regimes that run counter to academic norms, where communication and collaboration across national boundaries is common. In 2014, for example, the United Kingdom enacted a copyright exemption for non-commercial text mining when a researcher has lawful access to the text. Meanwhile, the European Union has no such exemption and adheres to a 1996 copyright protection for databases and their creators that limits data mining. Even when text is out of copyright, it may be subject to license agreements or hidden behind paywalls, which restrict its lawful access and use and make it difficult to discern what text data can be mined.

Additionally, scholarly text data mining exists within an entangled network of researchers, publishers, universities, scholarly societies and professional organizations, and librarians. Librarians in particular have found themselves as brokers of textual data, an extension of their long term role acquiring content and supporting scholarship. Yet most TDM services in libraries are currently limited to ad hoc access negotiation with content providers on researchers' behalf (Miller, 2015).[3] The full range of issues relating to TDM with use-limited text datasets is poorly understood, and the library community on the whole has yet to develop service models for supporting the many facets of text data mining (Orcutt, 2015; Schwarcz, 2017). When libraries are able to confirm the right to mine a textual resource, that does not mean that the data are discoverable, in a format researchers desire, or able to be analyzed due to lack of skill or technological infrastructure. As a result, the right to do text and data mining, particularly on use-limited text, may be scant more than a "theoretical right" for many who want to use this research method.[4]

Data mining with use-limited texts is thus subject to a series of tensions, which include strain between legal approaches and philosophies, roles and role reversals, and competing models for facilitating TDM within and across stakeholder groups. Combined, these issues complicate scholars' ability to take up, librarians' ability to support, and content providers' ability to develop means for this research method. This paper explores these tensions and the legal, social, and technical barriers surrounding text data mining with use-limited text, with a particular focus on the role of libraries in the United States.[5] It describes the landscape for academic TDM with rights-restricted and use-limited text datasets, and discusses the preliminary findings of a study that analyzes the perspectives of librarian, content provider, legal expert, and scholarly stakeholders in TDM research. This paper foregrounds the "National Forum on Text Mining with Limited-Access Data" funded by the Institute for Museum and Library Services, and the stakeholder perspectives it describes are those of forum attendees. It intends to establish a shared vocabulary for forum discussion, highlight key themes we have heard from

[3] This is not to overlook library initiatives that support both the pedagogical and the technical aspects of text data mining and analysis, such as 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/).
[4] We have adopted this framing of the "theoretical right" from a librarian participant of the National Forum.
[5] Portions of this paper were first presented at the 2018 International Digital Curation Conference (http://hdl.handle.net/2142/99026).

participants over the past six months, and outline strategies for engaging participants during the upcoming forum.

Methods

Efforts leading up to the National Forum began with a two-part research initiative: a literature review and semi-structured interviews with participating stakeholders. During summer 2017, efforts began with 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. The review was limited to works in English from 2000-2017. While primarily focused on the United States, the team also included scholarship that addressed other legal jurisdictions, including Canada, Australia, the United Kingdom, and the European Union. To ensure coverage across multiple disciplines, the team elected to conduct initial searches in prominent databases for Law, Library and Information Science, Computer Science, Linguistics, eScience, Digital Humanities, and Computational Social Science.

Prior to the initial review period, the team outlined criteria for inclusion and exclusion and agreed upon a selection of search terms to use in a range of combinations until results became uniformly redundant. For the purposes of this study, the team included any materials that focused on providing library services, developing computational workflows, and addressing issues related to data sharing. The scope, however, was limited to unstructured textual data with intellectual property considerations that took the form of copyright protections, data provisioned through licensing agreements, and web scraping online text accessed via a paywall and subject to terms of service. In choosing to focus on the legal aspects of intellectual property and text data mining, this review excluded scholarship on working with privacy-restricted data and the ethical implications of using data on the open web. Further exclusions filtered out text data mining focused solely on indexing for search and retrieval and a body of literature focused on patent analysis. An initial database search returned 103 results across seven categories, with 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. The literature review is ongoing through the conclusion of the grant project, and recommendations from forum participants are both welcome and encouraged.

Potential stakeholders were identified through the literature review and subsequent snowball sampling, and the final set of 25 forum participants includes representatives of professional societies; 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. Each participant agreed to prepare a forum statement and an analysis of Strengths, Weaknesses, Opportunities, and Threats (SWOT) prior to the event. Interview protocols were designed to assist participants in preparing their materials, while also providing an opportunity to speak extemporaneously and confidentially with the project team during the early phase of the project. Upon completion of all interviews, the project team reviewed notes and interview transcripts for prominent themes and then selectively coded each interview for a set of 26 thematic codes divided into six categories. Using the codebook, the team conducted a conventional qualitative content analysis of the transcribed interviews to identify key topics and establish cross-cutting themes and tensions identified by participants from across different stakeholder communities (Hsieh & Shannon, 2005).

Findings presented in this paper are based on analysis of coded transcripts as well as participants' forum statements and SWOT analyses. Using a similar approach to that used for the interview analysis, we performed an informal qualitative analysis of the SWOTs assigning labels to each comment in each SWOT according to the type of stakeholder (e.g., librarian, publisher, legal expert, professional society) and then aggregating these in a spreadsheet. We then "mined" the aggregate spreadsheet using a qualitative approach where we identified and coded each comment using one of six common themes that ran through the comments.[6] These materials were also used to inform the final agenda for the National Forum, which is discussed in detail below.

Literature Review

The literature review uncovered relatively few articles in science, humanities, and social science journals where authors explicitly reference rights issues in relation to text data mining research. Mentions in the literature were generally limited to meta-commentary, for example articles about open science or about reproducibility in the social sciences (e.g. Contreras and Reichman, 2015, Baiocchi, 2007), as well as those advocating for corpus creation (e.g. Garfinkel, et al., 2009) or text mining as a research method (e.g. Cohen and Hersh, 2005). We also found the occasional footnote or brief mention that the authors had used certain kinds of data due to intellectual property limitations, such as relying on abstracts instead of full articles (e.g. Li, et al., 2011), and descriptions of data sources tended to be most robust in digital humanities literature (e.g., Algee-Hewitt, et al., 2016). While the paucity of such articles discovered in our literature review may be the result of our search strategy, we believe the silence is telling. It may indicate, among other factors, a reliance on open datasets, an unwillingness to describe use of data that may be protected, or an unawareness of the implications for use-restricted data on downstream research.

In the legal literature, much of the discussion of text mining has focused on it as a research method made possible by mass digitization, where it is often presented as both a promise of and justification for mass digitization projects (Sag, 2012, Anderson, 2012). 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 (Jockers, et al., 2012, Amended Settlement Agreement, 2009). Nevertheless,

[6] Six themes that emerged from the SWOT comments included:
Business models: statements that focused on whether and how TDM can be made accessible, including early interest in exploring access models; models for monetizing access; models for access, and potential audiences for TDM, including scholars, business, and private citizens.
Content: comments that focused on identifying types of content (public domain, in-copyright), the value of content aggregation in research; the depth of coverage of digitized collections; and scholarly needs for access to content across publishing and distribution platforms.
Legal & policy: comments that focused on legal or policy aspects of TDM
Library roles: statements that referenced roles or perceived roles, perspectives, power, and responsibilities of libraries in providing access to and supporting TDM by researchers.
Publisher/content provider roles: comments that referenced roles or perceived roles, perspectives, power, and responsibilities of publishers and content providers.
Research process: statements that discussed the process, workflows, methods, impacts, and potential contributions of TDM to research
Technical: comments that referred to use of or need for technical expertise, information protocols and standards, APIs and technology services to facilitate TDM

the legal literature pays minimal 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. One 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 orientation toward practice 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.

Background

Text Data Mining

The terms 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. While several forum participants encouraged us to use more general terms, such as content mining, so as not to preclude computational analysis of video, audio, and images, in order to manage the scope of the forum, we have chosen to focus on *textual* data in particular.

Focusing on textual data allows us to raise questions about intellectual property and use-restrictions not applicable to numerical or tabular data, which, in the United States at least, are generally not copyrightable. Text, conversely, is nearly always protected by copyright from the moment it is fixed in a tangible medium. The concept of text as (digital) data complicates long-held copyright protections for literature and scholarly and news articles, and challenges traditional norms for access, resale, and use. Additionally, narrowing the scope of the National Forum to uses of textual data limited by copyright and licenses allows us to mostly sidestep important privacy and ethics issues. While these issues are more prevalent in sensitive, human-subjects data, several participants have indicated that these issues do impact published textual data as well. We limit the forum to intellectual property for purposes of scope and feasibility; however, we welcome discussions that better help us understand the intersection of privacy, intellectual property, and ethical research.

Use-Limited Data

The phrase use-limited data, which we employ throughout this paper, also runs the risk of being misunderstood or variously interpreted. As we have previously described, we are interested in textual data where use and access are limited, or potentially limited, due to copyright, licensing, and other contractual terms. Issues of researchers' data access, use, and reuse abound in data-driven scholarship, and so in order to manage the scope of the project, we initially sought to distinguish data subject to intellectual property restrictions from data that are restricted due to the ethical and privacy concerns surrounding human subjects by emphasizing difficulties related to data acquisition. Early project documents, and in fact the name of the funded project, 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. We have come to believe "use-limited" better describes the more restrictive facet of research with these data: how they may be used. This limitation encompasses a spectrum of activities ranging from modes of access to redistribution for validation and re-use.

In terms of copyright restrictions, 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 and 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 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 open 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.

Access Strategies for Use-Restricted Data

There are a variety of strategies for provisioning access to these use-restricted 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. Some content is available on the open web but nevertheless copyrighted or subject to terms of use that prohibit web scraping.[7] These terms are made more complicated by the fact that some 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) (Brehm and Lee, 2015).

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. 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).

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 point. 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 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

[7] One illustrative example is Rotten Tomatoes, an online movie review aggregations site, which is subject to Terms of Service that assert the content is under copyright and may only be used for personal and non-commercial uses (https://www.fandango.com/policies/terms-of-use).

and its fitness for use.

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 researchers throughout the TDM process and provides specific recommendations for dealing with texts protected by intellectual property rights. These services includes provisioning access 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 project builds on work such as FutureTDM in Europe and and the Australian Government's Productivity Commission Data Availability and Use inquiry. It is an early step in the United States at least 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 has convened 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.

Findings: Themes and Tensions

The complex legal and socio-technical environment in which researchers attempt text mining gives rise to a series of tensions around text data access and use. These tensions were apparent within and across participant interviews and written statements, and also within and across the stakeholder communities represented in this project. They relate to the competing legal approaches and philosophies for TDM with use-restricted text, the way roles for facilitating and sustaining research with these data are challenged and may reverse, and how labor is divided between librarian, content provider, and researcher stakeholders. In the following section, we describe these tensions and their impact on TDM with use-restricted text.

Legal approaches and philosophies

One of the most prominent points of divergence among forum participants is over legal approaches and philosophies surrounding TDM, most acutely between those who advocate for including TDM within the parameters of existing licenses or establishing a common-license mechanism, and those who prefer to rely on fair use justifications for text mining. For the last decade, libraries, responding to user requests for access to textual data for mining, have moved to integrate text mining clauses into their licenses. These clauses give explicit permission for scholars to engage in text mining research whereas earlier licenses were likely silent on the issue, and access to text data underlying the content was limited or unavailable. Based on our conversations with forum stakeholders, it is not certain whether uses of licensed content can be restricted to those expressly stated in the license, or if omission of a use in a license by default indicates its allowability. Librarians who promote the use of TDM licenses believe that they mitigate uncertainty and are likely to prevent situations where access to a resource is cut-off due to a researcher's attempt to text mine. Some librarian participants have shared and promoted standard model-licensing clauses for TDM that would make it easier for other libraries to adopt similar terms.

Not all participants are advocates for text mining licenses. Several participants acknowledge that lengthy license negotiation processes hamper productivity, and one researcher expressed frustration about the chains of communication required to begin the process at all. Participants also discussed the multiplication of effort that occurs when gaining access to many, discrete text datasets from different sources. Additionally, both librarians and data providers 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. Likewise the library may find it difficult to provision access to the data, which may be unwieldy, poorly described, or in an undesirable format. The license model may also encourage the kind of publisher-hosted TDM platforms that a number of stakeholder participants, including librarians and researchers, said were insufficient or subject to substantial privacy concerns for researchers.

Those who prefer to rely on fair use to justify TDM tend to be concerned that the increase in library-negotiated license agreements for text mining, particularly in the United States, compounds so-called "permissions culture," where users and license negotiators increasingly rely on license terms in situations of ambiguous permissions. They fear that such agreements run counter to legitimately fair uses, and also risk contracting away rights that would otherwise be assumed. One participant advocated that testing the limits of fair use was an opportunity for librarians and other stakeholders to bring clarity to the process, though participants more commonly considered the possibility of legal action a threat rather than an opportunity. Others raised the concern that only public non-profit institutions may be protected by fair use, and that private academic institutions may not technically be shielded by it.

Forum participants also spoke about potential rewards and risks of seeking legislative solutions to resolve TDM restrictions. Advocating for a text mining exemption could result in legislation that would formalize permitted uses and remove legal uncertainty that is likely stunting scholarship and scholarly communication. Pushing for legislative solutions could, however, ultimately result in more restrictive rules than are in place now, particularly in the United States: Many participants who mentioned legislation also noted that more permissive TDM laws are not a foregone conclusion of future modifications to rules for text mining. Some advised against creating opportunities for such a risky result, preferring to lean heavily on fair use justifications, common use licenses, or memoranda of understanding. In addition to risk as a downside to legislative reform, several participants noted it would be formidable to enact, requiring extensive lobbying and long-term commitment to the issue.

In the United States, the fair use argument for text mining frequently centers on the concept of non-consumptive research, which, though defined in the 2010 rejected settlement agreement in Authors Guild v. Google, in practice is more complicated than it first appears. In particular, the boundary between consumptive uses and non-consumptive research is under-developed, and the line between checking results, which is permissible according to the Google Books decision, and human reading is not bright. One legal expert cautioned that although they do not expect any "backsliding" on the core principle of non-consumptive use being fair, should the concept of "transformative use" fall into disfavor, this could lead in turn to a shrinking of the zone of permissible TDM, citing the TVEyes case.[8] Content providers, including publishers and digital libraries, have explored ways to facilitate TDM that provide varying levels of access to the data, from the hard drives of data and APIs that do not limit human reading to building platforms that bar any exposure to the text. Additionally, tensions arise for researchers around content for which downloading and human (consumptive) reading is allowed, but access to the text as data for machine (nonconsumptive) reading is disallowed. Some have taken up the phrase "the right to read is the right to mine" as a way of pushing back on the disparate permissions sometime ascribed to human reading versus computational mining.

Outside of legal concerns, text mining with use-restricted data is often at odds with the FAIR data principles, a best practice guide that stipulates data should be Findable, Accessible, Interoperable, and Reproducible and which was mentioned by a number of participants. When researchers cannot access the same data, cannot aggregate related data from across content providers, and cannot share their input data, some fear that it hinders sustainable, reproducible research. Researcher stakeholders reported feeling uncertain about what results they are permitted to share, and many respondents expressed concern over inaccessible data existing in non-compatible systems and formats. Sharing extracted data or derivative results and developing standardized access protocols, respectively, were offered as solutions to these issues. Overall, the scholarly impact of data-driven research that by its nature relies on data that run counter to the FAIR principles is yet to be fully explored.

Participants spoke of the chilling effect of use restrictions on TDM research. We noted a distinction between those who said use-restrictions stiffle research even before a project is underway as scholars avoid these limited data out of fear of legal repercussions, and those who said that researchers continue to do work with use-restricted data, but then do not openly communicate their methods and data sources. A further concern exists among researchers about their inability to support growing replication and transparency requirements if they use copyrighted or licensed data, particularly in the social sciences. These scenarios are troubling because of their implications, and likely have limited the noticeable impact of TDM research. At whatever point it occurs in the research process, the threat of legal action seems to have a strong effect on TDM, and participants discussed the fatigue and anxiety they felt as a result of wanting to mine use-restricted content.

Roles and role reversals

As libraries have moved to mediate licenses for text data mining, their role as research facilitators is increasingly in tension with their role monitoring copyright and license infringements. We found that librarian participants tended to feel as though negotiating text mining licenses and advocating for TDM was a way to remain involved in the research process, and that it was an extension of their role in content acquisition. Several researcher participants reported feeling as though their library was undermining their research, however, via the licenses and terms for text mining they had negotiated.

[8] See, for example, Katherine Trendacosta of the Electronic Frontier Foundation "Second Circuit Court Gouges TVEyes with Terrible Fair Use Ruling." 27 Feb. 2018. URL: https://www.eff.org/deeplinks/2018/02/second-circuit-gouges-tveyes-terrible-fair-use-ruling

More than one researcher participant voiced the frustration that librarians often appear as obstructing TDM by acting in what they perceive as the role of the "copyright police," responsible for enforcing publisher license requirements. One participant advocated for the library to take a more active role in promoting copyright literacy as a part of library instruction, with the goal of developing self-sufficient researchers who view the library as ally and not enforcer.

We found there tends to be unequal balance between how responsibility and authority are vested in the process of facilitating TDM with use-restricted text data. On any academic campus there are generally a range of stakeholders in TDM, data use and re-use, and content licensing, including diffuse expertise within the library, and researchers reported their struggle to find a point of contact for requesting and analyzing use-restricted data. And while digital scholarship librarians, who are likely most knowledgeable in their libraries about scholarly practices and preferences, may have a limited voice in data acquisition decisions, and they often lack agency to get data on terms and in formats they know scholars desire. Furthermore, the absence of clarity around allowable TDM also elicits feelings of concern from researchers and academic systems-developers who are responsible for fair use decisions, or who rely on students to make those decisions, and feel under-equipped to do so. The researchers responsible for making fair use determinations may resist asking for assistance from those with expertise in interpreting legal contracts, such as university counsel or librarians, who they perceive as risk averse and likely to block their uses.

Participants also expressed tension regarding the role of commercialization in text mining services. Some fear that, if they have not already, universities will lose ground to large corporations, such as Google, who will serve as data brokers for researchers instead of libraries. Others noted that publishers' interest in data mining extends beyond building TDM platforms and provisioning data access, but also to mining journal content for internal business purposes. This raises a question asked by some participants about who has become the data provider and who the data miner. And while libraries are also interested in building text mining applications to improve search and discovery, participants demonstrated the greatest anxiety over publisher-developed systems. This concern relates the the broad theme of author's rights and researcher privacy, which was apparent in multiple participant interviews. Considerations we heard ranged from the importance of authors' preferences for how their text and data are shared for mining to arguments against publisher-provided TDM platforms that collect usage data and services that mine researcher content, including articles and bibliographies, and sell profile data back to universities.

Models for Text Data Mining

Participant responses also highlight tension over the various models for facilitating academic TDM. There are differences of opinion among researchers, librarians, and content providers about the best way to provide access to use-restricted data. 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. As we have previously described, models for providing access to these data include moving it via hard drive or file transfer protocols, as well as models where researchers run analysis on a platform using off-the-shelf tools. There are costs and benefits to each model. For example, when data are moved from a publisher to local servers, it may come with a range of metadata and in various formats, and the receiver must find ways to build services for discovery and use. We found that even well-resourced universities struggle to provide access to content that has been delivered in such a manner. Nevertheless, participants indicated that researchers are less likely to be satisfied with platforms where only results and not input data are moved from provider to researcher, and they want the freedom and flexibility to control their analytic workflows afforded by locally-hosted data.

While standardized access protocols and the elimination of data silos were common refrains we heard from participants, there is strain also between data aggregation and disaggregation. 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. Still, respondents who work on data aggregation projects noted the difficulty they face maintaining up-to-date and standardized metadata. Additionally, some content providers who publish primarily aggregations assembled from various sources find it difficult to separate content and clear permissions for the streams of data they manage. They find that contributing institutions, which are often libraries, are concerned about the decontextualization of their content if it is provided via the publisher for TDM in a disaggregated form.

A number of participants presented varying viewpoints on the nascent and emerging business models for TDM. While some noted that licensed data sets are a source of economic viability, and therefore a way to extend a thriving publishing industry, a number of stakeholders voiced concerns that by monetizing access for mining purposes, publishers, especially the commercial sector (e.g., Pharma, life sciences, biotechnology and biomedical) could shape the arrangements for everybody, making TDM cost-prohibitive for most. There is a palpable tension, both within and across stakeholder groups, around the concept of including TDM as a value-added (and extra cost) option for libraries at the point of licensing. Among the current strengths in this landscape, participants identified projects that are committed to developing business and policy cases (e.g. EUH2020 and FutureTDM). They also noted that since TDM is widely used by “mega-corporations” like Google and IBM, this pushes the drive to extend TDM to other sectors and it facilitates the trickle-down of the technology. A number of participants across several stakeholder groups cited concern that high development costs would either discourage demand from researchers or be too high for publishers to manage. Some participants voiced concern that publishers would develop TDM as an added cost to journal and database licensing, a model that has already been embraced by some publishers and content providers.

Seeding Discussion: Toward a National Forum

To collectively and productively address the themes and tensions identified above, the National Forum will be organized into three segments, each with its own directive: listen and learn, seek collaborative opportunities, and make commitments. In lieu of conventional forum structures (e.g., presentations and open discussions), the team has adopted multiple strategies from the liberating structures menu to facilitate early engagement and encourage concrete outcomes within a relatively condensed period of time.[9] The first day is structured around a set of small group activities, each concluding with a W^3

[9] For more on liberating structures, see http://www.liberatingstructures.com/.

debrief in which participants reflect on the outcomes of the sessions to articulate what happened, why it mattered, and what actions should follow. This is colloquially referred to as “What? So what? Now what?”

Listen and Learn

After a brief introduction by the project team, the bulk of the morning will be dedicated to a fishbowl storytelling session. The room will be arranged with an inner circle of five chairs surrounded by a U-shaped conference table configuration. Participants will be organized into five groups of five by stakeholder affiliation with groups ordered from the most individual toward the institutional and then the collective: researchers, librarians, content providers, professional societies, and legal experts.

Each group in the inner will have ten minutes to discuss their perspective on challenges their stakeholder group encounters when text data mining and intellectual property intersect. Each 10-minutes discussion will then be followed by five minutes of Q&A with participants outside the circle. The directive for the morning is to listen and learn about each stakeholder group with the intention of better understanding the challenges different participants face within this context:

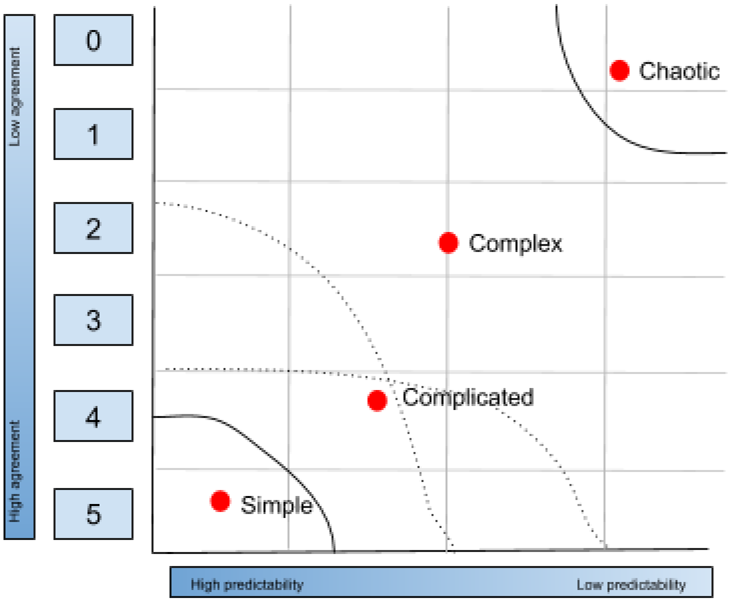
- **Researchers**: discuss the actual work of researchers conducting text data mining to identify bottlenecks and support requirements.
- **Librarians:** discuss current infrastructure for TDM services in academic libraries to understand the current roles that librarians assume in in facilitating TDM and how services might be improved.
- **Content Providers:** discuss your business model in relation to TDM with a focus on what potential viability and known deal breakers when working with subject to intellectual property.
- **Professional Societies:** discuss the degree to which TDM is an embedded practice within your community and what community members conducting TDM might need in terms of support and advocacy.
- **Legal Experts:** discuss the legal infrastructure for TDM to gain clarity on rights, options, and arguments that affect research with text data subject to copyright, licenses, contracts, technical protection measures, etc. when working within the US and with collaborators abroad.

Seek Collaborative Opportunity

Following a working lunch, new groups will convene around three thematic topics that represent different parts of a researcher’s workflow: finding and getting data, conducting analysis, and communicating results. This session begins a process of cross-stakeholder engagement intended to continue through the rest of the forum with the goal of seeking collaborative opportunities among members of the group. Drawing on themes emerging from initial analysis of interviews and forum statements, the facilitators have crafted a list of topics pertinent to each group. Because each group represents a snapshot of a continuous and cyclical process, the organizers anticipate that topics from one group may overlap with another.

The ***finding and getting data*** group anticipates thematic coverage that may include developing metadata and data remediation, understanding terms (both contractual and cross-disciplinary), advocating for fair use vs. licensing, articulating distinctive differences in data acquisition vs. hosted platforms, and/or strategizing workflows and processes for authorization. The ***conducting analysis*** group anticipates thematic coverage that may include outlining workflows and processes for conducting analysis, strategizing the development of data transfer standards and protocols, formalizing processes for research documentation, and/or confronting reproducibility issues in use-limited data contexts. The ***communicating results*** group anticipates thematic coverage that may include articulating strategies for peer review, planning for data reuse and the creation of derived datasets, publishing data and alternative documentation, understanding authors' copyright and licensing options, and crafting statements on privacy protections of content creators and data users. While discussions may include large-scale and long-term planning, participants will also be asked to identify small actions that can be taken immediately to trigger momentum within and across stakeholder communities. This will encourage initial action in light of current resources and circumstances, and so-called 15% solutions will be recorded as part of the W^3 debrief.

The third group session of the day is designed to quickly sort recommendations drawn from forum statements according to degrees of agreement and the predictability of the action's outcomes. Participants will be organized into five groups of five, and each group will be given an identical deck of index cards. Printed on each card is a potential action drawn from interview transcripts and forum statements. Within their groups, participants will first work together to determine a level of agreement for each action through a simple show of hands, sorting each card into one of six vertically arranged piles. Next, the team will move each card horizontally based on how certain or predictable the anticipate the outcomes of the proposed action to be. This will provide a bird-eye view of the complexity of the actions proposed.



Once the sorting activity is complete, participants will be asked to review the matrix and choose 2-5 priority items. Because the afternoon's directive is seeking collaborative opportunity, the project team recommends selecting from items that score high in agreement, but we are agnostic about whether they score high or low in terms of predictability. For each prioritized action, participants will be asked to identify the allies needed to make progress (either within or beyond the present group) and determine whether the action is a short-, medium- or long-term initiative. As time permits, attendees may then begin to outline discrete steps toward accomplishing the actions they prioritized.

After the closing discussions of day one, participants are invited to attend an evening reception, which will include a crowdsourcing exercise designed to seed the next morning's Birds of a Feather (BoF) topics. Each attendee will be given an index card and asked "If you were ten times bolder, what big idea would you recommend? What first step would you take to get started?". After writing their ideas on cards, participants will be asked to mingle and pass cards, but will be instructed not to read the until the bell sounds. At the sound of the bell, participants will read the card in their hand and mark it with a score from 1 ("I would absolutely not be interested in discussing this further at a BoF session") to 5 ("I would definitely like to participate in a BoF session on this idea"). Cards will be passed and rated four more times, each by a different attendee. After all five rounds, participants will be asked to tally the scores on the back of the card that they are currently holding. The top five high-scoring cards will then be selected for the next day's BoF session. Before the end of the reception, a team member will be assigned as note take for each group, and a participant will be selected as the session facilitator from among those participants whose initial ideas were not selected.

Make Commitments

The directive for day two shifts from seeking collaborative opportunities to making commitments. The day will kick off with attendee-driven BoF sessions where goals, strategies, and themes are fully determined by participants. The project team only asks that the final few minutes of each session are spent documenting concrete commitments related to the topic at hand. The BoF session will then be followed by an awareness building campaign. Alone or in groups, participants are asked to spend 45 minutes on an effort to expand the forum conversation beyond the group. Participants are encouraged to write a letter, draft an op-ed, create wikipedia entries for key topics, or outline abstracts for papers, panels, or funding proposals. Participants will be asked to complete their exercise after the forum and update the project team on their results.

The final two sessions of the forum will be devoted to plenary-style conversation and decision making. Participants will first provide feedback to shape an ACRL White Paper on Text Data Mining with Use-Limited Datasets, focusing explicitly on situating library action within a broader landscape. Participants will deliberate one what libraries need to know, what concrete actions libraries should take, what collaborative relationships libraries should foster, and what issues are pertinent but fall outside the auspices of library action. Following the white paper brainstorming session, participants will expand their focus to the full spectrum of potential stakeholders. Discussions will focus on tangible next steps, concrete commitments, and sustainable strategies for ongoing communication and engagement.

Conclusion

Text data mining and analysis methods hold strong potential to enable transformative and meaningful scholarly inquiry. Forum participants nearly unanimously mentioned the significant impact text mining has had and will continue to have on a variety of disciplines, from humanities to biomedicine. For this promise to be realized, though, clear, sustainable workflows for researchers who want to do TDM with use-restricted data need to be established. These changes are critical: The pervasive complexities and challenges surrounding access and use of textual data mean that softening the barriers that surround use-restricted data will likely result in overall improved research outcomes for scholars who seek to use data mining methods on any text.

Participants offered a number of solutions to improve the situation surrounding research with use-restricted data. The diversity of responses indicate that 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. Libraries are well positioned to facilitate text mining as part of digital scholarship and research data services, and their success in doing so depends on a coordinated effort with the relevant stakeholders. The National Forum will help 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. We would also like to thank Kate McDowell for her thoughtful advice on structuring the forum as well as Jana Diesner and Matt Jockers for initial feedback on research design protocols.

Bibliography

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).

Algee-Hewitt, Mark, Allison, Sarah, Gemma, Marissa, Heuser, Ryan, Moretti, Franco, & Walser, Hannah. (2016). Literary Lab Pamphlet 11: Canon/Archive. Large-scale Dynamics in the Literary Field. Retrieved from https://litlab.stanford.edu/pamphlets/

Anderson, I., Crews, K., Kaufman, R., & Maher, W. (2012). What Should Be the Conditions on Libraries Digitizing, Maintaining and Making Available Copyrighted Works. *Colum. JL & Arts*, 36, 587.

Baiocchi, G. (2007). Reproducible research in computational economics: guidelines, integrated approaches, and open source software. *Computational Economics, 30(1)*, 19–40. https://doi.org/10.1007/s10614-007-9084-4

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/

Brehm, Alison S., & Brehm, Lee, Cathy D. (2015). “Click Here to Accept the Terms of Service.” Communications Lawyer 31(1), Retrieved from: https://www.americanbar.org/publications/communications_lawyer/2015/january/click_here.html

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. Briefings in Bioinformatics, 6(1), 57–71.

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

Jorge L. Contreras and Jerome H. Reichman. (2015). Sharing by design: Data and decentralized commons. *Science*, *350(6266)*, 1310–1312. https://doi.org/10.1126/science.aad8071

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.

Garfinkel, S., Farrell, P., Roussev, V., & Dinolt, G. (2009). Bringing science to digital forensics with standardized forensic corpora. *Digital Investigation, 6, S2–S11*. https://doi.org/10.1016/j.diin.2009.06.016

Green, Harriett, Dickson, Eleanor, Nay, Leanne & Zegler-Poleska, Ewa. (2017). Scholarly Needs for Text Analysis Resources: A User Assessment Study for the HathiTrust Research Center. *Proceedings of the Charleston Library Conference*. http://dx.doi.org/10.5703/1288284316464

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

Li, Y., Hu, X., Lin, H., & Yang, Z. (2011). A Framework for Semisupervised Feature Generation and Its

Applications in Biomedical Literature Mining. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2), 294–307. https://doi.org/10.1109/TCBB.2010.99

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.

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

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

Publishing Research Consortium (2016). Text Mining
Of Journal Literature, 2016: Insights from researchers worldwide. Retrieved from http://publishingresearchconsortium.com/index.php/130-prc-projects/research-reports/text-mining-of-journal-literature-2016/170-text-mining-of-journal-literature-2016

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

Sag, Matthew. (2012). Orphan Works As Grist For The Data Mill. *Berkeley Tech. L.J., 27(3),* https://doi.org/10.15779/Z387M5B

Jockers, Matthew and Sag, Matthew and Schultz, Jason. (2012). Brief Of Digital Humanities And Law Scholars As Amici Curiae In Partial Support Of Defendants' Motion For Summary Judgment Or In The Alternative Summary Adjudication. http://dx.doi.org/10.2139/ssrn.2102542

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. H*arvard Journal of Law and Technology, 27(1),* 135-202.

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-consumptive use 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