TRANSFORMING LIBRARY SERVICES FOR COMPUTATIONAL RESEARCH WITH TEXT DATA

Environmental Scan, Stakeholder Perspectives, and Recommendations for Libraries

A Report from the IMLS National Forum on Data Mining Research Using In-Copyright and Limited-Access Text Datasets

Megan Senseney, University of Arizona
Eleanor Dickson Koehl, University of California, Los Angeles
Beth Sandore Namachchivaya, University of Waterloo
Bertram Ludäscher, University of Illinois at Urbana-Champaign

TRANSFORMING LIBRARY SERVICES FOR COMPUTATIONAL RESEARCH WITH TEXT DATA

Environmental Scan, Stakeholder Perspectives, and Recommendations for Libraries

A Report from the IMLS National Forum on Data Mining Research Using In-Copyright and Limited-Access Text Datasets

Megan Senseney, University of Arizona
Eleanor Dickson Koehl, University of California, Los Angeles
Beth Sandore Namachchivaya, University of Waterloo
Bertram Ludäscher, University of Illinois at Urbana-Champaign

Association of College and Research Libraries
A division of the American Library Association

Chicago, Illinois 2021

ACRL Advancing learning
Transforming scholarship
Association of College & Research Libraries
A division of the American Library Association

Citation: Association of College and Research Libraries. *Transforming Library Services for Computational Research with Text Data: Environmental Scan, Stakeholder Perspectives, and Recommendations for Libraries*. Megan Senseney, Eleanor Dickson Koehl, Beth Sandore Namachchivaya, and Bertram Ludäscher. Chicago: Association of College and Research Libraries, 2021.

Association of College & Research Libraries
A division of the American Library Association

Chicago, Illinois 2021

CONTENTS

EXECUTIVE SUMMARY

In April 2018, a group of twenty-five stakeholders selected from among researchers, librarians, legal experts, content providers, and representatives of scholarly and professional societies convened for a *National Forum on Data Mining Research Using In-Copyright and Limited-Access Text Datasets*. The initiative was funded by the Institute for Museum and Library Services to build a shared understanding of the issues and challenges associated with the legal and socio-technical logistics of conducting computational research with text data. This paper reports on preparatory activities leading up to the forum and its outcomes to (1) provide academic librarians with a set of recommendations for action and (2) establish a research agenda for the LIS community.

While responsibility for addressing the challenges of conducting text data mining (TDM) research with proprietary and IP-protected data does not fall solely on the shoulders of librarians, academic libraries have a key role to play in establishing a thriving scholarly ecosystem for TDM research. By working directly with researchers, communicating across units within the library, establishing campus-wide partnerships, and building coalitions with other external stakeholders, librarians can enact the recommendations outlined in this paper. In total, there are twenty-three recommendations organized along the six dimensions described below:

- **Fair use and licensing.** Where possible, avoid agreeing to license terms that limit the use of public domain materials or otherwise limit fair use, including TDM. The terms of institutional licenses should also be clearly communicated and shared with the community to which they apply, while individual licenses from scholars' data acquisition should be collected and stored to improve institutional memory and avoid duplicative effort.

- **Communication, outreach, and instruction.** Adopt a "collections as data" mind-set and facilitate the use of digital data through both formal and informal training and instruction.

- **Workforce development.** Establish text mining and legal literacies as core competencies for librarians working in the areas of digital scholarship and scholarly communication, create professional development opportunities for in-service professionals, and recruit information professionals with deep knowledge of TDM to work in academic libraries.

- **Research and governance.** Convene a campus-wide task force to address issues of data governance and risk management, establish institutional workflows for acquiring and using proprietary data, clarify the role of librarians as mediators and facilitators empowered to support TDM research, and document case studies of TDM research from data acquisition through analysis and dissemination.

- **Advocacy.** Work collaboratively with external stakeholders to develop a best practice guide for fair use in TDM, streamline scholar-initiated license negotiations; build awareness of TDM within scholarly and professional communities, and advocate for broad data access rights in matters of policy and legislation.

- **Infrastructure.** Participate in standards-making efforts to establish shared strategies for data interchange, establish partnerships to support large-scale data storage and high-performance computing (HPC) initiatives with the library's data, and explore opportunities for innovating data repositories to address the legal dimensions of data-intensive research along with its dissemination, preservation, and reuse.

INTRODUCTION

In an era of expectations that any and all data can be combined and mined, restrictions on the use of texts for data mining research may close off an area of inquiry before it even begins. Text data mining (TDM) includes the computational processes for employing statistical methods to discover new information and reveal patterns in unstructured text, and it often requires access to large amounts of machine-readable textual data. Legal agreements such as licenses and terms of use, intellectual property restrictions, socio-technical barriers, and economic limitations complicate the use of text data for mining, even at a moment when digital text abounds. These obstacles prompt scholars instead to rely on works that are free of copyright or licensing restrictions, resulting in analysis that excludes widely read or contemporary texts more truly representative of the scholarly or popular themes they had intended to examine—if they continue their research at all. On college and university campuses, librarians play a key role in facilitating access to the textual material researchers may intend to use as data by negotiating licensing agreements, purchasing content, and developing infrastructure and services for computationally intensive research.

The 2018 National Forum on Data Mining Research Using In-Copyright and Limited-Access Text Datasets, funded by the Institute for Museum and Library Services, brought together key stakeholders in conducting and facilitating TDM with use-limited data to grapple with these challenges and recommend solutions to ensure more successful outcomes for TDM researchers and the scholarly environments in which they work. During the forum, these twenty-five experts convened in Chicago for a day-and-a-half-long meeting. The forum project situated librarians as a key part of the diverse social network that encompasses TDM. The invited librarians represented a variety of specializations, along with affiliates of library-related organizations, all of whom hold a range of outlooks and priorities. Through activities and guided discussions, attendees of the forum moved toward next-step actions to improve the state of academic TDM.

Data mining methods have been in development for decades, but several factors have recently pushed TDM into the forefront as a key research tool across academic disciplines. The availability of significant digital text corpora (e.g., Google Books, HathiTrust, Internet Archive), coupled with accessible high-performance computing resources, have fueled significant momentum to use TDM as a research tool in the scholarly community. During the same period, research libraries have established digital collections by digitizing uniquely valuable scholarly content in their own collections; participating in initiatives to aggregate public domain and in-copyright content, such as HathiTrust and the Digital Public Library of America; and purchasing digital backfiles of journals, books, and other formerly print publications. Often these library-supported digital collections are offered to scholars through means unconducive to TDM, such as "page-turner" reading applications. At the same time, libraries have become ever more reliant on vendor-licensed digital content. The negotiated terms for some licensed materials may expressly make TDM unachievable by prohibiting bulk downloads and computational collection of data, and the technical infrastructure through which the materials are presented may discourage bulk access and use. As interest in TDM has grown, publishers' tendencies to present text datasets through systems that promote content as static, immutable, and contextually dependent has not yielded commensurately mineable digital textual data—all of this despite legal decisions affirming data mining as a fair use in the United States.[1]

To date, library and information science professionals' knowledge of TDM tools and practice has been largely limited to pockets of expertise among specialists working in concentrated ways with individual scholars and research institutes. Increasingly, librarians are involved in TDM workflows at the point where scholars ask them to arrange access to text data. Most TDM services in libraries are currently limited to ad hoc negotiation with content providers on researchers' behalf.[2] Other text mining initiatives in libraries have focused on training, such as the text mining and scholarly communications workflow at the University of California, Berkeley, or the Digging Deeper, Reaching Further initiative led by the University of Illinois, which has developed and disseminated a "train the trainer" curriculum on TDM for library and information professionals.[3] Librarians have also begun to develop guides and documentation that explain what text data are available to researchers and in what format.[4] Overall, the full range of issues relating to TDM is poorly understood, and the library community on the whole has yet to develop service models for supporting the many facets of this research.[5] Even when libraries successfully negotiate the right to mine a textual resource, that does not necessarily mean that the data are discoverable or in a format researchers desire; in addition, researchers may be unable to analyze the data due to lack of skill or technological infrastructure. As a result, the right to do TDM, particularly on use-limited text, may be scant more than a "theoretical right" for many scholars.[6]

As academic libraries integrate digital scholarship tools and methods into mainstream library services, there is a strategic opportunity to grow core skills in data mining, data analysis, and computational research methods within the LIS profession and in our service portfolios. Doing so would fuel innovative approaches to the ways in which research libraries make information available. In 2015, Yasmeen Shorish wrote of library research data services that "the absence of a holistic approach to data can result in the propensity to separate data from the corpus of information for which librarians already provide stewardship."[7] The same could be said now for the textual content scholars increasingly seek to mine as data, especially as the understanding of what it means for content to be accessible continues to shift from mere discoverability to flexible computational use. Projects such as Always Already Computational—Collections as Data encourage libraries to promote "computational use of digitized and born digital collections" in a way that would close the perceived gap between a library's collection and that which can be mined.[8] In the current environment, however, librarians working with scholars on text mining projects often find themselves cast in uncelebrated roles: beleaguered broker, bearer of bad news, or "copyright police."

Librarians are just one node in the tangled network of stakeholders grappling with these limitations, which also includes researchers, publishers, university administrators, professional organizations, and legal experts. In this environment, roles and responsibilities are not clearly delineated. Librarians have the opportunity to become partners, advocates, and leaders in facilitating data-intensive research, and doing so would require coordination across their campuses and within their libraries, as well as coordination with outside groups such as publishers and scholarly societies. As it stands, libraries' (and librarians') attitudes toward TDM may be perceived as reflecting their parent institutions' degree of legal risk aversion, hesitating to embrace TDM research as a right on behalf of their community and ceding rights to publishers via license agreements that restrict TDM and other forms of computational analysis. Within this network of stakeholders, individual librarians find themselves bound by competing expectations, restrictive agreements, and uncertainty over the rules and their ability to provide guidance on them.

This paper is based on the outcomes of the 2018 National Forum on Data Mining Research Using In-Copyright and Limited-Access Text Datasets. It describes the key takeaways of the event and the research project that accompanied it.[9] What follows are recommendations for academic libraries to integrate TDM practice into user-facing programs, technical and curatorial practice, and ongoing professional development.

DEFINITIONS

At the heart of the project are two concepts—text data mining and use-limited data—that are often underspecified or poorly defined. Below is a set of working definitions that team members developed and refined over the course of the project to articulate a shared understanding of the terms within the context of this project.

Text Data Mining

Text mining is the process of using computers to analyze and discover knowledge within text. 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 or an analytical approach.[10] Several forum participants encouraged the use of more general terms, such as *content mining*, so as not to preclude computational analysis of video, audio, and images, though other stakeholders believed the term *text data mining* is sufficiently broad to cover analysis of those other media. In order to manage the scope of the forum, we chose to focus on *textual* data in particular, and our use of the term related to that content specifically. Doing so allowed 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 allowed us to mostly sidestep important privacy and ethics issues. While these issues are more prevalent in sensitive human-subjects data, they do impact published textual data as well. Nevertheless, we limited the forum to intellectual property for purposes of scope and feasibility.

From Limited-Access to Use-Limited Data

We use the term *use-limited data* to refer to textual data where use and access are limited, or potentially limited, due to copyright, licensing, or other contractual terms. Throughout the course of the project, our terminology shifted from *limited-access* to *use-limited* in order to better capture the nature of the restrictions researchers are likely to face. In the early stages of our research, we 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. This limitation encompasses a spectrum of activities ranging from modes of access to redistribution for validation and reuse. Other terms considered by the project team include *proprietary data*, which was determined to be narrower than our working definition of *use-limited data* and more likely yield an emphasis on data-as-object rather than the role of these data within a more comprehensive research process. Emphasizing the way textual data are incorporated into research expands the conversation beyond data acquisition to include consideration of how the law impacts analytic workflows and reproducibility of TDM results while observing use limitations related to redistribution. As described above, we exclude data that are restricted due to the ethical and privacy concerns surrounding human subjects.

In terms of intellectual property restrictions in the United States, 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 either exceeded their copyright period, been released into the public domain by their creator, or been created under conditions such that they are born into the public domain (e.g., federal government documents). Because these works fall outside the protections of copyright, many scholars presume that they are unrestricted for TDM purposes. Within the United States, however, contracts may supersede questions of copyright, and it is common to enter into contractual agreements as a condition for accessing texts that have been digitized, organized, or otherwise maintained by third parties and licensed to intermediaries. Where licensing agreements are silent on the question of TDM, researchers are often unsure what activities are allowable. In-copyright works considered within the framework of this study include those that are in copyright but openly licensed for use and redistribution via schemes such as Creative Commons, texts digitized from a legally acquired print copy, digital copies of texts that have been lawfully purchased but include technical protection measures, texts that sit behind a paywall for which access has been licensed, and in-copyright text that is freely available on the open web but subject to terms of use, such as conditions that may apply to the use of a social media platform.

PROJECT DESIGN

Data Mining Research Using In-Copyright and Limited-Access Text Datasets was conducted from July 1, 2017, to December 31, 2018, and it was funded as a national forum under the IMLS National Leadership Grants program. This effort was led by a team of coinvestigators (Bertram Ludäscher, Beth Namachchivaya, Megan Senseney, and Eleanor Dickson Koehl) in consultation with a ten-member local advisory committee comprising experts brought together from across the University of Illinois at Urbana-Champaign (see appendix A). The forum was designed to engage a group of key stakeholders and scholars to explore broadly applicable approaches that support computational research with in-copyright and use-limited data and to make recommendations for best practices and policy to guide libraries as they develop TDM services (see appendix B). The overarching goal of the project was to articulate a research agenda for the LIS community and to provide an action framework for libraries to facilitate research access, implement best practices, and mitigate issues associated with methods, practice, policy, security, curation, reproducibility, and replicability in research that incorporates a broad spectrum of text datasets.

The project was composed of four discrete phases: (1) conducting a literature review and environmental scan, (2) recruiting and interviewing participants, (3) planning and conducting the national forum, and (4) analyzing forum outcomes. Each phase informed the development of subsequent phases; detailed methods for each phase are documented in appendix C. The literature review was conducted systematically and informed the initial identification of forum participants, complemented with subsequent snowball sampling. During fall 2017, the project team conducted semi-structured interviews with each participant to incubate forum statement ideas and model the process of completing a SWOT (strength, weakness, opportunity, and threat) analysis. Forum planning was grounded in the desire to elicit perspectives from diverse stakeholders, motivate them to interact, identify common ground, and work collaboratively to make a difference in TDM research capabilities. The project team organized the forum around three key directives: listen and learn, seek collaborative opportunity, and make commitments. Through a series of structured activities adapted from Liberating Structures,[11] we sought to reveal participants' perspectives and create opportunities for collaboration, even in the absence of consensus. Throughout the project, the team utilized approaches drawn from qualitative analysis to synthesize a large body of textual materials, including interview transcripts, participants' forum statements and SWOT analyses, and detailed notes taken over the course of the forum. Results from qualitative analysis informed the forum agenda, shaped the pre-forum discussion paper, and are presented below as a set of recommendations for academic libraries.

LITERATURE REVIEW

Background

Scholarly publishing on data mining has been on the rise since 1996, with growth increasing rapidly through the first decade of the twenty-first century.[12] With growing interest in data-intensive computational work, the number of tools and platforms to support TDM research have also proliferated. Jovic, Brkic, and Bogunovic have developed a useful matrix-style comparison of six free tools: RapidMiner, R, Weka, KNIME, Orange, and scikit-learn,[13] and other open source initiatives such as ContentMine and Voyant Tools continue to diversify the field.[14] Meanwhile a range of content providers (e.g., Elsevier, Gale, JSTOR, and HathiTrust) have begun developing a variety of services including APIs, web-based portals, computing environments, and other forms of support for analyzing their content as data. Acknowledging the rise of TDM as a valuable (and increasingly viable) research strategy, several library associations have released statements on TDM over the past five years.[15] While this project focuses explicitly on text data, the digital objects with which scholars work are increasingly complex and heterogeneous, including structured datasets, audiovisual materials, and multimodal environments, suggesting the need to adapt socio-technical strategies for dealing with more complex datasets.[16] In response, some scholars have shifted in favor of the term *content mining* over *TDM* to encompass this variety.[17]

Following the rise of TDM, discussions of legal issues related to TDM gained traction within the past decade, with a notable increase in scholarly publications around 2012. The graph in figure 1 illustrates this trend by mapping the publication dates of the eighty-nine articles included in this literature review. These dates correspond with two watershed events in the English-speaking world that called attention to questions of copyright and TDM: (1) a set of lawsuits initiated by the Authors' Guild against Google, Inc., and HathiTrust and (2) the release of a UK-based report on opportunity costs and the potential economic impact of copyright law in the digital era.

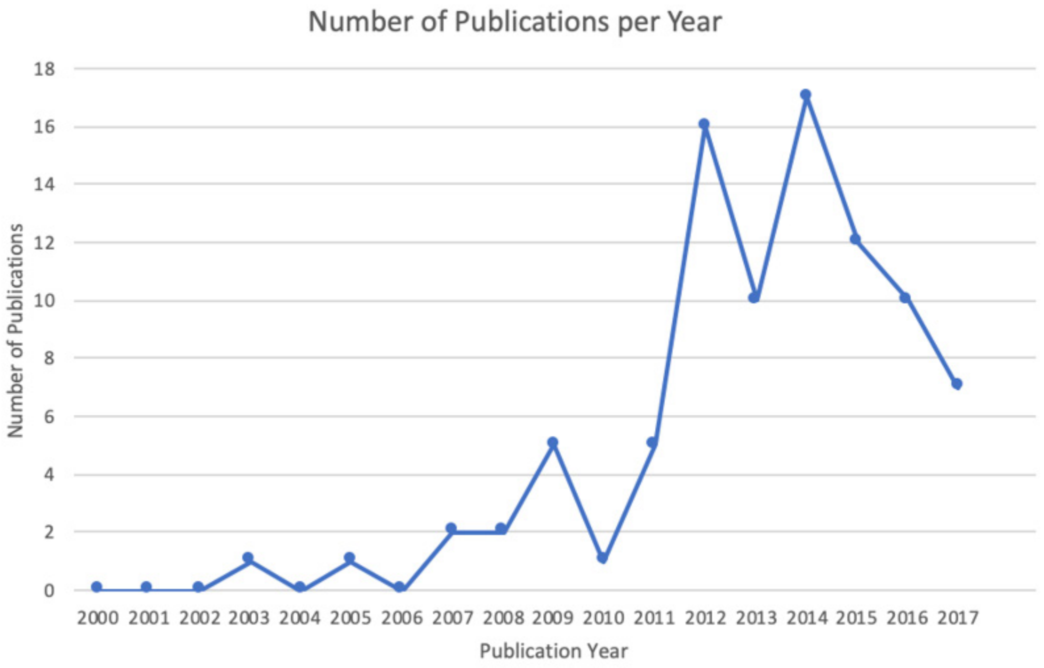

Figure 1. Number of items represented in literature review by year of publication

Authors' Guild Lawsuits

From 2005 to 2015, the Authors' Guild was engaged in a set of lawsuits brought against Google and HathiTrust for copyright infringement in response to an initiative to digitize millions of books. Working at scale, Google established a set of partners, including a number of academic libraries, to systematically scan copies of print books, convert the images to text using optical character recognition, and store and index the resulting corpus. The HathiTrust Digital Library, launched in 2008, is a consortial repository of the books that were sourced primarily from research libraries and digitized by Google. The legal decisions hinged on whether the activities of Google and HathiTrust met the legal requirements for fair use, and the high-profile cases sparked a set of legal reflections on digitization, derivative work, orphan works, and non-display uses.[18] In an amicus brief that directly addressed the question of copy-reliant technologies and TDM, Jockers, Sag, and Schultz argued that scanning works constitutes a non-expressive use and that text mining is a process of information extraction in which aspects of the original text become factual metadata about the text.[19] The Second Circuit ultimately decided in favor of Google and HathiTrust, deeming the uses "highly transformative" and thus fair.[20]

Hargreaves Report and Subsequent Copyright Reform

A major watershed moment in Europe was the release of *Digital Opportunity: A Review of Intellectual Property and Growth*, commissioned by Prime Minister David Cameron and released in 2011 by Ian Hargreaves. It concluded that an outdated intellectual property system in the United Kingdom would lead to negative economic impacts by stifling digital innovation. Among other recommendations, Hargreaves proposed making licensing easier, creating a specific and mandatory exception for TDM, and significantly reforming copyright for the digital age.[21] In 2014, UK researchers were granted the right to conduct data mining for noncommercial purposes on any text a user otherwise has access to, but the reform also introduced a set of new complications: namely, that it did not address the challenges of gaining permission and it remains ambiguous about what constitutes noncommercial use.[22] Following the Hargreaves report, there was also a concerted push for the European Commission to adopt a text mining exemption.[23]

Legal Precedents and Ongoing Uncertainty

Copyright and the Doctrine of Fair Use

In the legal literature, much of the discussion of text mining in the United States 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.[24] Concepts such as "non-consumptive" and "non-expressive" use emerged from cases where US courts ruled in favor of text and data mining as fair uses of digital libraries.[25] Notably, non-expressive use is just one of several kinds of potentially transformative uses (along with productive, orthogonal, and creative uses) that may determine whether a given use is fair.[26] While some have criticized the so-called "transformation of transformative use,"[27] the recent trend favoring fair use interpretations in court findings may empower librarians to consider a bolder approach toward asserting fair use when developing service models, providing patron consultations, or developing institutional risk management strategies.[28] Fair use is a flexible standard adjudicated through a four-factor framework that focuses on the purpose of the use, the reasonableness of the use in light of that purpose, and the likely effect of the use on relevant copyright markets. Courts also focus on whether and to what extent a given use of a work is transformative, or whether the new work created is sufficiently different in "context, meaning, or message" when determining whether the new use is fair. The flexibility of fair use is commonly regarded as one of its virtues in a world of fast-changing technology, but some scholars argue that such flexibility may have a chilling effect for risk-averse educational institutions.[29] Recent attempts to create bright-line rules have, however, been dismissed in court.[30]

Contract Law, CFAA, DMCA, and Other Legal Considerations

The relationship between copyright and contract law complicates the implementation of recent fair use determinations. Fair use presumes prior access, but access to desired texts often hinges on the terms and conditions of a license agreement, which are governed by contract law.[31] At present, contracts supersede copyright, creating conditions for access that may restrict uses otherwise deemed lawful and appropriate.[32] Further complicating the legal uncertainty surrounding TDM are unresolved questions regarding the potential application of the Computer Fraud and Abuse Act (CFAA) to terms of service violations related to excessive downloading and the Digital Millennium Copyright Act's (DMCA) criminalization of technologies that circumvent technical protection measures implemented as part of digital rights management strategies.[33] Finally, the imprecise use of the term *data mining* in legal writing and policy creates legal risk when terms are not agreed upon; a taxonomy of data mining and its relationship to other subfields has been proposed to help clarify technical terminology for the legal community, which may, in turn, reduce risks for technologists.[34] Ultimately, the combined effects of the current legal environment stand to hamper twenty-first-century research and innovation.[35] Acknowledging this risk, scholars have called for reforms that ensure accepted uses of in-copyright works cannot possibly be overridden by contracts,[36] and others have advocated for shifting to open access publishing with permissive licensing schemes, encouraging authors to move away from large commercial publishers.[37] Without further intervention, greater adoption of open access scholarly publishing would certainly benefit many communities, but it would not resolve copyright and contractual issues related to text mining previously published works or nonscholarly content (e.g., commercial publications).

International Context

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. US-based scholars must consider the international context for TDM when working with collaborators located in other legal jurisdictions or seeking content from providers based outside the United States. Issues surrounding moral rights, database law, and fair dealing may create further legal difficulties in designing and executing an internationally scoped TDM project.[38]

While an in-depth discussion of the full range of variation across international legal jurisdictions is outside the scope of this paper, many of the articles reviewed by the project team addressed the legal environment for TDM in Europe, Canada, and Australia and compare fair dealing (found in many common law jurisdictions) with the United States' decidedly more flexible concept of fair use. Despite the UK's adoption of a new copyright exception for computational research in 2014,[39] many scholars and librarians remain uncertain and hesitant to exercise their rights, leading to calls for librarians to "be bold about the advice they give to researchers."[40] To date, the European Union has no such exemption and adheres to a 1996 copyright protection for databases and their creators that limits data mining, but debate about a proposed TDM exception in the European Union is active and ongoing, with a copyright harmonization effort known as the Directive on Copyright in the Digital Single Market under discussion through the end of 2018. In developing potential policy frameworks for an EU copyright exception, members of the Future TDM project have focused their efforts on articulating technology-neutral strategies for framing the nature and purpose of reproduction, focusing instead on language clarifying that any case where reproductions are made for the sole purpose of extracting unprotected ideas from a work should not fall within the scope of the author's exclusive right.[41] Some scholars have also expressed concerns about the restricted personal scope of the proposal and the emphasis on public research institutions as well as concerns about technical protection measures that might go against the spirit of the exception and make mining impracticable for many.[42]

Scholarship on Access, Use, and Dissemination

When scholars publish the results of their TDM research, they are often silent on the processes by which they obtained their underlying corpora and vague regarding details of their methodology that might expose them or their organizations to legal liability. This ambiguity is likely related to the legal uncertainty in which they operate but undermines community values expressed through validation and reproducibility. Considering these issues at a level of remove, it is also worth noting how copyright shapes the collecting policies of large-scale digital library initiatives, such as the Biodiversity Heritage Library, which may serve as vital sources for text data.[43] Similarly, contractual considerations shape initiatives like the Public Library of Science (PLOS), which actively encourages text and data mining across the entirety of its collections.[44] Occasionally, scholars directly address how data access issues impose inherent or potential limitations upon their studies, modeling a practice that other scholars would do well to adopt.[45]

Within the body of literature identified for this review, articles have tended to focus on questions of access as opposed to potential nuances related to use or later dissemination. In figure 2, the frequency with which articles are tagged as discussing access issues is denoted in orange. Notably, the most recent articles in this review include more frequent discussions of use (gray), and several serve as exemplary case studies (green), suggesting a welcome trend toward more soup-to-nuts analyses of how legal issues affect research in TDM. This section follows a framework of access, use, and dissemination to discuss contributions from disciplinary and library-based literature on the topic of TDM with use-limited data.

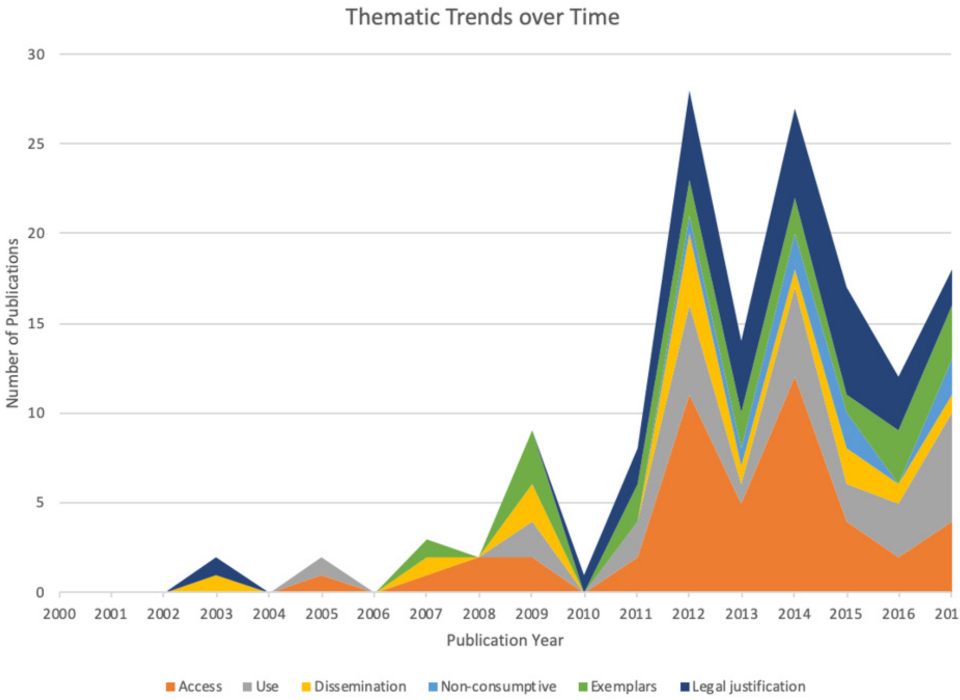

Figure 2. Thematic trends identified in the literature over time

Access

Lack of access to texts due to licensing and copyright materially shapes the corpora gathered and subsequently used by scholars. While negotiating access to proprietary data for individual research projects can severely impact work plans and result in significant delays, contractual agreements should be specified and settled prior to the start of any TDM project.[46] In a panel discussion with scholars and librarians, one content vendor suggested that scholars improve their success with negotiating access by modeling data requests that "describe how the content would be used, the scope of the content being requested, and who will be involved in the research project."[47] In response to the complications inherent to securing permissions, some scholars have advocated for the development of a clearinghouse that provides a single point of access for both researchers and publishers.[48] To that end, both Crossref and the Copyright Clearance Center have made strides in aggregating existing rights metadata but do not participate in brokering additional rights and permissions.[49]

Despite the fact that many publishers ultimately allow some form of text analysis for content in subscription databases, further challenges arise when provisions require that a researcher work within a given publisher's confined environment rather than gaining direct access to the data through bulk download or another transfer protocol.[50] Often, researchers must combine data from multiple sources to satisfy their research question, rendering platform-based research inadequate and compounding opportunity costs as researchers approach multiple publishers to build a corpus.[51] Some wonder whether such complex licensing permissions and access controls on the part of the content provider ultimately serves to hamper legitimate uses among scholars who are careful to comply with license agreements without thwarting uses that violate terms.[52]

For researchers, web crawling is a common mode of gaining access to content on the open web, often governed by a robots exclusion protocol in which websites communicate with crawlers about any content that is disallowed and may also include expectations meant to throttle use through crawl delays. Neylon addressed the concerns that publishers have raised about researchers whose content crawling of their sites has purportedly put undue stress on publisher servers by suggesting that a combination of social and technical approaches can be employed to reasonably manage increased use.[53] In a discussion of the legal and ethical considerations concerning web scraping for research purposes, Black modeled a scraping protocol that encourages researchers to directly communicate their intentions by writing tools with user-agent strings that include "their name and the address to a web page that explains the goals of their research, details their web scraping procedures, and provides contact information for feedback about their procedures or for requests to be excluded from the project."[54] When researchers apply these techniques to subscription databases provisioned through their institutions, however, they risk violating contractual agreements that were negotiated on their behalf. Content providers may respond by cutting off access to the entire institution, and librarians are responsible for responding to license breaches from systematic downloading that the publisher deems excessive.[55] Indeed, librarians often first learn about TDM activity on campus through notifications of unusual behavior or outright database shutdowns.[56]

Some scholars may not realize that librarians are responsible for negotiating access to key scholarly databases such as LexisNexis or JSTOR, and many librarians have already elected to incorporate provisions for TDM through licensing. The scope of library licensing efforts is significant; in 2014, for example, Stanford University managed about 600 different contracts for licensed content.[57] Individual licenses may take as long as a year to negotiate and finalize, and some libraries have developed licensing checklists to streamline the process across multiple negotiations over long periods of time.[58] Specific terms for TDM are often negotiated case by case, and librarians make a good faith effort to ensure that terms are favorable for research in practice and that the data formats are amenable to TDM.[59] The California Digital Library, the Center for Research Libraries, and the Canadian Research Knowledge Network have led the way in developing and disseminating model licensing clauses for TDM in an effort to improve both efficiency and consistency.[60] Based on lessons learned from past experiences with overly restrictive journal licensing agreements, the Association of Research Libraries has developed similar language for e-books, suggesting that authorized users can do anything "consistent with exceptions and limitations of copyright."[61] Perhaps most importantly, while the licenses that librarians negotiate on scholars' behalf may disallow uses that are otherwise deemed fair, there is little evidence that librarians are clearly communicating the terms of these licenses back to scholars.[62] Indeed, functional and subject librarians—who work closely with patrons but do no participate in license negotiation—are often unclear on the specifics of these terms and commonly express general uncertainty regarding TDM as a fair use.[63]

Use

Even when access to data is secured, significant barriers to use remain. As scholars combine data gathered from multiple sources, difficulties arise due to a range of technological barriers, from the lack of standardized data formats across publishers to the lack of appropriate personnel and infrastructure for data transfer.[64] Reflecting on problems with standardization, Reilly argued that librarians are well situated to use their leverage with content providers to push for structuring content such that it is amenable for TDM.[65] Confronted with the realities of potential challenges related to library staffing, Orcutt outlined models for basic, moderate, and intensive library engagement with TDM research.[66] The latter two models include services beyond basic license negotiation and access, where moderate support focuses on basic training and instruction for novice researchers and intensive support adopts a high-touch collaborative approach in individual TDM projects.

In an example of intensive library engagement, Vanderbilt University established a model premised on cooperation between liaison librarians and functional specialists.[67] At Vanderbilt, the librarians' primary focus remains on licensing, acquisitions, and curatorial concerns, but a formal partnership with researchers is established via a memorandum of understanding that provides scope for the TDM project and articulates the extent of library support and engagement throughout the research process. At the University of Toronto, curatorial support has taken the form of corpus scoping, data normalization, and deduplication, and Dyas-Correia and Alexopoulos have noted how new training opportunities emerge around curatorial issues related to tool selection and customization.[68] Documentation also arises as a persistent theme among participants in Digging into Data, a grant program sponsored by several leading research funders from around the world to address how "big data" changes the research landscape for the humanities and social sciences. In particular, participants reflected on the growing importance of articulating the interconnection of the research question, data, and algorithm as well as the need to demystify black-box algorithmic strategies.[69] While some exemplars have emerged in the past few years that detail the process of securing access to text data and supporting TDM researchers,[70] there remains a great need for further publication of real-world, soup-to-nuts use cases that follow the data through to analysis and dissemination.[71]

NON-CONSUMPTIVE PARADIGM

In some cases, scholars can use text data without ever gaining full access. One of the provisions in the amended settlement agreement between the American Association of Publishers, the Authors' Guild, and Google was the creation of a research corpus for a range of computational research methods, including TDM. While the settlement was ultimately rejected, it planted the seed for the subsequent creation of the HathiTrust Research Center (HTRC).[72] Since 2011, HTRC has played a vital role in exploring research strategies that align with non-consumptive (or non-expressive) use, which, according to HTRC's Non-consumptive Use Research Policy, means "research in which computational analysis is performed on one or more volumes (textual or image objects) in the HT collection, but not research in which a researcher reads or displays substantial portions of an in-copyright or rights-restricted volume to understand the expressive content presented within that volume."[73] HTRC has experimented with several approaches for allowing computational analysis against the collections gathered by the HathiTrust Digital Library.[74] Currently, non-consumptive research can be conducted against works in copyright in one of three ways: using HTRC's web-accessible data analysis and visualization tools, downloading derived datasets of so-called "extracted features" (which constitute factual information about the corpus), and working directly in a secure research environment called the HTRC Data Capsule. A data capsule is a virtual computer where researchers can perform computational analysis on the texts in the HTRC corpus using their own algorithms and incorporating other resources. Security is managed through a combination of policy and technical constraints. Notably, the model of the "data capsule" also bears resemblance to recent information retrieval systems that have emerged to allow computation against privacy-restricted or industry datasets,[75] suggesting the potential for more widespread adoption.

Consistent with findings from Digging Into Data, a recent HTRC case study on conducting research in the data capsule environment highlighted the importance of clearly documenting the research process and discussed possibilities for automating documentation through data provenance.[76] Explicit documentation is vital for providing coherent, interpretable exports and communicating findings to the broader community. This documentation can be achieved by capturing and describing (1) information about the data selected for analysis; (2) computational workflows, including inputs, transformations, and outputs; and (3) the final results, which are packaged as "non-consumptive exports." As the open science community continues to lead the movement toward sharing data, code, and execution environments along with published papers in an effort to increase transparency,[77] scholars working with use-limited data will need to develop adequate alternatives that are likely to take the form of community norms for process documentation.

Dissemination

Across multiple domains, researchers have expressed concerns about how authors choose to license their scholarly content to publishers and the public, which may affect potential reuse for TDM. For example, even when authors retain copyright, many publishers seek exclusive commercial reuse rights as part of their standard agreement with authors, which may enable publishers to exert more control over text mining.[78] Open licensing schemes such as Creative Commons have developed simple tools for more flexible licensing, but the popular adoption of noncommercial licensing is also becoming difficult to navigate for university researchers who are engaged in public-private partnerships or who benefit from industry funding.[79] Such concerns led Hrynaszkiewicz and Cockerill to propose novel combinations of a Creative Commons Attribution license and Creative Commons CC0 waiver for scholarly publications.[80]

Publishing the results of TDM projects that analyze use-limited data also raises questions about appropriate strategies for data sharing. Copyright generally prohibits—and contractual terms tend to disallow—republishing extracts of text data on the open web, which creates difficulties in supporting trends toward data sharing and calls for greater reproducibility in data-intensive research.[81] In a groundbreaking paper on culturomics, Michel and colleagues described the restrictions on providing unmediated access to an underlying corpus of in-copyright texts and chose instead to distribute a dataset derived from the in-copyright texts they used.[82] The data took the form of a frequency table that counted the occurrences of single words, successfully meeting the data-sharing requirements of *Nature* while respecting copyright restrictions. Sharing innovative forms of derivative data has since become more common among individual scholars,[83] and it is one of the non-consumptive strategies adopted by the HathiTrust Research Center, which generates large-scale extracted features datasets from the HathiTrust corpus as a service to researchers.[84]

Libraries certainly have a role to play in facilitating data archiving for scholarly communication, research stewardship, and preservation.[85] While the early library literature on TDM focused on needs assessment and license negotiation,[86] recent discussions have shifted toward digital pedagogy and data curation.[87] Comprehensive library service models will take a life-cycle approach to TDM, providing support that includes training and instruction for use along with long-term curation to enable discovery and reuse of TDM corpora and accompanying outputs. Making recommendations based on challenges encountered during the first round of Digging into Data awards, Williford and Henry advocated for more "explicit, long-term agreements" that offset the burden of investment required to ensure the necessary resources, skills, and services for computationally intensive research and permit data sharing across institutional boundaries.[88] Still others have begun advocating for the development of knowledge commons with shared rights infrastructures featuring a "common data access and use policy."[89] By creating conditions of possibility for data sharing, librarians can, in turn, improve access for future generations.

FORUM STAKEHOLDERS' PERSPECTIVES

Throughout the course of the national forum project, the cohort of participating stakeholders had multiple opportunities to define and clarify their viewpoints on the issues surrounding TDM with use-restricted text. Prior to the forum, we synthesized their written statements, SWOT analyses, and interviews into a position paper that drew out areas of tension that had emerged during our analysis and established baseline perspectives and expectations across stakeholder groups.[90] As the attendees represented a microcosm of the identified stakeholder groups (researchers, librarians, legal experts, content providers, and professional organizations), putting their comments and experiences in juxtaposition with one another afforded us the opportunity to identify the degrees of consensus or polarization within and across participant groups on several aspects of TDM, including legal, technical, financial, and material concerns. We shared the paper, along with the written statements and SWOT analyses, with attendees in order to seed discussion at the forum. We used the outcomes of our pre-forum analysis also to inform the development of the forum program and activities (see appendix B). During the course of the program, certain ideas from our earlier analysis emerged as recurrent topics, while others were scarcely discussed. Following the forum, we analyzed our written notes and other by-products of forum activities to explore the most salient topics of conversation, agreed-upon actions, and areas of disagreement.

Going into the forum, attendees understood that it is difficult to reduce challenges and find mutually beneficial solutions for the issues surrounding research with use-limited data. Their perspectives addressed many competing concerns, such as the open access movement, as well as principles such as FAIR data, which stipulates that data should be findable, accessible, interoperable, and reproducible.[91] They also noted the way in which potential actions could draw national, or international, government-level attention through laws, treaties, or judicial decisions that would address, among other concerns, international data standards and copyright harmonization across state boundaries. Where there was pre-forum consensus on key points from most of the groups (for example, that current US copyright law has not evolved in pace with computational analysis in the digital age), there were differences of opinion within and across stakeholder groups about how to remediate such issues. Through the course of the forum activities, we found that different stakeholder groups came closer to consensus on certain themes than they had been prior to the forum.

In particular, the first forum activity—called a "fishbowl conversation"—set the tone for the event (see appendix C).[92] The issues raised in the fishbowl established topics that persisted throughout the forum. They focused attention toward library services and the library-researcher relationship and away from solutions reliant on content providers or on governmental action. In one notable example, a researcher pressed a librarian about whether he could in fact access a textual dataset the librarian had claimed his library made available; the librarian demurred.

Taken together, the continuum of forum activities demonstrates the way in which several of the tensions highlighted in our pre-forum discussion paper were areas of evolving thought through the course of the project, while others were points of entrenchment left unresolved by the end of the forum. In the following section, we highlight stakeholders' changed and unchanged viewpoints by describing the salient pre-forum tensions and resultant forum discussions.

Evolving Perspectives

1a. Pre-forum Tension: Library Licenses for TDM

While the existing literature in librarianship tends to focus on negotiating licenses to establish TDM access for textual data, in their pre-forum statements, stakeholders from the librarian and legal expert groups initially disagreed over whether it was preferable to (1) advocate for including TDM within the parameters of existing licenses, (2) establish a common-license mechanism, or (3) rely on fair use justifications for text mining, especially as the licenses may apply to public domain content. Some participants believed that licensing for TDM is useful in mitigating uncertainty, while others expressed concern that these licenses compound "permissions culture," where users and license negotiators increasingly rely on license terms in situations where the right to perform an activity is ambiguous. They argued that such agreements run counter to legitimate fair uses and also risk contracting away rights that would otherwise be assumed.

1b. Forum Discussion: Text Data Discovery and Access

Access to content and library licenses for TDM were a prevalent topic of conversation at the forum. The researcher and librarian stakeholder groups related to one another with perhaps the most initial discord, and the most resultant empathy, as they sought to find common ground despite researchers notably faulting librarians for the difficulties they have faced in accessing textual data. Researchers continued to emphasize that they are weary of being told to wait patiently for their library to negotiate access, while librarians described feeling stuck between researchers' wants and what they are able to accomplish with regard to TDM licenses and infrastructure. Librarians in attendance expressed the desire to develop sustainable TDM services that go beyond case-by-case access negotiation.

Opinions in the room trended toward frustration with the licensing model. Researchers and librarians began to ally themselves as the researchers recognized that librarians may not feel supported by faculty to push back against license terms unfavorable to TDM. Many participants began to favor promoting and relying on "fair use over licensing," and one legal expert who works in a library committed to encouraging researchers to rely on fair use when conducting TDM. Incorporating the clause "Notwithstanding the foregoing, nothing in these license terms restricts fair uses of this content" was suggested for licenses that include TDM.[93]

Forum discussion drew attention to the shifting expectations of what access means for library-held content, which has trended from availability, to discoverability, to flexibility over time. After hearing the researchers' frustration about not knowing what data their libraries make available, several librarian participants committed to making text data resources discoverable and services findable. While some supported the idea of a data clearinghouse, others noted that researchers are unlikely to search the library catalog for text datasets—which suggests ongoing uncertainty about the best way to convey what data are available, in what formats, and for whom.

Some saw preprint repositories or open access provisions as solutions that would move textual data away from the current paradigm for licensing content, especially with regard to text data from journals or other scholarly publications. Still, attendees acknowledged that open access does not provide a solution for commercially produced, non-scholarly, or historical data, and some expressed concern that tying TDM to the open access movement would deter publisher support for the method. One group of participants from across stakeholder groups recommended that libraries seek to regain control of mineable text by focusing on unencumbered resources and ceasing to invest in resources that cannot be used for computational research.

2a. Pre-forum Tension: Conflicting Obligations

Participants' pre-forum statements suggested that as libraries are increasingly involved in mediating licenses for TDM, their role as research facilitators is in tension with their obligation to monitor copyright and license infringements among users. Librarian participants perceived license negotiation and advocacy for TDM as a way to support the research process, seeing it as an extension of their role in content acquisition. Nevertheless several researcher participants reported feeling as though their library was undermining their research via the licenses and terms for text mining they had negotiated. More than one researcher participant voiced frustration that librarians often appear to obstruct TDM by acting in what the researchers perceive as the role of the "copyright police," responsible for enforcing publisher license requirements. Conversely, one librarian participant advocated for the library to cultivate a perception of itself as ally and not enforcer by taking a more active role in promoting copyright literacy as a part of library instruction.

Forum attendees' statements also emphasized the unequal balance between how responsibility and authority are vested in the process of facilitating TDM with use-limited text data. Researchers reported their struggle to find a local point of contact for requesting and analyzing use-limited data. Digital scholarship librarians, who likely have the most knowledge in their libraries about scholarly practices and preferences for TDM, oftentimes are not deeply engaged with library licensing activities and may not be experienced with negotiating for data acquisition.

These librarians may lack agency to get data on terms and in formats they know scholars desire. Furthermore, researchers are both uncomfortable with performing and unprepared to perform their own fair use analyses that would enable their TDM work. Researchers may also 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.

2b. Forum Discussion: Shared Responsibilities

While in their pre-forum statements, stakeholders made suggestions as to who should be primarily responsible for spearheading efforts to reform the issues surrounding use-limited data, at the forum it became clear that there are cross-stakeholder common goals that point to shared responsibility. For example, one attendee wrote in his statement that researchers should take the lead on advocating for changes because their understanding of the issues and how those issues relate to their research agendas would be necessary for gaining traction. Then, during the forum, researcher and librarian participants found mutual support for the challenges presented by use-limited data. Nevertheless, lingering questions remained about the best path forward. Responsibility for advancing solutions is still complex, and solutions involve the actions and decisions of multiple groups. It is important to note, however, that among the forum participants, including the content providers present, no one rejected the idea of text mining access for use-limited data, and all seemed willing to seek next steps to improve outcomes for TDM research.

With regard to library responsibilities specifically, attendees largely agreed that libraries should make it easier for researchers to find and mine textual data. The proposal to "improve library TDM services" was received with broad consensus, although there were a range of ideas about how such services could be successfully implemented. These proposals included librarians providing guidance and assistance for data preparation and normalization; creating shareable, derived datasets; and providing access to content through APIs, potentially through the use of a collective data standard for mineable text data. Attendees who proposed strategies for improving library services generally agreed that libraries need to build capacity to assist with data-driven research by prioritizing in-library expertise. One librarian participant suggested a holistic shift in library services such that that libraries would "reprioritize data-driven research" and make data use and reuse an integral part of a library's service model.

3a. Pre-forum Tension: Uptake of TDM Methods

Whether there has been adequate uptake of TDM research methods to justify development of service models that would support it was a concern frequently expressed by content providers in their pre-forum statements. Indeed, research studies have shown that interest in TDM has outpaced its uptake by scholars.[94] In their statements, researcher and librarian participants described the chilling effect of use restrictions on TDM research, and our literature review revealed relatively scant references to rights or use restrictions related to datasets in papers where TDM was a research method. Researcher and librarian participants were divided between those who said use restrictions stifle research 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-limited data but then do not openly communicate their methods and data sources. At whatever point it occurs in the research process, the threat of legal action seems to have a strong effect on TDM, and researcher participants discussed the fatigue and anxiety they felt as a result of wanting to mine use-limited content. Prior to the forum, we had identified this incongruous framing of the problem as one of uptake versus one of discouragement, and the stories stakeholders shared with one another in the forum assisted in demystifying perspectives across their respective groups.

3b. Forum Discussion: Advocacy and Storytelling

Forum attendees largely agreed that improved communications about TDM with use-limited data—messaging that both described challenges and demonstrated success—could lead to better outcomes for this area of research. One small group proposed developing an international policy statement—a "Chicago Statement," described below—that would build awareness about the challenges faced by text data miners. Participants described the need to communicate these issues not just to the public to garner support, but also to content providers, who may not fully understand the extent of the limitations their policies have imposed. Attendees also felt it is important to build support for TDM by sharing use cases and success stories that would demonstrate the need to invest in infrastructure and services for data-driven research. One stakeholder suggested a practical phased approach to first generate grassroots support before advocating at a national or supranational level in order to develop proofs-of-concept and use cases to support a larger advocacy initiative.

4a. Pre-forum Tension: Ambiguity over Allowable Activities

One commonly shared point in statements from librarians, researchers, and others was persistent confusion over what is and is not allowed with regard to TDM and use-limited data. Researchers expressed their frustration over content for which downloading and human (consumptive) reading is allowed, but access to the text as data for machine (non-consumptive) reading is disallowed, whether due to contractual agreements or to the technical limitations of content platforms. They demonstrated shared alignment with the phrase "the right to read is the right to mine." Legal experts, on the other hand, tended to view the current environment as much less ambiguous than their fellow participants believed: They described the current legal precedent in the United States as a navigable and knowable, if imperfect, framework with various ways by which researchers can accomplish their goals.

Still, pre-forum statements from legal experts and content providers did draw attention to areas of uncertainty. Legal experts suggested in their statements that the boundary between consumptive uses and non-consumptive research in particular is underdeveloped, and the line between checking results and human reading is not bright. For example, Google Books held that the display of three-line snippets to add context to book search results was transformative in purpose and that it was reasonable in proportion to that purpose.[95] Those snippets allowed a user to verify that a book suggested by the search engine was in fact relevant to her interests. In addition, the snippets were so brief that they did not pose any risk of fulfilling the reader's demand for the original expression of the underlying manuscripts. The practice was thus found to be fair use. In Google Books, three-line snippets were shown to users of the book search engine to provide context for the search results. The length of the snippets and the limits on how they could be combined made it clear that this was the predominant function the snippets served. The Google Books case provided some, but nevertheless limited, guidance on the use of snippets from the underlying text. While one participant advocated that testing the limits of fair use was an opportunity for librarians and other stakeholders to bring clarity to the process, attendees more commonly considered the possibility of legal action a threat. During the forum, librarian and researcher stakeholders were eager to learn more from the legal experts in the room, and for many it seemed these conversations provided valuable clarity.

4b. Forum Discussion: Establishing Clarity around TDM Research

During the forum, there was wide support for creating better "norms and guidelines" for TDM that would provide clarity for the widespread confusion surrounding TDM. Participants committed to creating improved written documentation and guidelines in the form of blogs, LibGuides, and publications that could be presented as a workflow for decision-making. The idea to create a best practice guide for TDM similar to others that already exist, such as ARL's "Code of Best Practices in Fair Use,"[96] was one of the most widely agreed-upon solutions coming out of the agreement-predictability matrix activity. Several attendees suggested that such guidelines could bolster the legal defense of TDM as a permissible act under fair use in the United States by reinforcing community norms and standards.

Another strategy to alleviate confusion around TDM with use-limited data focused on training opportunities. From the perspective of some legal expert stakeholders, one of the biggest challenges facing TDM research is that scholars are unsure what their rights are. Participants recognized that the need for training extends to researchers as well as librarians. Several attendees committed to training-specific actions, including developing instruction to build literacies and confidence in researchers engaged in TDM, as well as crafting an educational road show. While participants from across stakeholder groups favored training initiatives, librarian attendees were most likely to commit to training-related actions.

One area where the legal-expert stakeholders provided clarity for forum attendees was around the allowability of sharing derived datasets, which they asserted was a permissible activity. With that encouragement, there was relatively broad support for mechanisms by which scholars could make available derived datasets created during the course of their research. Researcher stakeholders expressed interest in data repositories and data journals where datasets and valuable context about their creation, respectively, could be deposited. The idea to create derived datasets, which obscure the original expression of the full text, was acknowledged as an opportunity for both researchers and libraries to sidestep intellectual property protections that might otherwise prevent data sharing.

Entrenched Perspectives

Whereas the thematic areas presented above represented topics where stakeholders' opinions shifted from before the forum to after it, those described below were points of disagreement throughout the course of the forum project. While attendees continued to discuss and offer new ideas about services and business models for TDM with use-limited data, these themes were persistently divisive.

1. How to Best Establish TDM Services

Forum participants' statements highlighted tension over the various models for facilitating TDM, and opinions remained relatively polarized through the forum. There were differences of opinion among researchers, librarians, and content providers about the best way to provide access to use-limited data. As we have previously described, models for providing access to these data include moving them 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, they may come with a range of metadata and in various formats, and the recipient 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, and content providers worry about the security of data in this scenario. Participants also indicated that researchers are less likely to be satisfied with platform-based solutions where only results and not input data are moved from provider to researcher because they want the freedom and flexibility afforded by locally hosted data to control their analytic workflows. This concern was shared among content providers, researchers, and librarians. Even within stakeholder groups, attendees disagreed on whether the best next steps are to seek actions that meet majority of researchers' TDM needs (referred to as "80% solutions") or to focus on the minority of high-end researchers ("20% solutions"), assuming that the research community will eventually catch up with those now on the cutting edge. By the end of the forum, stakeholders seemed no closer to finding consensus for this open question.

Participants from across stakeholder groups expressed concern also over the lack of standards for facilitating TDM. They noted a gap in 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. While stakeholders brainstormed ideas for how to ameliorate these issues, we did not see any proposal generate substantial buy-in from the rest of the participants.

2. Business and Funding Models

During the forum, a number of participants presented varying viewpoints on the nascent and emerging business models for TDM. While some noted that licensed datasets 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., pharmaceuticals, life sciences, biotechnology and biomedical), could shape the arrangements for everybody, making TDM cost-prohibitive for most. There was 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. Content providers tended to express concern over the viability of TDM services and the cost required to build services and provide text as data. Still, several stakeholders cautioned against the risks to reproducibility as well as to equity if only scholars at well-resourced institutions could afford to engage in TDM. A number of participants across stakeholder groups cited concern that high development costs would either discourage demand from researchers or be more costly than content providers would be willing to bear. Among the related strengths identified in this area, participants mentioned projects that are committed to developing business and policy cases (e.g., EUH2020 and FutureTDM).[97]

Participants also expressed tension regarding commercialization of text mining services. Some shared fear that, if they have not already, universities will lose ground to large corporations, such as Google, which 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. Similarly, some participants warned against publisher-provided TDM platforms that collect usage data and services that mine researcher content, including articles and bibliographies, to then sell profile data back to universities. 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 to the theme of researcher privacy, which was apparent in multiple participant interviews, as well as in our literature review. While the topic of funding and business models was addressed at the forum, the diversity of strongly held positions on the costs associated with TDM led to few points of consensus among stakeholders.

FORUM ACTIONS AND NEXT STEPS

On the second day of the forum, participants self-selected into four birds-of-a-feather groups that engaged in deeper conversation: Practical Library Recommendations, Low-Hanging Fruit, Chicago Statement on Content Mining, and Legal Infrastructures. These groups were intended to build momentum around areas of shared interest. In the following section, we describe the outcomes of each birds-of-a-feather group and any resulting post-forum actions they have taken.

Practical Library Recommendations

The Practical Library Recommendations group consisted primarily of librarians who were interested in further synthesizing the outcomes of two forum activities—agreement-certainty matrix and 10× bolder ranking process—in order to make recommendations for attainable library actions. Their recommendations include

1. Develop guidelines for appropriate computational uses of textual resources, including copyrighted and licensed materials.
2. Build capacity in library staff to support researchers in TDM activities.
3. Support researchers by offering preprocessing services for TDM activities.
4. Collaborate with the institution and external service providers to offer storage, computing, and publishing support for TDM research.
5. Engage in technical standardization for transfer protocols between content providers, APIs, publishing services, and preservation.

The group recognized that some of their recommendations warrant further consideration. For example, they grappled with how much vetting and how controversial a best practices guidelines document would be. They also noted that building capacity to support TDM from the library could take the form of re-skilling through professional development, strategic hiring, or both.

Low-Hanging Fruit

The Low-Hanging Fruit group was largely made up of practitioners, some of whom had preexisting relationships from activities not directly tied to the national forum. They included content providers, librarians, and others who are engaged in providing access for TDM. They discussed services and initiatives to facilitate

TDM with a particular focus on business models for TDM. Several of the themes they discussed included the need to move forward despite recognizing that there is no solution to satisfy all parties, publishers' concerns about security risks involved in moving data, and the existing market for value-added TDM services. They discussed models for content-provider-supported TDM they have seen in practice. Moreover, two content providers in this group deepened their commitment to join a planning effort to provide secure access to in-copyright content for TDM.

Post-forum work: Several members of this group presented at the 2018 Charleston Conference in a "lively discussion" session,[98] and they are pursuing other ongoing collaborations. Representatives from two content provider organizations have participated in independent follow-up meetings to plan a project that would explore providing scholars with secure access to in-copyright content for TDM research.

Chicago Statement on Content Mining

This Chicago Statement on Content Mining group proposed writing a statement modeled on the Hague Declaration on Knowledge Discovery In the Digital Age.[99] The group comprised researchers, legal experts, and representatives of professional organizations. During the forum, members of the group began drafting their Chicago Statement on Content Mining. By the end of the forum, they had a working draft that they committed to completing in the coming months. Their draft included recommendations for libraries and data centers, several of which include

1. Move away from gatekeeping responsibilities and foster an environment favorable to TDM.
2. Be transparent about which datasets are mineable and under what conditions.
3. Licenses should contain a provision on TDM.
4. Maintain and sustain data in open, mineable formats.
5. University-vendor contracts must not allow third-party vendors to monitor library users' behavior.

Post-forum work: A draft version of the Chicago Statement is under review and slated for future release.

Legal Infrastructures

Members of the Legal Infrastructures group consisted of researcher, librarian, and legal expert stakeholders who were concerned with legal issues, not limited just to copyright, and how they intersect with teaching and learning. During the forum, they discussed opportunities for legal advocacy to raise awareness of the issues faced when doing TDM. They also talked about the need to train researchers and librarians about their rights and the risks for TDM with use-limited data. They identified several post-forum activities:

- Case studies from scholars about their pain points in TDM research to inform instruction and facilitate advocacy
- A best practices guide, its audience, and how to best approach drafting this kind of document
- A follow-on grant proposal for training on TDM's legal concerns, how to structure such a learning opportunity, and who should be involved

Post-forum work: This group's post-forum work involved communication with related groups, including the National Association of College and University Attorneys (NACUA), Authors Alliance, and other law clinics, and biweekly meetings to discuss follow-on activities, including grant proposals and journal articles.[100]

RECOMMENDATIONS

In an increasingly digital scholarly communication landscape, collections of textual data are both the output and the object of research, and TDM has emerged as a powerful method for scholars across disciplines to conduct analysis. While some national forum attendees expressed considerable uncertainty about permissible activities and what their libraries can do or can direct researchers to do, legal experts in the room articulated far more certainty about exercising fair use. Undoubtedly gray areas remain, but what we know is allowable is (1) able to be communicated and (2) worth universities "owning" in order to exercise the rights they are afforded. For these reasons, we believe the key ideas that surfaced in the forum and through our research are the creation of a best practices guide, as well as efforts that would empower and encourage educational institutions to make use of their TDM rights. Legal uncertainty and institutional uncertainty will remain as long as librarians, scholars, and their institutions hesitate to build robust digital scholarship practice with data mining and analysis as a key component of the digital scholarship life cycle.

Drawing upon our analysis of the literature, semi-structured interviews, participant forum statements and SWOT analyses, and documentation captured from the IMLS National Forum on Data Mining Research Using In-Copyright and Limited-Access Text Datasets, the project team has identified the following recommendations that library and information science professionals should consider as they integrate TDM and a collections-as-data mind-set into mainstream library programs. Situating TDM as a core research method within the larger digital scholarship services landscape bears significant promise for integrating this practice into mainstream scholarship.

Within the Library

Fair Use and Licensing

1. Where licenses are silent on TDM, have librarians educate users to exercise fair use.
2. Develop an internal strategy and set of guidelines for negotiating more favorable terms for existing licenses that contain restrictive language for TDM to ensure fair uses are preserved.
3. Centralize and share licensing agreements with the broader campus community in a secure, password-protected environment; clearly communicate terms of agreements to students, faculty, and staff.
4. Expand copyright and licensing services to help authors understand individual licensing agreements for TDM access, and establish a hub of institutional memory regarding these agreements to reduce duplicate effort in the future.
5. Establish and sustain communication channels and workflows across acquisitions, licensing, and e-resources librarians and library professionals who work directly with scholars to resolve questions of access to use-limited text datasets.

Communication, Outreach, and Instruction

6. Adopt a "collections as data" mind-set, and develop strategies for making library collections more computationally amenable (e.g., by developing library-based APIs for library-held content).
7. Integrate TDM and accompanying legal literacies into librarian-led information sessions and instructional workshops.
8. Develop partnerships with faculty, departments, and research centers that support scholars in building competency in the areas of TDM and analysis.
9. Partner with faculty seeking to develop TDM modules for graduate and undergraduate courses.

10. Facilitate informal learning by hosting librarian office hours, research groups, or coding sessions.

Workforce Development

11. Establish facility with both text mining and related legal literacies as a core professional competency for librarians working in digital scholarship, research data management, scholarly communication, and cognate specialties.[101]
12. Create opportunities for practicing librarians and other information professionals to build TDM skills into their professional portfolios through training, coursework, or praxis.
13. Recruit information professionals who have deep skill sets in TDM and related digital scholarship tools and methods (e.g., predictive modeling and data visualization) to work in academic libraries.

The Library in Collaboration with Other Stakeholders

Research and Governance

14. Convene a campus-level task force to address data governance and risk management (including issues around TDM) that brings together campus administrators, legal counsel, technologists, librarians, and faculty representatives.
15. Clarify the role librarians play as mediators between research and content providers with faculty, administration, and legal counsel to ensure that librarians are sufficiently empowered as advocates and negotiators.
16. Collaborate with university counsel and data governance groups to establish institutional procedures for securing TDM access to use-limited data as a means to keep from duplicating effort. Designate an institutional point of contact, and create internal documentation about prior individual and institutional engagement on campus.
17. Develop a suite of user stories by collaborating with faculty on individual case studies that document the entire TDM research process, from data acquisition through use and dissemination.

Advocacy

18. Collaborate with professional organizations to commission a best practices guide for text data mining (see, for example, fair use, software preservation).
19. Partner with researchers and their scholarly professional organizations to reach existing audiences and to cultivate new audience awareness of TDM and its use in research.
20. Collaborate with scholars, legal counsel, and technology security services to develop templated guidelines for scholars negotiating access to stand-alone datasets for TDM.
21. Identify issues, policy, and legislation specific to TDM, and work with professional organizations to advocate for broad access rights to use digital texts.

Infrastructure

22. Work with content providers and standards-making bodies (e.g., W3C) to formalize data format standards and transfer protocols to improve interoperability.
23. Partner with scholars, academic departments, research groups, and high-performance computing organizations to identify computing resources and environments that can facilitate mounting standalone datasets and TDM computational analysis, including storage.

24. Participate in large-scale efforts to establish innovative repositories, such as a TDM data clearinghouse or knowledge commons.
25. Support scholars' use of data journals and accompanying repositories focused on derived data and methods papers for TDM with use-limited data.[102] Advocate with scholars to make the changes in data submission guidelines that support reproducibility of TDM research with use-limited datasets.

FUTURE RESEARCH

The National Forum on Data Mining Research Using In-Copyright and Limited-Access Text Datasets set out to generate both recommendations and a next-phase research agenda for the LIS community around TDM with use-limited text. How can libraries support and partner with scholars to use TDM effectively to deepen scholarship, developing and implementing methods and programs that expand the spectrum of research using text collections as data? We recommend a straightforward and collaborative approach: Observe researchers in practice and ask direct questions about their access and content needs, develop case studies that document both successes and failures, and share those with scholars to develop solutions together.

Below we propose a research agenda that would refine our findings and bring clarity to questions raised during our study by focusing on the limitations, needs, and behaviors of scholars who are or would like to be engaged in TDM research.

1. **Conduct further user studies to learn more about scholarly needs.** What data formats, tools, or systems do researchers require in order to conduct TDM? Where should libraries and content providers focus their attention when developing infrastructure and capacity for TDM?
2. **Research provenance, workflows, and related socio-technical infrastructures.** What workflows do TDM researchers follow, where do they get their data, and how do they interact with existing socio-technical systems for accessing, analyzing, and sharing data? Where does TDM with use-limited data intersect with ongoing initiatives to improve reproducibility for data-driven research?[103]
3. **Measure TDM success and failure rates.** How many researchers have given up on a TDM project because of data issues, or how many have done work they have not reported out of legal concern? What is the cost, material or otherwise, to scholarship when scholars are unable to complete their research as planned or spend significant time navigating data access rules and restrictions?
4. **Understand the role of training LIS students.** As programs in data science mature, there is increased overlap with the skills required by data scientists to manage, analyze, and create metadata and derived data for datasets. What are library and information school students learning about TDM, data manipulation, and statistics? How does their training complement what graduate students in other disciplines are learning, and in what ways could LIS professional knowledge support other areas of expertise on academic campuses?

CONCLUSION

The LIS community, working with researchers, content providers, the legal community, and scholarly societies, has the potential to make substantial inroads to establishing TDM practice as one of the fundamentals of digital scholarship, one that is part of a robust data life cycle encompassing effective discovery, access, preservation, and reuse. While during the course of the national forum project the critical role libraries can and will be able to play in regard to data-driven scholarship was clear, reducing the challenges libraries face developing services for TDM will require collaboration with the academic community writ large. While universities may be centers of cutting-edge research, they also tend to be risk-averse on an institutional level. Within this institutional context, those who wield the most power are commercial publishers that have the leverage to push for restrictive licensing agreements and university administrations that seek to avoid risking legal challenges. Librarians have ended up with limited agency and as easily blamed figures for the challenges their local scholars encounter when attempting TDM research. Solutions to these issues necessitate buy-in from both the top down—at the level of professional organizations and university administration—and the grassroots—at the level of researchers and academic professionals. Such a holistic, systems-level approach is required in particular for improving outcomes for TDM due to the way that computational text analysis with use-limited data touches on a myriad of hot-button concerns, from the open access movement and profit-driven publishing industry, to the reproducibility crisis and ethics concerns over research involving commercial data, to campus data governance and privacy.

Promoting computational research and enacting a collections-as-data mind-set in libraries will not be easy, not least because it requires coordination between digital scholarship, technical services, and acquisitions librarians in potentially new and frictive ways. The stakes are high for libraries if we continue to invest in digital collections but neglect the transformative possibilities they enable in data-driven research, both for the relevancy of the library and also for its return on investment. During the forum, one participant framed the challenge for libraries as moving beyond discovery and delivery of content to making content "useful." While those may be vexing words to librarians who work in constrained environments to make collections available to scholars, they speak to the evolving and generally unmet needs of the research community. Libraries have made significant updates to their access and service models as they have come to terms with the so-called digital turn,[104] and the time has come for us to take stock of these initiatives, our accomplishments, and the opportunities that remain.

APPENDIX A: FORUM PROGRAM AND ATTENDEES

Program

Day One	
Listen and Learn	
9:30–10:15	Welcome and setting the scene
10:15–10:30	Break
10:30–12:00	Fishbowl storytelling
12:00–12:30	Debrief
Seek Collaborative Opportunity	
12:30–1:30	Working lunch
1:30–2:15	Group discussions by theme
2:15–2:35	Debrief
2:35–2:50	Break
2:50–3:35	Five by five by five
3:35–3:55	Debrief
3:55–4:30	Close out and setting the scene for day two
5:00–7:00	Reception for informal networking

Day Two	
Make Commitments	
9:30–9:45	Plenary
9:45–11:45	Deep engagement
	Toward a Declaration: Process for What We Want and How to Communicate It
	Toward a Road Map: Prioritizing Action Item by Cost-Impact
	Toward a Legal Infrastructure
	The Library Today (Toward a Better Tomorrow): Small Wins and Low-Hanging Fruit
11:45–12:00	Break
12:00–12:30	White paper brainstorming
12:30–12:40	Closing remarks

Table 1. Forum program

Attendees

- **Scott Althaus,** Merriam Professor of Political Science, Professor of Communication, and Director of the Cline Center for Advanced Social Research at the University of Illinois at Urbana-Champaign

- **Christine L. Borgman,** Distinguished Professor and Presidential Chair in Information Studies at the University of California, Los Angeles
- **Brandon Butler,** Director of Information Policy at the University of Virginia Library
- **Beth Cate,** Associate Professor at the School of Public and Environmental Affairs (SPEA) at Indiana University Bloomington
- **Marc Cormier,** Head of Humanities Publishing at Gale-Cengage
- **Krista Cox,** Director of Public Policy Initiatives at the Association of Research Libraries
- **Mary Ellen K. Davis,** Executive Director of the Association of College and Research Libraries
- **J. Stephen Downie,** Professor and Associate Dean for Research at the University of Illinois School of Information Sciences and the Illinois Co-director of the HathiTrust Research Center
- **Patricia Feeney,** Head of Metadata at Crossref
- **Lucie Guibault,** Associate Professor at the Law and Technology Institute of the Schulich School of Law at Dalhousie University
- **Wolfram Horstmann,** Director of Göttingen State and University Library
- **Clifford Lynch,** Executive Director of the Coalition for Networked Information
- **Peter Murray-Rust,** Reader Emeritus in Molecular Informatics at the University of Cambridge and Co-founder of ContentMine
- **Darby Orcutt,** Assistant Head, Collections and Research Strategy, at North Carolina State University Libraries
- **Thomas Padilla,** Visiting Digital Research Services Librarian at the University of Nevada Las Vegas
- **Michelle Paolillo,** Digital Curation Services Lead in the Department of Digital Scholarship and Preservation Services at Cornell University Library
- **Andrew Piper,** Professor and William Dawson Scholar in the Department of Languages, Literatures, and Cultures at McGill University
- **Matthew Sag,** Georgia Reithal Professor of Law at Loyola University of Chicago School, where he is also the Associate Director for Intellectual Property of the Institute for Consumer Antitrust Studies
- **Rachael G. Samberg,** Scholarly Communication Officer, University of California, Berkeley, Library
- **Jean P. Shipman,** Vice President of Global Library Relations for Elsevier
- **George Strawn,** Director of the Board on Research Data and Information (BRDI) at the National Academies of Sciences, Engineering, and Medicine
- **Paul F. Uhlir,** independent consultant in information policy and management
- **Günter Waibel,** Associate Vice Provost and Executive Director of the California Digital Library
- **Kate Wittenburg,** Managing Director of Portico
- **Glen Worthey,** Digital Humanities Librarian at the Stanford University Libraries

APPENDIX B: DETAILED METHODS

This appendix details methods for each of the project's four successive phases.

Phase 1: Literature Review

We performed a systematic 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 assertions. The review was limited to works in English from 2000 to 2017. While we primarily focused on the United States, we also included scholarship that addressed other legal jurisdictions, including Canada, Australia, the United Kingdom, and the European Union. To ensure coverage across multiple disciplines, we elected to conduct initial searches in prominent databases for law, library and information science, computer science, linguistics, e-science, digital humanities, and computational social science (tables 2 and 3). Prior to the initial review period, we outlined criteria for inclusion and exclusion, and we agreed upon a selection of search terms that we would use in a range of combinations until results became uniformly redundant (table 4).

ACM Digital Library
IEEE
INSPEC
LexisNexis
Library and Information Science Source
Scopus
Web of Science

Table 2. Databases queried

DH Quarterly
Journal of Digital Humanities
Computational Social Networks Science

Table 3. Individual journals queried

"text mining"	+	copyright
"data mining"		"intellectual property"
"text analysis"		licens*
"text analytics"		"terms of service"
"data analytics"		"web scraping"

Table 4. Search term combinations used in first phase of literature review

For our purposes, we included any materials that focused on providing library services, developing computational workflows, or addressing issues related to data sharing. We limited our interest in textual data to copyright-protected data, data provisioned through licensing agreements, and text scraped from web pages that were accessed via a paywall or subject to terms of service. Over the course of our search, we further refined our criteria by excluding TDM focused solely on indexing for search and retrieval and a body of literature focused on patent analysis.

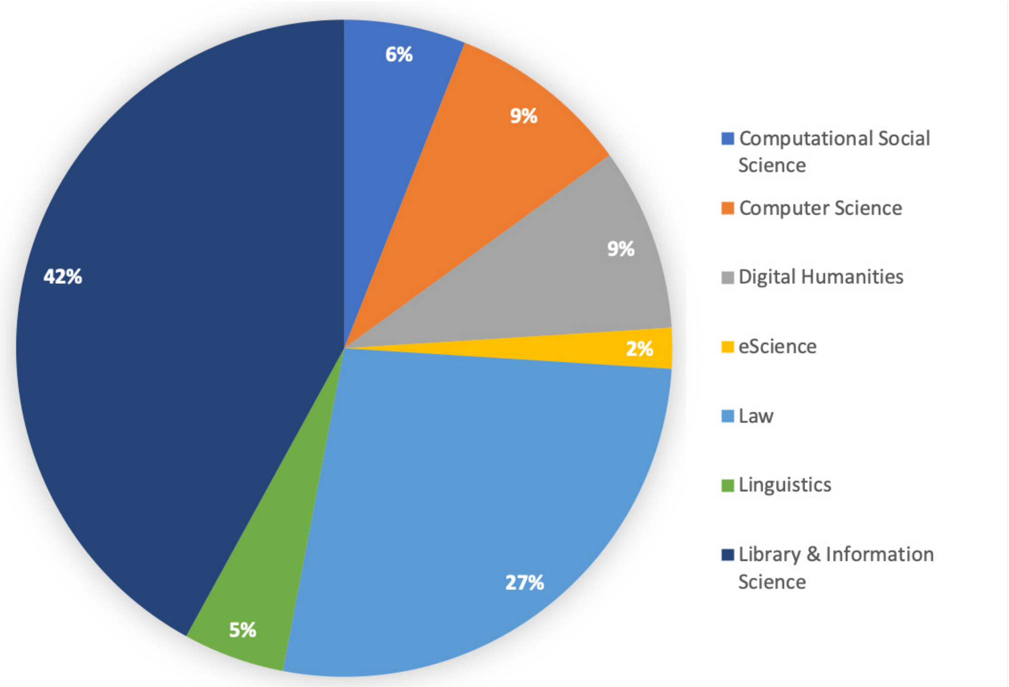

Figure 3. Disciplinary coverage of literature review

Our initial database search returned 103 results across seven categories, with the majority of articles discovered in library and information science (42%) or law (27%); see figure 3. In our second phase of review, we reviewed the cited references from articles identified in the first phase, accumulating references to 49 more items, which ranged from formal articles to blog posts to news announcements and press releases. From this list of 152 items, two members of the team selected a subset of 89 articles to read in full. In addition to taking structured notes, reviewers assigned each item one or more of the following tags: priority, access, use, dissemination, non-consumptive, exemplars, and legal justification.

Phase 2: Participant Selection and Pre-forum Interviews

Potential stakeholders were identified through the literature review and subsequent snowball sampling, and the final set of twenty-five forum participants included 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 two-page forum statement and an analysis of strengths, weaknesses, opportunities, and threats (SWOT) prior to the event.

The research team also designed a semi-structured interview protocol intended 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 (see appendix D). Each interview lasted no longer than an hour and covered seven questions that aligned with participants' research backgrounds, probed on the four components of the SWOT analysis, and elicited the participants' ideas about their priorities and potential strategies for action. 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 twenty-six thematic codes divided into six categories (table 5). Using an informal codebook, the team conducted an initial 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.[1]

Characterizations of TDM	Preference for different terminology
	Sources of text data (e.g., journal articles, social media, websites, novels, newspapers, historical archives)
	Future-orienting TDM
Risks	Inaction due to uncertainty
	Importance of critical mass/concern about uptake
	Chilling effect on scholarship
	Uneven access affecting quality
	Data security
Law	International aspects (e.g., copyright harmonization, choice of law)
	Testing the limits of fair use
	Certainty/uncertainty about lawfulness and what is allowable
	Licenses for text analysis
	Terms of service
	Ambiguity about ownership
Policy and advocacy	Risk aversion and institutional inertia
	Market value (e.g., balancing profit with security, customer demand, creating revenue streams)
	Cross-stakeholder exchange
	Raising awareness
	Seeking government-led solutions
Scholarship and scholarly communication	Technical training (e.g., skills, competencies, literacies)
	Added complexity for communicating research
	Supporting documentation and reproducibility
	Role of the library
Standardization and access workflows	Standardizing data formats
	Lack of shared terminology
	Codifying access procedures

Table 5. Informal codebook for qualitative content analysis of semi-structured interviews

Pre-forum findings 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 by assigning labels to each item 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 analyzed the aggregate spreadsheet by coding each item using one of seven common themes (table 6). Outcomes of initial coding were used to inform the final agenda for the national forum, which we discuss in detail below.

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

Table 6. Codebook for seven themes that emerged from SWOT analyses

Phase 3: Forum

To collectively and productively address the themes and tensions revealed during pre-forum analysis, the project team organized the national forum 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 adopted multiple strategies from the Liberating Structures menu to facilitate early engagement and encourage concrete outcomes within a relatively condensed period of time.[2] We selected Liberating Structures because it provides a framework of novel methods aimed at increasing individual participation in group conversations. The first day was structured around a set of small-group activities, each concluding with a W^3 debrief in which participants reflected 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 first morning was dedicated to a fishbowl storytelling session. Facilitators arranged the room with an inner circle of five chairs surrounded by a U-shaped conference table configuration. Participants were organized into five groups of five by stakeholder affiliation, with groups ordered from those with the most individual perspective to those with institutional and then collective perspectives: researchers, librarians, content providers, professional societies, and legal experts.

Each group was given ten minutes in the inner circle of the fishbowl to discuss their perspective on challenges their stakeholder group encounters when TDM and intellectual property intersect. Each ten-minute discussion was then followed by five minutes of Q&A with participants outside the circle. The directive for the morning was 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 TDM 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 facilitating TDM and how services might be improved.
- **Content providers.** Discuss your business model in relation to TDM with a focus on what makes it viable and what are known deal breakers when working with other stakeholders.
- **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 convened around three thematic topics that represent different parts of a researcher's workflow: finding and getting data, conducting analysis, and communicating results. This session transitioned toward 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 crafted a list of topics pertinent to each group. Because each group represented a snapshot of a continuous and cyclical process, the organizers rightly anticipated that topics from one group would overlap with another. Toward the end of a wide-ranging discussion, participants were asked to shift their focus back to the individual level and identify small actions that could be taken immediately to trigger momentum within and across stakeholder communities. Within the parlance of Liberating Structures, these are called "15% solutions," and they are designed encourage initial action in light of current resources and circumstances.

The third group session of the day was organized to quickly sort a set of sixteen recommendations that had been drawn from participants' two-page forum statements according to degrees of agreement and the predictability of the action's outcomes. Participants were reorganized into five groups of five, and each group was given an identical deck of index cards with one recommendation printed on each card. Within their groups, participants worked 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 moved each card horizontally based on how certain or predictable they anticipated the outcomes of the proposed action to be. The resulting matrix provided a bird's-eye view of the complexity of the actions proposed (figure 4).

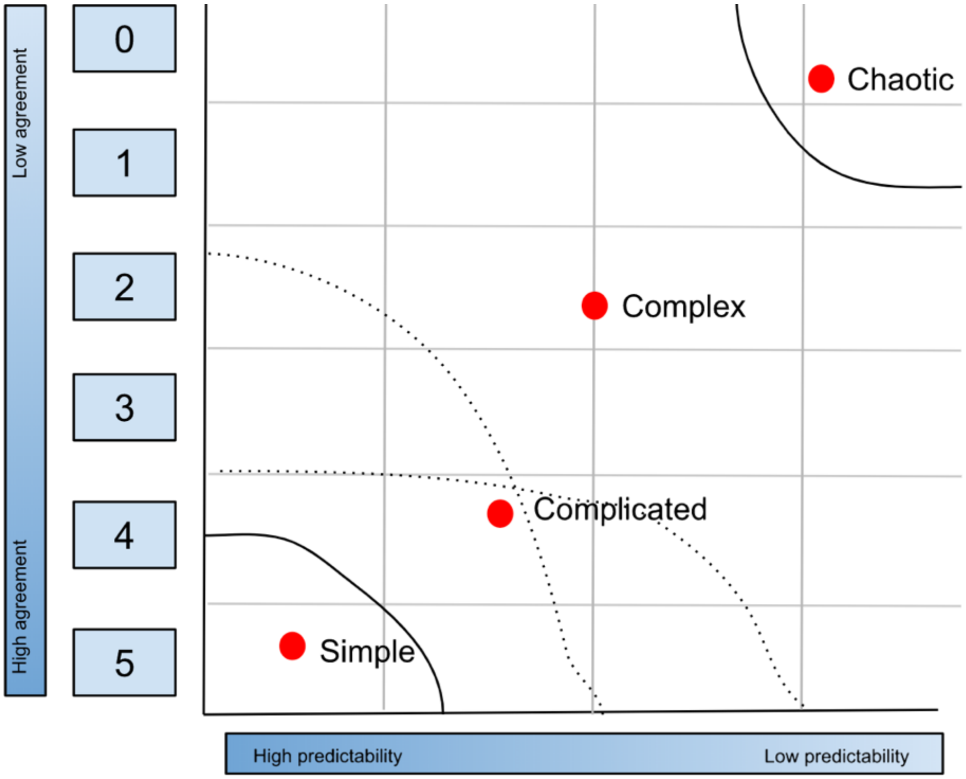

Figure 4. Agreement-predictability matrix

Once the sorting activity was complete, participants were asked to review the matrix and choose two to five priority items. Because the afternoon's directive was seeking collaborative opportunity, the project team recommended selecting from items with high agreement but let the groups decide how to prioritize predictability in their selection. For each prioritized action, participants were 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.

At the close of day one, participants regrouped to review the initial plans for the second day and reorganize activities based on emerging commitments and affinities. Afterward, participants attended an evening reception, which included a crowdsourcing exercise meant to gather and prioritize another, more ambitious set of recommendations. Each attendee was 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 were asked to mingle and pass cards but were instructed not to read the card until the bell had sounded. At the sound of the bell, participants read the card in their hand and marked it with a score from 1 ("I would absolutely not be interested in discussing this further") to 5 ("I would definitely like to discuss this idea further").[3] Cards were passed and rated four more times, each by a different attendee. After all five rounds, a facilitator tallied the scores on the back of each card, and the team shared the results at the following morning's plenary session.

Make Commitments

The directive for day two shifted from seeking collaborative opportunities to making commitments. For the bulk of the morning, participants regrouped into attendee-driven Birds of a Feather sessions, where goals, strategies, and themes were determined via small-group consensus. The project team asked that participants spend the final few minutes of each session documenting concrete commitments related to the topic at hand. Group affinities ranged from the pragmatic to the ideal and pursued approaches related to education, infrastructure, and advocacy.

The final two sessions of the forum were devoted to plenary-style conversation and decision-making. One session was dedicated to gathering participants' feedback on key themes for this ACRL white paper on TDM with use-limited datasets, focusing explicitly on situating library action within a broader landscape. Participants deliberated on 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 expanded their focus to the full spectrum of potential stakeholders while focusing on tangible next steps, concrete commitments, and sustainable strategies for ongoing communication and engagement.

Phase 4: Post-forum Analysis

The project team systematically gathered data throughout the two days of the forum. A dedicated notetaker captured the outcomes of each plenary session and small-group discussion. Stacks of index cards were also scattered throughout the room, and organizers gathered handwritten materials from participants in the form of 15% solutions, discussion debriefings, and miscellaneous notes. Following the forum, the project team transcribed all handwritten materials and combined them with the typewritten notes taken during each session. These, in turn, were combined with the initial pre-forum interview transcripts and SWOT analyses to form a corpus for qualitative analysis in ATLAS.ti. Using the forum documentation, the project team conducted open coding on twenty-two separate documents, generating a list of 129 unique codes. Using the results of open coding and the preliminary codebooks used during pre-forum analyses, the project team formalized a codebook with eleven code families and thirty-nine codes (table 7). Drawing on approaches from grounded theory,[4] the project team sought to uncover relationships among codes that shed light on the library's current and potential role as a facilitator of—and an advocate for—rigorous text mining with proprietary data.

Access	"Just give me the data."	Use for instances where people talk about unmediated access.
	General access	Use as a catchall for any kind of access that doesn't fall under open access—terms and conditions, black box "access," illegal downloads, etc.
	Open access	Use for any mentions of open access publishing or the philosophy of open access.
	Use and reuse	Use for examples or speculation on activities that occur after the point of access.
Conflicts	(Mis)communication	Use for all instances of challenges relating to communication.
	Critique of terminology	Use for any instance in which a given term is discussed as problematic or a preferable term is proposed.
	International issues	Use for discussions of international collaboration, copyright harmonization, legal aspects of TDM outside the US.
	Opportunity costs	Use for occasions when current policy/practice prohibits beneficial outcomes (e.g., innovation, new partnerships, etc.).
	Uncertainty	Use whenever uncertainty is expressed about what is allowable (legally, ethically, technically).
Financial aspects	Financial concerns	Use for all discussion of costs, funding, budgets, or other matters relating to money.

Legal aspects	"The right to read is the right to mine."	Use for instances in which the principle that human readability and machine readability should be treated as legally/morally equivalent.
	Fair use	Use primarily for discussions of the fair use doctrine, but also use for copyright issues more generally.
	Legal infrastructure	Use for overarching set of legal considerations and processes; include relevant laws such as CFAA, DMCA; processes of legal determination; prior cases and precedents.
	Legal risks	Use for discussions of potentially problematic legal questions.
On data	"Collections as data"	Use for discussions of collections of data AND other forms of collections that might be used as data.
	Data acquisition	Use for the technical aspects of getting access to data.
	Data governance	Use for policies about data, including data about faculty; sensitive data; and issues of privacy, security, and ethics.
	Data munging	Use for activities related to data transformation and remediation, including data cleaning, processing, and wrangling.
	Data quality	Use for issues including fitness for use, FAIR data principles, and discoverability.
On libraries	Librarian staffing and skills	Use for instances of what kind of work librarians do, how librarians do it, and what they need to know in order to complete their work.
	Library licenses	Use for the terms of licenses or library licensing generally.
	Library relationship with scholars	Use for discussions of interactions between librarians and scholars, successful or otherwise.
	License negotiations	Use for the process of negotiation, with libraries or otherwise.
Organizational aspects	Business models	Use for discussions of how TDM work operates at a business level, or how decisions about TDM affect business.
	Policies	Use for local or institutional policies (nonlegal contexts).
	Related projects or initiatives	Use to capture references to work that is relevant to the IMLS national forum but conducted by other groups.
	Stakeholders, roles, and responsibilities	Use generally for discussion of different people who need to be involved in TDM research and decision-making.
Process	Time scales	Use for discussions of length and durations of TDM research, negotiations, etc.
	Understanding needs	Use for instances of consensus seeking and expressions of need from the perspective of individual scholars, libraries and institutions of higher education, and content providers.
	Workflows and documentation	Use for discussions of how the work of TDM proceeds and is captured, including researchers' strategies during TDM, library support processes, and decision-making among stakeholders.
Research practices	Research practices	Use for discussion of research questions, methods, and styles; (inter)disciplinary issues; scholarly publishing; and reproducibility.
Socio-technical issues	Service models	Use for discussions of overarching services that integrate both social and technical considerations.
	TDM tools	Use for discussion of specific tools and platforms used to conduct TDM.
	Technical infrastructure	Use for discussions of virtual machines, data transfer protocols, non-consumptive models, and other aspects of support that rely on technical solutions.

Strategies	Advocacy	Use for strategies that focus on getting the word out, lobbying for change, and empowering individuals and organizations to move forward with TDM.
	Norms, standards, and best practices	Use for strategies that suggest the development of best practice guides and the establishment of norms and standards (for disciplines, organizations, etc.).
	Risk management	Use for discussions of attitudes toward risk and how to manage it.
	TDM stories	Use for instances of storytelling in which people discuss their own or others' actual experiences doing TDM.
	Training and literacies	Use for strategies that focus on developing training opportunities as well as focusing on issues around information literacy.

Table 7. Formal codebook for post-forum analysis of all collected data

Notes

1. Hsiu-Fang Hsieh and Sarah E. Shannon, "Three Approaches to Qualitative Content Analysis," *Qualitative Health Research* 15, no. 9 (2005): 1277–88, https://doi.org/10.1177/1049732305276687.
2. Henri Lipmanowicz and Keith McCandless, Liberating Structures home page, accessed December 3, 2018, http://www.liberatingstructures.com/.
3. In practice, this scale also served as a proxy indicator for whether participants liked the idea.
4. Juliet M. Corbin and Anselm L. Strauss, *Basics of Qualitative Research*, 3rd ed. (Los Angeles: Sage, 2008).

APPENDIX C: INTERVIEW PROTOCOL

Data Mining Research Using In-Copyright and Limited-Access Text Datasets

Estimated length: 45 to 60 minutes

Introduction

Hello, my name is [ ] from the University of Illinois. Thank you for agreeing to be interviewed as part of our national forum project on text data mining of in-copyright and limited-access datasets.

First let me tell you about the study. The project team is headed by Bertram Ludäscher at the School of Information Sciences at the University of Illinois at Urbana-Champaign with Megan Senseney, Beth Namachchivaya, and Eleanor Dickson leading the research initiative. This project is supported by a National Leadership Grant from the Institute for Museum and Library Services (IMLS).

The full research project consists of an environmental scan, this set of interviews, SWOT analyses generated by key stakeholders, a day-and-a-half-long meeting, and an ACRL white paper. This interview is designed to learn more about you as a national forum participant and assist you in the development of your forum statement and SWOT analysis.

Your participation in this interview is entirely voluntary, with no risks besides those of everyday life. You are free to discontinue participation in the study at any time, and you are free to request that we cease recording at any time. Your responses today are confidential; our goal is to help you prepare materials that you will share with other forum participants. Summary data and de-identified information from interview may also be used in aggregate for future project disseminations and reports.

Before we begin, let me review the consent form and ask for your verbal consent.

Interview Questions

Background

1. Could you start by telling us a bit about your research or professional experiences with text data mining?
2. Could you walk us through a time when your have dealt with texts that are under copyright or licensed or otherwise restricted due to intellectual property rights in relation to data mining.

As you may recall, forum participants are drafting SWOT (strength, weakness, opportunity, and threat) analyses around text mining with copyrighted or licensed data. The following questions will walk through the key areas in a SWOT analysis and ask you to reflect on each.

Strengths, Weaknesses, Opportunities, and Threats

3. Let's discuss some existing strengths from your perspective [as a researcher, as a content provider, as a librarian] of accessing and conducting text mining on copyrighted or licensed texts.

PROMPTS

- Are there services or strategies that have proven useful?
- For researchers: Are there positive impacts in your research outcomes?
- For librarians and content providers: Are there perceived strengths in including these services as part of your model for providing research support?

4. Maintaining your perspective [as a researcher, as a content provider, as a librarian], where are the current weaknesses in the processes of accessing [or providing access to] in-copyright and licensed texts and utilizing text mining techniques for research?

PROMPTS

- What barriers to text mining with limited-access data have you encountered?
- What challenges does your sector face with regard to [providing access/gaining access] to copyrighted or licensed data?
- How does working with these data impact scholarly communication?

5. What kinds of opportunities are apparent [for libraries, for your research or discipline, for your industry] in using in-copyright and limited-access datasets in text mining research?

PROMPTS

- How might text data mining of copyrighted or licensed texts advance scholarship?
- What benefits to your sector do you see for providing access for text data mining to copyrighted or licensed texts?

6. Could you describe some potential threats that text data mining with in-copyright and limited-access datasets might pose for you personally or your field more generally?

PROMPTS

- How might the threat of legal action impact your work?
- How might these threats impact your productivity?

Synthesis

7. Now that we've reflected on the TDM landscape with regard to copyright, licensing, and intellectual property, what would you identify as the highest priorities for establishing a research and development agenda?
8. In cases where you perceive threats and weaknesses, could you think out loud a bit about potential strategies for addressing those issues?
9. Is there anything else on your mind about text data mining and intellectual property that we haven't covered today?

Closing

Thank you for your time. We hope this exercise has helped seed ideas for your forum statement and SWOT analysis. We would also be happy to share a copy of this recording with you if you would find that helpful. As a reminder, your statement and SWOT analysis will be due on [date]. These will be shared with participants for open discussion and review in advance of our meeting in April. We hope these documents will help set the agenda for the event and allow us to utilize our face-to-face time effectively.

[NOTE: In accordance with common interview practice, we expect to make minor adjustments to the instrument for individual participants during the course of each interview based on their responses, the relevance of questions to their research, and how the conversation evolves.]

APPENDIX D: GLOSSARY

Collections as data: A concept that promotes computational or algorithmic uses of library collections, for example, for data mining or building digital maps. The notion of collections as data has been popularized by the IMLS-funded Always Already Computational and Mellon Foundation-funded Collections as Data: Part to Whole, as well as by the Library of Congress's National Digital Initiatives group.[1]

Computer Fraud and Abuse Act: A law that prohibits unauthorized access of computers and computer systems as well as restricts access in excess of an otherwise authorized use, as defined in 18 U.S. Code § 1030. As some researchers use web-scraping or otherwise gather data from websites and databases, there is concern that such behaviors could open them to legal risk under the Computer Fraud and Abuse Act if they violate terms of use or exceed the intended, authorized use for the resource.[2]

Digital Millennium Copyright Act: A law that, among other restrictions, prohibits actions that break technological protection measures, such as digital rights management (DRM), as stipulated in 17 U.S. Code § 512. Researchers who seek data from certain sources may be legally liable if they thwart technological protection measures in order to do so, for example, by breaking DRM on an e-book to access the text within.[3]

Digital scholarship: A broad term that for purposes of this paper relates to the variety of ways in which "digital evidence" and computational methods are incorporated into academic research. Generally speaking, it is used to evoke scholarly behavior and practice in which digital objects, data, or workflows (i.e., analysis techniques) are a primary component of the research project.[4]

Fair dealing: A legal framework found in the Commonwealth of Nations, including Canada, that enumerates possible defenses for uses of copyrighted material without permission from the copyright holder. Fair dealing may be applied to research, private study, education, parody, satire, criticism, review, and news reporting. In order to be considered a fair dealing, the use is evaluated by six-criteria test.[5]

Fair use: A United States legal doctrine that allows copyrighted material to be used without permission from the copyright holder in certain cases, as described in Section 107 of the US Copyright Code. Such cases include scholarship, research, teaching, comment, criticism, and news reporting. Determinations about fair use are made through application of a four-factor test that weighs the nature, purpose, and potential outcome of the use.[6]

Model license: Recommended language for libraries to consider during negotiation with vendors that may include the entire license or clauses. The California Digital Library, Center for Research Libraries, and Canadian Research Knowledge Network provide notable examples.[7]

Non-consumptive research: First defined in a rejected 2010 settlement agreement in Authors Guild v. HathiTrust. Later operationalized by HathiTrust in its Non-consumptive Use Research Policy. Defined by HathiTrust as "research in which computational analysis is performed on one or more volumes (textual or image objects)…, but not research in which a researcher reads or displays substantial portions of an in-copyright or rights-restricted volume to understand the expressive content presented within that volume."[8] Also called non-expressive use.

Open access: According to the Scholarly Publishing and Academic Research Consortium (SPARC), "Open Access is the free, immediate, online availability of research articles combined with the rights to use these articles fully in the digital environment."[9] As reflected in this definition, full use would entail text data mining applications.

Public copyright license/open license: Mechanism for copyright holders to optionally assign a license to their work that grants a blanket exception for their work to be reused without permission, sometimes within certain parameters such as for noncommercial purposes or with attribution. Notable examples include Creative Commons licenses and the GNU General Public License.[10]

Reproducibility: Defined by Peng as "an attainable minimum standard for assessing the value of scientific claims, particularly when full independent replication of a study is not feasible. The standard of reproducibility calls for the data and the computer code used to analyze the data be made available to others."[11] While replication would require that comparable data be independently gathered and analyzed, reproducibility is a means for ensuring that computational research can be externally verified using the same data and method for analysis.

Research workflow: The series of steps a scholar follows in the course of conducting research. While a comprehensive workflow may cover early steps, such as seeking funding, through later steps, such as presenting at conferences, for computationally intensive research the key parts of the workflow include the actions taken and code used to gather, prepare, analyze, and share the data used or generated.[12]

Structured/unstructured data: Structured data is stored in a database or other organized format where the elements of the dataset are separated and distinguished and may be searched. Unstructured data is not as highly organized, and the elements within have not been distinguished; thus, search is problematic. A spreadsheet of temperature measurements from a series of experiments would be structured data, whereas a text file containing the contents of a journal article would be unstructured data.[13]

Text data mining: According to Hearst, it is "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation."[14] More than search and discovery, text data mining uses data science techniques to uncover patterns in text that reveal information about the text. Also referred to as text mining, text analysis, or computational text analysis.

Notes

1. "The Santa Barbara Statement on Collections as Data," version 2, Always Already Computational—Collections as Data, 2018, https://collectionsasdata.github.io/statement/; Library of Congress, *Collections as Data: Stewardship and Use Models to Enhance Access*, executive summary, accessed September 20, 2019, http://digitalpreservation.gov/meetings/dcs16/AsDataExecutiveSummary_final.pdf.
2. See "18 U.S. Code §1030. Fraud and Related Activity in Connection with Computers," Legal Information Institute, Cornell Law School, accessed September 20, 2019, https://www.law.cornell.edu/uscode/text/18/1030.
3. See "17 U.S. Code §512. Limitations on Liability Relating to Material Online," Legal Information Institute, Cornell Law School, accessed September 20, 2019, https://www.law.cornell.edu/uscode/text/17/512.
4. See Association of Research Libraries, "Digital Scholarship Profiles," accessed September 20, 2019, https://www.arl.org/focus-areas/scholarly-communication/digital-scholarship.
5. See "Learn More about Fair Dealing," Fair Dealing ©anada, accessed September 20, 2019, https://fair-dealing.ca/.
6. See "More Information on Fair Use," Copyright.gov, accessed September 20, 2019, https://www.copyright.gov/fair-use/more-info.html.
7. Mihoko Hosoi, "CDL Model License Revised," California Digital Library, January 25, 2017, https://www.cdlib.org/cdlinfo/2017/01/25/cdl-model-license-revised/; Center for Research Libraries, "Model Licenses," LIBLICENSE: Licensing Digital Content, accessed September 14, 2019, http://liblicense.crl.edu/licensing-information/model-license/; Canadian Research Knowledge Network, "Model License," accessed September 20, 2019, http://www.crkn-rcdr.ca/en/model-license.
8. HathiTrust Research Center Task Force for Non-consumptive Research Use Policy, "Non-consumptive Use Research Policy," February 20, 2017, https://www.hathitrust.org/htrc_ncup.
9. "Open Access," SPARC, accessed September 20, 2019, https://sparcopen.org/open-access/.
10. Creative Commons home page, accessed September 20, 2019, https://creativecommons.org/; "GNU General Public License," version 3, June 20, 2007, GNU Operating System, https://www.gnu.org/licenses/gpl.html.

11. Roger D. Peng, “Reproducible Research in Computational Science,” *Science* 334, no. 6060 (December 2, 2011), 1226-1227, https://doi.org/10.1126/science.1213847.
12. See Susan Kroll and Rick Forsman, *A Slice of Research Life* (Dublin, OH: OCLC Research, June 2010), https://www.oclc.org/content/dam/research/publications/library/2010/2010-15.pdf.
13. See Amir Gandomi and Murtaza Haidar, “Beyond the Hype: Big Data Concepts, Methods, and Analytics,” *International Journal of Information Management* 35, no. 2 (April 2015): 137–44, https://doi.org/10.1016/j.ijinfomgt.2014.10.007.
14. Marti Hearst, “What Is Text Mining?” unpublished paper, October 17, 2003, http://people.ischool.berkeley.edu/~hearst/text-mining.html.

NOTES

1. Authors Guild v. HathiTrust 755 F.3d 87 (2d Cir. 2014).
2. Hillary Miller, “Securing Text and Data Mining Rights for Researchers in Academic Libraries” (master’s thesis, University of North Carolina, 2015), https://cdr.lib.unc.edu/record/uuid:704c0c1e-e103-4242-85d7-d3abf5b25835.
3. “Text Mining,” Berkeley Library Scholarly Communication Services, accessed December 3, 2018, http://www.lib.berkeley.edu/scholarly-communication/publishing/copyright/text-mining; “Digging Deeper, Reaching Further: Libraries Empowering Users to Mine the HathiTrust Digital Library Resources,” HathiTrust Research Center, accessed December 3, 2018, https://teach.htrc.illinois.edu/.
4. Examples of library guides to TDM include University of Chicago, “Text and Data Mining,” accessed September 14, 2019, http://guides.lib.uchicago.edu/textmining; University of Cambridge, “Text and Data Mining: Home,” accessed September 14, 2019, https://libguides.cam.ac.uk/tdm/home; and Penn State University, “Text Mining: Web-Based Resources,” accessed September 14, 2019, https://guides.libraries.psu.edu/textmining/web.
5. Darby Orcutt, “Library Support for Text and Data Mining,” *Online Searcher* 39, no. 3 (June 5, 2015): 27–30; Andras Schwarcz, “Text and Data Mining: A New Service for Libraries?” *European Parliamentary Research Service Blog*, October 20, 2017, https://epthinktank.eu/2017/10/20/text-and-data-mining-a-new-service-for-libraries/.
6. This framing of the “theoretical right” was suggested by a librarian participant of the national forum.
7. Yasmeen Shorish, “Data Data Everywhere …but Do We Want to Drink?” *ACRL TechConnect* (blog), July 16, 2015, https://acrl.ala.org/techconnect/post/data-data-everywherebut-do-we-want-to-drink/.
8. Principle 1 from “The Santa Barbara Statement on Collections as Data,” version 2, Always Already Computational—Collections as Data, 2017, https://collectionsasdata.github.io/.
9. The project website is available at “Data Mining Research Using In-Copyright and Limited-Access Text Datasets,” Text Mining with Limited Access Data 2018 National Forum, accessed September 14, 2019, https://publish.illinois.edu/limitedaccess-tdm/; tweets from the forum can also be accessed at https://twitter.com/ using the hashtag #TDMForum18.
10. Casey M. Bergman, Lawrence E. Hunter, and Andrey Rzhetsky, “Announcing the PLOS Text Mining Collection,” *PLOS ONE Community Blog*, April 17, 2013, https://blogs.plos.org/everyone/2013/04/17/announcing-the-plos-text-mining-collection/; Bernard F. Reilly, When Machines Do Research, Part 2: Text-Mining and Libraries,” *Charleston Advisor* 14, no. 2 (October 2012): 75–76, https://doi.org/10.5260/chara.14.2.75.
11. Henri Lipmanowicz and Keith McCandless, Liberating Structures home page, accessed December 3, 2018, http://www.liberatingstructures.com/.
12. Hsu-Hao Tsai, “Global Data Mining: An Empirical Study of Current Trends, Future Forecasts and Technology Diffusions,” *Expert Systems with Applications* 39, no. 9 (July 2012): 8172–81, https://doi.org/10.1016/j.eswa.2012.01.150.
13. Alan Jovic, Karla Brkic, and Nikola Bogunovic, “An Overview of Free Software Tools for General Data Mining,” in *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO): Proceedings*, ed. Petar Biljanovic, Zeljko Butkovic, Karolj Skala, Stjepan Golubic, Marina Cicin-Sain, Vlado Sruk, Slobodan Ribaric, et al. (Rijeka, Croatia: Croatian Society for Information and Communication Technology, Electronics and Microelectronics—MIPRO, 2014), 1112–17, https://doi.org/10.1109/MIPRO.2014.6859735.
14. ContentMine home page, accessed December 3, 2018, http://contentmine.org/; Stéfan Sinclair and Geoffrey Rockwell, Voyant Tools home page, accessed December 3, 2018, https://voyant-tools.org/.
15. See, for example, International Federation of Library Associations (IFLA), *IFLA Statement on Text and Data Mining* (The Hague, Netherlands: IFLA, 2013), https://www.ifla.org/publications/node/8225; Association of Research Libraries (ARL), *Text and Data Mining and Fair Use in the United States*, issue brief (Washington, DC: ARL, June 2015), https://www.arl.org/wp-content/uploads/2015/06/TDM-5JUNE2015.pdf; All European Academies (ALLEA), *The Text and Data Mining Exception and the Enhancement of Access to Scientific Information in Europe* (Berlin: ALLEA, November 2017), https://www.knaw.nl/shared/resources/internationaal/bestanden/allea_pwgipr_statement_tdm_2017; and Ligue des Bibliothèques Européennes de Recherche (LIBER), “Text and Data Mining,” accessed September 14, 2019, https://libereurope.eu/text-data-mining/.
16. Jennifer Guiliano and Mia Ridge, “The Future of Digital Methods for Complex Datasets: An Introduction,” *International Journal of Humanities and Arts Computing: A Journal of Digital Humanities* 10, no. 1 (March 2016): 1–7, https://doi.org/10.3366/ijhac.2016.0155.
17. Orcutt, “Library Support for Text and Data Mining”; Jennifer Molloy, Maximillian Haeussler, Peter Murray-Rust, and Charles Oppenheim, “Responsible Content Mining,” in *Working with Text: Tools, Techniques, and Approaches*

for Text Mining, ed. Emma L. Tonkin and Gregory J. L. Tourte (Cambridge, MA: Chandos, 2016), 89–109.

18. Sag, Matthew. "The Google Book Settlement and the Fair Use Counterfactual." *New York Law School Law Review* 55, no. 1 (2010/2011): 19–75, https://digitalcommons.nyls.edu/nyls_law_review/vol55/iss1/2/; Ivy Anderson et al., "What Should Be the Conditions on Libraries Digitizing, Maintaining and Making Available Copyrighted Works," *Columbia Journal of Law and the Arts* 36, no. 4 (2013): 587–606, https://heinonline.org/HOL/P?h=hein.journals/cjla36&i=603; Angel Siegfried Diaz, "Fair Use and Mass Digitization: The Future of Copy-Dependent Technologies after Authors Guild v. HathiTrust," *Berkeley Technology Law Journal* 28 (2013), https://doi.org/10.15779/Z38X700; Samuelson, Pamela. "The Quest for a Sound Conception of Copyright's Derivative Work Right." *Georgetown Law Journal* 101, no. 6 (2013): 1505–64. https://dx.doi.org/10.2139/ssrn.2138479; Matthew Sag, "Orphan Works as Grist for the Data Mill," *Berkeley Technology Law Journal* 27, no. 3 (2012), article 9, https://doi.org/10.15779/Z387M5B; Jennifer M. Urban, "How Fair Use Can Help Solve the Orphan Works Problem," *Berkeley Technology Law Journal* 27, no. 3 (2012): 1379–492, http://btlj.org/data/articles2015/vol27/27_3_S/27-berkeley-tech-l-j-1379-1430.pdf; Hansen, David R., Kathryn Hashimoto, Gwen Hinze, Pamela Samuelson, and Jennifer M. Urban. "Solving the Orphan Works Problem for the United States." *Columbia Journal of Law and the Arts* 37, no. 1 (2013): 1–55. https://doi.org/10.7916/jla.v37i1.2145; Maurizio Borghi and Stavroula Karapapa. "Non-display Uses of Copyright Works: Google Books and Beyond." *Queen Mary Journal of Intellectual Property* 1, no. 1 (April 2011), https://dx.doi.org/10.2139/ssrn.2358912.
19. Matthew L. Jockers, Matthew Sag, and Jason Schultz, Brief of Digital Humanities and Law Scholars as Amici Curiae in Authors Guild v. Google (August 3, 2012), http://dx.doi.org/10.2139/ssrn.2102542.
20. Authors Guild v. HathiTrust 755 F.3d 87 (2d Cir. 2014); Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015).
21. Ian Hargreaves, *Digital Opportunity* (Newport, South Wales, UK: UK Intellectual Property Office, 2011), https://www.gov.uk/government/publications/digital-opportunity-review-of-intellectual-property-and-growth; Ian Hargreaves et al., *Standardisation in the Area of Innovation and Technological Development, Notably in the Field of Text and Data Mining* (Luxembourg: Publications Office of the European Union, 2014), https://doi.org/10.2777/71122.
22. Andrés Guadamuz and Diane Cabell, "Data Mining in UK Higher Education Institutions: Law and Policy," *Queen Mary Intellectual Property Review* 4, no. 1 (2014): 3–29, https://papers.ssrn.com/abstract=2446447.
23. Richard Van Noorden, "Text-Mining Spat Heats Up," *Nature News* 495, no. 7441 (March 21, 2013): 295, https://doi.org/10.1038/495295a.
24. Sag, "Orphan Works as Grist for the Data Mill"; Anderson et al., "What Should Be the Conditions on Libraries Digitizing, Maintaining and Making Available Copyrighted Works"; Urban, "How Fair Use Can Help Solve the Orphan Works Problem."
25. Jockers, Sag, and Schultz, Brief of Digital Humanities and Law Scholars as Amici Curiae in Authors Guild v. Google; Authors Guild v. Google Inc., 770 F. Supp. 2d 666 (S.D.N.Y. 2011).
26. Brandon C. Butler, "Transformative Teaching and Educational Fair Use after Georgia State," *Connecticut Law Review* 48, no. 2 (2015): 473–530, https://docs.wixstatic.com/ugd/17f3ef_48f887e2d77a4848bda319e2ccf57fdc.pdf.
27. Jane C. Ginsburg, "Fair Use for Free, or Permitted-but-Paid," *Berkeley Technology Law Journal* 29, no. 3 (2015): 1383–1446, https://doi.org/10.15779/Z38GW14.
28. Jonathan Band, "What Does the HathiTrust Decision Mean for Libraries?" *Research Library Issues,* no. 285 (January 2015): 7–13, https://doi.org/10.29242/rli.285.3; Brandon C. Butler, "Fair Use Rising: Full-Text Access and Repurposing in Recent Case Law," *Research Library Issues,* no. 285 (2015): 3–6, https://doi.org/10.29242/rli.285.2; Hansen et al., "Solving the Orphan Works Problem for the United States."
29. Niva Elkin-Koren and Orit Fischman-Afori, "Rulifying Fair Use," *Arizona Law Review* 59, no. 1 (2017): 161–200, http://arizonalawreview.org/rulifying-fair-use/; Diaz, "Fair Use and Mass Digitization."
30. Butler, "Transformative Teaching and Educational Fair Use after Georgia State."
31. Peter B. Kaufman and Jeff Ubois, "Good Terms—Improving Commercial-Noncommercial Partnerships for Mass Digitization: A Report Prepared by Intelligent Television for RLG Programs, OCLC Programs and Research," *D-Lib Magazine* 13, no. 11/12 (November 2007), https://doi.org/10.1045/november2007-kaufman; Dillon, Cy. "Transformative Use: An Update on the Google Books Case with Jonathan Band." *Virginia Libraries* 60, no. 2 (August 1, 2014). http://doi.org/10.21061/valib.v60i2.1296; Borghi and Karapapa, "Non-display Uses of Copyright Works."
32. Molloy et al., "Responsible Content Mining."
33. For a brief discussion of CFAA and text mining research, see Brett Currier, "You May Be Sued …or Arrested," *Intersections* (blog), Scholarly Communications at the University of North Carolina, Chapel Hill, April 30, 2015, https://blogs.lib.unc.edu/intersections/index.php/2015/04/30/you-may-be-sued-or-arrested/; see also Sandvig v. Sessions, 315 F. Supp. 3d 1 (U.S.D.C. Dist. Columbia 2018). With regard to DMCA, see Krista Cox, "Internation-

al Copyright Developments: From the Marrakesh Treaty to Trade Agreements," *Research Library Issues,* no. 285 (2015): 14–22, https://doi.org/10.29242/rli.285.4.

34. Liane Colonna, "A Taxonomy and Classification of Data Mining," *Southern Methodist University Science and Technology Law Review* 16, no. 2 (Fall 2013): 309–70, https://scholar.smu.edu/scitech/vol16/iss2/4/.
35. Jerome H. Reichman and Paul F. Uhlir. "A Contractually Reconstructed Research Commons for Scientific Data in a Highly Protectionist Intellectual Property Environment," *Law and Contemporary Problems* 66, no. 1/2 (2003): 315–462, https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=1283&context=lcp; Diane McDonald and Ursula Kelly, *Value and Benefits of Text Mining*, report (London: Jisc, March 2012, upd. August 5, 2019), https://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining.
36. Borghi and Karapapa, "Non-display Uses of Copyright Works."
37. Jerome H. Reichman and Ruth L. Okediji, "When Copyright Law and Science Collide: Empowering Digitally Integrated Research Methods on a Global Scale," *Minnesota Law Review* 96, no. 4 (April 2012): 1362–1480, http://www.minnesotalawreview.org/wp-content/uploads/2012/08/ReichmanOkediji_MLR1362.pdf.
38. Maarten Truyens and Patrick Van Eecke, "Legal Aspects of Text Mining," *Computer Law and Security Review* 30, no. 2 (April 2014): 153–70, https://doi.org/10.1016/j.clsr.2014.01.009.
39. United Kingdom Intellectual Property Office, *Exceptions to Copyright: Research* (Newport, UK: Intellectual Property Office, March 2014), https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf.
40. Neil Stewart et al., "Liberating Data: How Libraries and Librarians Can Help Researchers with Text and Data Mining," *London School of Economics Impact Blog*, July 12, 2016, http://blogs.lse.ac.uk/impactofsocialsciences/2016/07/12/how-libraries-and-librarians-can-help-with-text-and-data-mining/.
41. Marco Caspers and Lucie Guibault, "A Right to 'Read' for Machines: Assessing a Black-Box Analysis Exception for Data Mining," *Proceedings of the Association for Information Science and Technology* 53, no. 1 (December 27, 2016): 1–5, https://doi.org/10.1002/pra2.2016.14505301017.
42. Matěj Myška, "Text and Data Mining of Grey Literature for the Purpose of Scientific Research," *Grey Journal* 13 (January 2, 2017): 32–37.
43. Indra N. Sarkar, "Biodiversity Informatics: Organizing and Linking Information across the Spectrum of Life," *Briefings in Bioinformatics* 8, no. 5 (September 2007): 347–57, https://doi.org/10.1093/bib/bbm037.
44. Sheryl P. Denker, "Unrestricted Text and Data Mining with allofPLOS," *The Official PLOS Blog*, November 28, 2017, https://blogs.plos.org/plos/2017/11/unrestricted-text-and-data-mining-with-allofplos/.
45. See, for example, Aaron M. Cohen and William R. Hersh, "A Survey of Current Work in Biomedical Text Mining," *Briefings in Bioinformatics* 6, no. 1 (2005): 57–71, https://www.ncbi.nlm.nih.gov/pubmed/15826357; Yanpeng Li et al., "A Framework for Semisupervised Feature Generation and Its Applications in Biomedical Literature Mining," *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 8, no. 2 (March 2011): 294–307, https://doi.org/10.1109/TCBB.2010.99; Andrée J. Rathemacher, "Developing Issues in Licensing: Text Mining, MOOCs, and More," *Serials Review* 39, no. 3 (2013): 205–10, https://doi.org/10.1080/00987913.2013.10766397; Matthew L. Jockers and Ted Underwood, "Text-Mining the Humanities." In *A New Companion to Digital Humanities*, ed. Susan Schreibman, Raymond G. Siemens, and John Unsworth (Malden, MA: John Wiley and Sons, 2015), 291–306, https://doi.org/10.1002/9781118680605.ch20.
46. Melissa M. Terras, "The Potential and Problems in Using High Performance Computing in the Arts and Humanities: The Researching e-Science Analysis of Census Holdings (ReACH) Project," *Digital Humanities Quarterly* 3, no. 4 (2009), http://www.digitalhumanities.org/dhq/vol/3/4/000070/000070.html.
47. Beth Bernhardt et al., "Revolutionizing Scholarship: A Panel Discussion on Text and Data Mining," *Serials Review* 41, no. 3 (July 2015): 184–86, https://doi.org/10.1080/00987913.2015.1064514.
48. Clark, Jonathan. *Text Mining and Scholarly Publishing*. Publishing Research Consortium, 2012. https://www.stm-assoc.org/2012_01_01_PRC_Clark_Text_Mining_and_Scholarly_Publishing.pdf.
49. Rachael Lammey, "CrossRef's Text and Data Mining Services," *Learned Publishing* 27, no. 4 (October 2014): 245–50, https://doi.org/10.1087/20140402.
50. Richard Van Noorden, "Trouble at the Text Mine," *Nature News* 483, no. 7388 (March 8, 2012): 134, https://doi.org/10.1038/483134a.
51. Michelle Brook, Peter Murray-Rust, and Charles Oppenheim, "The Social, Political and Legal Aspects of Text and Data Mining (TDM)," *D-Lib Magazine* 20, no. 11/12 (November 2014), https://doi.org/10.1045/november14-brook.
52. Eric Kansa, "Openness and Archaeology's Information Ecosystem," *World Archaeology* 44, no. 4 (December 2012): 498–520, https://doi.org/10.1080/00438243.2012.737575.
53. Cameron Neylon, "Best Practice in Enabling Content Mining," *PLOS Opens* (blog), March 9, 2014, accessed July

31, 2017, http://blogs.plos.org/opens/2014/03/09/best-practice-enabling-content-mining/ (page discontinued).

54. Michael L. Black, “The World Wide Web as Complex Data Set: Expanding the Digital Humanities into the Twentieth Century and Beyond through Internet Research,” *International Journal of Humanities and Arts Computing* 10, no. 1 (March 2016): 95–109, https://doi.org/10.3366/ijhac.2016.0162.
55. Jane Smith and Eric Hartnett, “The Licensing Lifecycle: From Negotiation to Compliance,” *Serials Librarian* 68, no. 1–4 (January 2015): 205–14, https://doi.org/10.1080/0361526X.2015.1017707.
56. Leslie A. Williams et al., “Negotiating a Text Mining License for Faculty Researchers,” *Information Technology and Libraries* 33, no. 3 (September 2014): 5–21, https://doi.org/10.6017/ital.v33i3.5485; Orcutt, “Library Support for Text and Data Mining”; Miller, “Securing Text and Data Mining Rights for Researchers in Academic Libraries.”
57. Tim Bowen et al., “Using Computing Power to Replace Lawyers: Advances in Licensing and Access,” *Serials Librarian* 66, no. 1–4 (January 2014): 232–40, https://doi.org/10.1080/0361526X.2014.881221.
58. Smith and Hartnett, “The Licensing Lifecycle.”
59. Ann Okerson, “Text and Data Mining—A Librarian Overview” (paper presented at the International Federated Library Association’s World Library and Information Congress, Singapore, August 2013), http://library.ifla.org/252/. Though outside the scope of this paper, it is also worth noting that protecting patron privacy is a major concern for librarians and often a point of tension during conversations with vendors who would like to use detailed patron use data for mining projects of their own (Orcutt, “Library Support for Text and Data Mining”).
60. California Digital Library, “CDL Model License,” February 13, 2019, https://cdlib.org/wp-content/uploads/2019/02/CDL_Model_License_2018.05.14_public.docx ; Center for Research Libraries, “Model Licenses,” LIBLICENSE: Licensing Digital Content, accessed September 14, 2019, http://liblicense.crl.edu/licensing-information/model-license/; Canadian Research Knowledge Network, “CRKN Model License,” accessed September 14, 2019, https://www.crkn-rcdr.ca/sites/crkn/files/2016-08/crkn_model_license_2016_final.pdf.
61. Charles B. Lowry and Julia C. Blixrud, “E-book Licensing and Research Libraries—Negotiating Principles and Price in an Emerging Market,” *Research Library Issues,* no. 280 (September 2012): 12, https://doi.org/10.29242/rli.280.3.
62. Sharon Dyas-Correia and Michelle Alexopoulos, “Text and Data Mining: Searching for Buried Treasures,” *Serials Review* 40, no. 3 (September 2014): 210–16, https://doi.org/10.1080/00987913.2014.950041; Miller, “Securing Text and Data Mining Rights for Researchers in Academic Libraries.”
63. Miller, “Securing Text and Data Mining Rights for Researchers in Academic Libraries.”
64. Okerson, “Text and Data Mining—A Librarian Overview”; Williams et al., “Negotiating a Text Mining License for Faculty Researchers.”
65. Bernard F. Reilly, “When Machines Do Research: Automated Analysis of News and Other Primary Source Texts,” special issue, “Redefining Relevance: Exceeding User Expectations in a Digital Age,” *Journal of Library Administration* 49, no. 5 (July 2009): 507–17, https://doi.org/10.1080/01930820903090888.
66. Orcutt, “Library Support for Text and Data Mining.”
67. Clifford B. Anderson and Hilary A. Craiglow, “Text Mining in Business Libraries,” *Journal of Business and Finance Librarianship* 22, no. 2 (2017): 149–65, https://doi.org/10.1080/08963568.2017.1285749.
68. Dyas-Correia and Alexopoulos, “Text and Data Mining.”
69. Christa Williford and Charles Henry, *One Culture,* CLIR Publication 151 (Arlington, VA: Council on Library and Information Resources, June 2012), https://www.clir.org/pubs/reports/pub151/.
70. Williams et al., “Negotiating a Text Mining License for Faculty Researchers”; Dyas-Correia and Alexopoulos, “Text and Data Mining”; Anderson and Craiglow, “Text Mining in Business Libraries.”
71. Bernhardt et al., “Revolutionizing Scholarship.”
72. Beth Plale, “Big Data Opportunities and Challenges for IR, Text Mining and NLP,” in *Unstructure NLP ’13: Proceedings of the 2013 International Workshop on Mining Unstructured Big Data Using Natural Language Processing* (New York: ACM, 2013), https://doi.org/10.1145/2513549.2514739; J. Stephen Downie, “The HathiTrust Research Center: Providing Analytic Access to the HathiTrust Digital Library’s 4.7 Billion Pages,” in *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries* (New York: ACM, 2015), 5, https://doi.org/10.1145/2756406.2771494.
73. HathiTrust Research Center Task Force for Non-consumptive Research Use Policy, “Non-consumptive Use Research Policy,” February 20, 2017, https://www.hathitrust.org/htrc_ncup.
74. See, for example, Guangchen Ruan et al., “TextRWeb: Large-Scale Text Analytics with R on the Web,” in *Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment* (New York: ACM, 2014), https://doi.org/10.1145/2616498.2616557; Jiaan Zeng et al., “Cloud Computing Data Capsules for Non-consumptive Use of Texts,” in *ScienceCloud ’14: Proceedings of the 5th ACM Workshop on Scientific Cloud Computing,* 9–16 (New York: ACM, 2014), https://doi.org/10.1145/2608029.2608031.

75. Peilin Yang et al., “Towards Privacy-Preserving Evaluation for Information Retrieval Models over Industry Data Sets,” in *Information Retrieval Technology: 13th Asia Information Retrieval Societies Conference, AIRS 2017, Jeju Island, South Korea, November 22–24, 2017, Proceedings*, ed. Won-Kyung Sung, Hanmin Jung, Shuo Xu, Krisana Chinnasarn, Kazutoshi Sumiya, Jeonghoon Lee, Zhicheng Dou, et al. (Cham, Switzerland: Springer, 2017), 210–21, https://doi.org/10.1007/978-3-319-70145-5_16.
76. Jaimie Murdock et al., “Towards Publishing Secure Capsule-Based Analysis,” in *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (Piscataway, NJ: IEEE, 2017), 261–64, https://doi.org/10.1109/JCDL.2017.7991585.
77. See, for example, the Whole Tale project (https://wholetale.org/), the Binder platform (https://mybinder.org/), and the Code Ocean repository (https://codeocean.com/).
78 Michael W. Carroll, “Why Full Open Access Matters.” *PLOS Biology* 9, no. 11 (November 29, 2011), e1001210, https://doi.org/10.1371/journal.pbio.1001210; Kansa, “Openness and Archaeology’s Information Ecosystem.”
79. Brook, Murray-Rust, and Oppenheim, “The Social, Political and Legal Aspects of Text and Data Mining (TDM).”
80. Iain Hrynaszkiewicz and Matthew J. Cockerill, “Open by Default: A Proposed Copyright License and Waiver Agreement for Open Access Research and Data in Peer-Reviewed Journals,” *BMC Research Notes* 5 (2012), article 494, https://doi.org/10.1186/1756-0500-5-494.
81. Patricia Cleary et al., “Text Mining 101: What You Should Know,” *Serials Librarian* 72, no. 1–4 (2017): 156–59, https://doi.org/10.1080/0361526X.2017.1320876; Bernhardt et al., “Revolutionizing Scholarship.”
82. Jean-Baptiste Michel et al., “Quantitative Analysis of Culture Using Millions of Digitized Books,” *Science* 331, no. 6014 (January 2011): 176–182, https://doi.org/10.1126/science.1199644.
83. Allison, Sarah. “Other People’s Data: Humanities Edition.” *Journal of Cultural Analytics*, December 8, 2016. https://doi.org/10.22148/001c.11822.
84. Boris Capitanu et al., *The HathiTrust Research Center Extracted Feature Dataset (1.0),* dataset, HathiTrust Research Center, 2016, last modified June 21, 2019, http://dx.doi.org/10.13012/J8X63JT3.
85. Reilly, “When Machines Do Research: Automated Analysis of News and Other Primary Source Texts.”
86. For example, Jill Emery, “Working in a Text Mine: Is Access about to Go Down?” *Journal of Electronic Resources Librarianship* 20, no. 3 (2008): 135–38, https://doi.org/10.1080/19411260802412745.
87. Cleary et al., “Text Mining 101.”
88. Williford and Henry, *One Culture*, 4.
89. Jorge L. Contreras and Jerome H. Reichman, “Sharing by Design: Data and Decentralized Commons,” *Science* 350, no. 6266 (December 11, 2015): 1314, https://doi.org/10.1126/science.aad8071.
90. To read our position paper and the attendees’ pre-forum statements, see: Eleanor Dickson, Megan Senseney, Beth Namachchivaya, and Bertram Ludäscher, “IMLS National Forum on Data Mining Research Using In-Copyright and Limited-Access Text Datasets: Discussion Paper, Forum Statements, and SWOT Analyses,” Center for Informatics Research in Science and Scholarship, University of Illinois at Urbana-Champaign, May 24, 2018, http://hdl.handle.net/2142/100055.
91. Mark D. Wilkinson et al., “The FAIR Guiding Principles for Scientific Data Management and Stewardship,” *Scientific Data* 3 (March 15, 2016), article 160018, https://doi.org/10.1038/sdata.2016.18.
92. For one attendee’s perspective on this activity, see: Andrew Piper, “Where’s the Data? Notes from an International Forum on Limited Use Text Mining,” .txtLAB, April 10, 2018, https://txtlab.org/2018/04/wheres-the-data-notes-from-an-international-forum-on-limited-use-text-mining/.
93. In its model license language, the California Digital Library includes a section entitled “No Diminution of Rights” to address fair use: California Digital Library, “CDL Model License.”
94. McDonald and Kelly, *Value and Benefits of Text Mining*.
95. Authors Guild v. HathiTrust 755 F.3d 87 (2d Cir. 2014); Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015).
96. Association of Research Libraries, “Code of Best Practices in Fair Use for Academic and Research Libraries: Designing the Public Domain,” January 2012, from “Designing the Public Domain,” *Harvard Law Review* 122 (2008–2009): 1489–1510, http://www.arl.org/focus-areas/copyright-ip/fair-use/code-of-best-practices.
97. “Horizon 2020,” European Commission, accessed December 5, 2018, https://ec.europa.eu/programmes/horizon2020/en/; FutureTDM home page, accessed December 5, 2018, https://www.futuretdm.eu/.
98. Ivy Anderson et al., “Standing Together before We Fall Apart: Solving the Content as Data Problem” (lively discussion, Charleston Library Conference, Charleston, SC, November 8, 2018), https://sched.co/GB39.
99. “The Hague Declaration on Knowledge Discovery in the Digital Age,” Hague Declaration, archived July 30, 2019, https://wayback.archive-it.org/12503/20190730165131/https://thehaguedeclaration.com/the-hague-declaration-on-knowledge-discovery-in-the-digital-age.
100. One participant expanded his initial forum statement into a full article, which has been deposited in SSRN and is

under review for publication: Matthew Sag, "The New Legal Landscape for Text Mining and Machine Learning," unpublished paper, February 9, 2019, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3331606.

101. NASIG includes TDM as part of its core competencies for scholarly communication librarians: NASIG, *NASIG Core Competencies for Scholarly Communications Librarians* (West Seneca, NY: NASIG, August 11, 2017), https://www.nasig.org/Competencies-Scholarly-Communication.

102. While data journals and accompanying data papers are increasingly common across multiple disciplines (Leonardo Candela et al., "Data Journals: A Survey," *Journal of the Association for Information Science and Technology* 66, no. 9 [2015]: 1747–62, https://doi.org/10.1002/asi.23358), there are few provisions in place for characterizing the in-depth methodological processes required for sharing and sufficiently contextualizing the data derived from an initial dataset for which redistribution is restricted. *Cultural Analytics*' new section of articles on corpus creation (each running 3,000–5,000 words) may serve as a promising model for documenting methods alongside derived datasets (See, for example, the Data Sets section of the *Journal of Cultural Analytics*: https://culturalanalytics.org/section/1579-data-sets).

103. See, for example, the Whole Tale project (https://wholetale.org/).

104. Wim Westera, *The Digital Turn* (Bloomington, IN: AuthorHouse, 2012).

BIBLIOGRAPHY

** Included in literature review and cited*

† Included in literature review but not cited

‡ Cited but not included in literature review

‡ All European Academies (ALLEA). *The Text and Data Mining Exception and the Enhancement of Access to Scientific Information in Europe*. Berlin: ALLEA, November 2017. https://www.knaw.nl/shared/resources/internationaal/bestanden/allea_pwgipr_statement_tdm_2017.

* Allison, Sarah. "Other People's Data: Humanities Edition." *Journal of Cultural Analytics*, December 8, 2016. https://doi.org/10.22148/001c.11822.

‡ Always Already Computational—Collections as Data. "The Santa Barbara Statement on Collections as Data," version 2. 2017. https://collectionsasdata.github.io/.

* Anderson, Clifford B., and Hilary A. Craiglow. "Text Mining in Business Libraries." *Journal of Business and Finance Librarianship* 22, no. 2 (2017): 149–65. https://doi.org/10.1080/08963568.2017.1285749.

‡ Anderson, Ivy, Marc Cormier, Mike Furlough, Darby Orcutt, Megan Senseney, and Kate Wittenberg. "Standing Together before We Fall Apart: Solving the Content as Data Problem." Lively discussion at the Charleston Library Conference, Charleston, SC, November 8, 2018. https://sched.co/GB39.

* Anderson, Ivy, Kenneth Crews, Roy Kaufman, and William Maher. "What Should Be the Conditions on Libraries Digitizing, Maintaining and Making Available Copyrighted Works." *Columbia Journal of Law and the Arts* 36, no. 4 (2013): 587–606. https://heinonline.org/HOL/P?h=hein.journals/cjla36&i=603.

† Arrow, Tom, Jenny Molloy, and Peter Murray-Rust. "A Day in the Life of *a Content Miner and Team*." *Insights* 29, no. 2 (July 5, 2016). https://doi.org/10.1629/uksg.310.

‡ Association of Research Libraries. "Code of Best Practices in Fair Use for Academic and Research Libraries: Designing the Public Domain." January 2012. From "Designing the Public Domain," *Harvard Law Review* 122 (2008–2009): 1489–1510. http://www.arl.org/focus-areas/copyright-ip/fair-use/code-of-best-practices.

‡ Association of Research Libraries. "Digital Scholarship Profiles." Accessed September 20, 2019. https://www.arl.org/focus-areas/scholarly-communication/digital-scholarship.

‡ Association of Research Libraries. *Text and Data Mining and Fair Use in the United States.* Issue brief. Washington, DC: Association of Research Libraries, June 2015. https://www.arl.org/wp-content/uploads/2015/06/TDM-5JUNE2015.pdf.

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

‡ Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015).

‡ Authors Guild v. HathiTrust 755 F.3d 87 (2d Cir. 2014).

* Band, Jonathan. "What Does the HathiTrust Decision Mean for Libraries?" *Research Library Issues* 285 (January 2015): 7–13. https://doi.org/10.29242/rli.285.3.

† Bell, Abraham, and Gideon Parchomovsky. "The Dual-Grant Theory of Fair Use." Faculty Scholarship at Penn Law 1594. From *University of Chicago Law Review* 83, no. 3 (Summer 2016): 1051–118. https://scholarship.law.upenn.edu/faculty_scholarship/1594.

‡ Bergman, Casey M., Lawrence E. Hunter, and Andrey Rzhetsky. "Announcing the PLOS Text Mining Collection." *PLOS ONE Community Blog*, April 17, 2013. https://blogs.plos.org/everyone/2013/04/17/announcing-the-plos-text-mining-collection/.

‡ Berkeley Library Scholarly Communication Services. "Text Mining." Accessed December 3, 2018. http://www.lib.berkeley.edu/scholarly-communication/publishing/copyright/text-mining.

* Bernhardt, Beth, Joel Herndon, Patrick Herron, Kevin Smith, Roger Strong, and Hillary Miller. "Revolutionizing Scholarship: A Panel Discussion on Text and Data Mining." *Serials Review* 41, no. 3 (July 2015): 184–86. https://

doi.org/10.1080/00987913.2015.1064514.

* Black, Michael L. "The World Wide Web as Complex Data Set: Expanding the Digital Humanities into the Twentieth Century and Beyond through Internet Research." *International Journal of Humanities and Arts Computing* 10, no. 1 (March 2016): 95–109. https://doi.org/10.3366/ijhac.2016.0162.

* Borghi, Maurizio, and Stavroula Karapapa. "Non-display Uses of Copyright Works: Google Books and Beyond." *Queen Mary Journal of Intellectual Property* 1, no. 1 (April 2011). https://dx.doi.org/10.2139/ssrn.2358912.

† Borgman, Christine L. "The Digital Future Is Now: A Call to Action for the Humanities." *Digital Humanities Quarterly* 3, no. 4 (2009). http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html.

* Bowen, Tim, Mimi Calter, Franny Lee, and Elizabeth Parang. "Using Computing Power to Replace Lawyers: Advances in Licensing and Access." *Serials Librarian* 66, no. 1–4 (January 2014): 232–40. https://doi.org/10.1080/0361526X.2014.881221.

* Brook, Michelle, Peter Murray-Rust, and Charles Oppenheim. "The Social, Political and Legal Aspects of Text and Data Mining (TDM)." *D-Lib Magazine* 20, no. 11/12 (November 2014). https://doi.org/10.1045/november14-brook.

* Butler, Brandon C. "Fair Use Rising: Full-Text Access and Repurposing in Recent Case Law." *Research Library Issues,* no. 285 (2015): 3–6. https://doi.org/10.29242/rli.285.2.

* ———. "Transformative Teaching and Educational Fair Use after Georgia State." *Connecticut Law Review* 48, no. 2 (2015). https://docs.wixstatic.com/ugd/17f3ef_48f887e2d77a4848bda319e2ccf57fdc.pdf.

‡ California Digital Library. "CDL Model License." February 13, 2019. https://cdlib.org/wp-content/uploads/2019/02/CDL_Model_License_2018.05.14_public.docx.

‡ Canadian Research Knowledge Network. "CRKN Model License." Accessed September 14, 2019. https://www.crkn-rcdr.ca/sites/crkn/files/2016-08/crkn_model_license_2016_final.pdf.

‡ Canadian Research Knowledge Network. "Model License." Accessed September 20, 2019. http://www.crkn-rcdr.ca/en/model-license.

‡ Candela, Leonardo, Donatella Castelli, Paolo Manghi, and Alice Tani. "Data Journals: A Survey." *Journal of the Association for Information Science and Technology* 66, no. 9 (2015): 1747–62. https://doi.org/10.1002/asi.23358.

‡ Capitanu, Boris, Ted Underwood, Peter Organisciak, Timothy Cole, Maria Janina Sarol, and J. Stephen Downie. *The HathiTrust Research Center Extracted Feature Dataset (1.0).* Dataset. HathiTrust Research Center, 2016, last modified June 21, 2019. http://dx.doi.org/10.13012/J8X63JT3.

† Carroll, Michael W. "Sharing Research Data and Intellectual Property Law: A Primer." *PLOS Biology* 13, no. 8 (2015): 31002235. https://doi.org/10.1371/journal.pbio.1002235.

* ———. "Why Full Open Access Matters." *PLOS Biology* 9, no. 11 (November 29, 2011), e1001210. https://doi.org/10.1371/journal.pbio.1001210.

* Caspers, Marco, and Lucie Guibault. "A Right to 'Read' for Machines: Assessing a Black-Box Analysis Exception for Data Mining." *Proceedings of the Association for Information Science and Technology* 53, no. 1 (December 27, 2016): 1–5. https://doi.org/10.1002/pra2.2016.14505301017.

‡ Center for Research Libraries. "Model Licenses." LIBLICENSE: Licensing Digital Content. Accessed September 14, 2019. http://liblicense.crl.edu/licensing-information/model-license/.

† Charlesworth, Andrew. "Digital Curation, Copyright, and Academic Research." *International Journal of Digital Curation* 1, no. 1 (2006): 17–32. https://doi.org/10.2218/ijdc.v1i1.3.

* Clark, Jonathan. *Text Mining and Scholarly Publishing.* Publishing Research Consortium, 2013. http://publishingresearchconsortium.com/index.php/prc-guides-main-menu/158-prc-guide-text-mining-and-scholarly-publishing.

* Cleary, Patricia, Kristen Garlock, Denise Novak, Ethan Pullman, and Sanjeet Mann. "Text Mining 101: What You Should Know." *Serials Librarian* 72, no. 1–4 (2017): 156–59. https://doi.org/10.1080/0361526X.2017.1320876.

* Cohen, Aaron M., and William R. Hersh. "A Survey of Current Work in Biomedical Text Mining." *Briefings in Bioinformatics* 6, no. 1 (2005): 57–71. https://www.ncbi.nlm.nih.gov/pubmed/15826357.

* Colonna, Liane. "A Taxonomy and Classification of Data Mining." *Southern Methodist University Science and*

Technology Law Review 16, no. 2 (Fall 2013): 309–70. https://scholar.smu.edu/scitech/vol16/iss2/4/.

‡ ContentMine home page. Accessed December 3, 2018. http://contentmine.org/.

* Contreras, Jorge L., and Jerome H. Reichman. "Sharing by Design: Data and Decentralized Commons." *Science* 350, no. 6266 (December 11, 2015): 1310–12. https://doi.org/10.1126/science.aad8071.

‡ Copyright.gov. "More Information on Fair Use." Accessed September 20, 2019. https://www.copyright.gov/fair-use/more-info.html.

‡ Corbin, Juliet M., and Anselm L. Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*, 3rd ed. Los Angeles: Sage, 2008.

* Cox, Krista. "International Copyright Developments: From the Marrakesh Treaty to Trade Agreements." *Research Library Issues,* no. 285 (2015): 14–22. https://doi.org/10.29242/rli.285.4.

‡ Creative Commons home page. Accessed September 20, 2019. https://creativecommons.org/.

‡ Currier, Brett. "You May Be Sued …or Arrested." *Intersections* (blog), Scholarly Communications at the University of North Carolina, Chapel Hill, April 30, 2015. https://blogs.lib.unc.edu/intersections/index.php/2015/04/30/you-may-be-sued-or-arrested/.

† De Guyter, Elise. "Inside Views: The Dilemma of Fair Use and Expressive Machine Learning: An Interview with Ben Sobel." *Intellectual Property Watch* (blog), August 23, 2017. http://www.ip-watch.org/2017/08/23/dilemma-fair-use-expressive-machine-learning-interview-ben-sobel/.

‡ Denker, Sheryl P. "Unrestricted Text and Data Mining with allofPLOS." *The Official PLOS Blog*, November 28, 2017. https://blogs.plos.org/plos/2017/11/unrestricted-text-and-data-mining-with-allofplos/.

* Diaz, Angel Siegfried. "Fair Use and Mass Digitization: The Future of Copy-Dependent Technologies after Authors Guild v. HathiTrust." *Berkeley Technology Law Journal* 28 (2013). https://doi.org/10.15779/Z38X700.

‡ Dickson, Eleanor, Megan Senseney, Beth Namachchivaya, and Bertram Ludäscher. "IMLS National Forum on Data Mining Research Using In-Copyright and Limited-Access Text Datasets: Discussion Paper, Forum Statements, and SWOT Analyses." Center for Informatics Research in Science and Scholarship, University of Illinois at Urbana-Champaign, May 24, 2018. http://hdl.handle.net/2142/100055.

* Dillon, Cy. "Transformative Use: An Update on the Google Books Case with Jonathan Band." *Virginia Libraries* 60, no. 2 (August 1, 2014). http://doi.org/10.21061/valib.v60i2.1296.

* Downie, J. Stephen. "The HathiTrust Research Center: Providing Analytic Access to the HathiTrust Digital Library's 4.7 Billion Pages." In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, 5. New York: ACM, 2015. https://doi.org/10.1145/2756406.2771494.

† Dunning, Alastair, Ian Gregory, and Andrew Hardie. "Freeing Up Digital Content with Text Mining: New Research Means New Licences." *Serials* 22, no. 2 (2009): 166–73. https://doi.org/10.1629/22166.

* Dyas-Correia, Sharon, and Michelle Alexopoulos. "Text and Data Mining: Searching for Buried Treasures." *Serials Review* 40, no. 3 (September 2014): 210–16. https://doi.org/10.1080/00987913.2014.950041.

* Elkin-Koren, Niva, and Orit Fischman-Afori. "Rulifying Fair Use." *Arizona Law Review* 59, no. 1 (2017): 161–200. http://arizonalawreview.org/rulifying-fair-use/.

‡ European Commission. "Horizon 2020." Accessed December 5, 2018. https://ec.europa.eu/programmes/horizon2020/en/.

‡ Fair Dealing ©anada. "Learn More about Fair Dealing." Accessed September 20, 2019. https://fair-dealing.ca/.

‡ FutureTDM home page. Accessed December 5, 2018. https://www.futuretdm.eu/.

‡ Gandomi, Amir, and Murtaza Haidar. "Beyond the Hype: Big Data Concepts, Methods, and Analytics." *International Journal of Information Management* 35, no. 2 (April 2015): 137–44. https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

* Emery, Jill. "Working in a Text Mine: Is Access about to Go Down?" *Journal of Electronic Resources Librarianship* 20, no. 3 (2008): 135–38. https://doi.org/10.1080/19411260802412745.

† Garfinkel, Simson, Paul Farrell, Vassil Roussev, and George Dinolt. "Bringing Science to Digital Forensics with Standardized Forensic Corpora." *Digital Investigation* 6, supplement (September 2009): S2–S11. https://doi.org/10.1016/j.diin.2009.06.016.

* Ginsburg, Jane C. "Fair Use for Free, or Permitted-but-Paid." *Berkeley Technology Law Journal* 29, no. 3 (2015): 1383–446. https://doi.org/10.15779/Z38GW14.

‡ GNU Operating System. "GNU General Public License," version 3. June 20, 2007. https://www.gnu.org/licenses/gpl.html.

† Grimmelmann, James. "Copyright for Literate Robots." *Iowa Law Review* 101, no. 2 (2016): 657–81. https://ilr.law.uiowa.edu/print/volume-101-issue-2/copyright-for-literate-robots/.

* Guadamuz, Andrés, and Diane Cabell. "Data Mining in UK Higher Education Institutions: Law and Policy." *Queen Mary Intellectual Property Review* 4, no. 1 (2014): 3–29. https://papers.ssrn.com/abstract=2446447.

* Guiliano, Jennifer, and Mia Ridge. "The Future of Digital Methods for Complex Datasets: An Introduction." *International Journal of Humanities and Arts Computing: A Journal of Digital Humanities* 10, no. 1 (March 2016): 1–7. https://doi.org/10.3366/ijhac.2016.0155.

‡ Hague Declaration. "The Hague Declaration on Knowledge Discovery in the Digital Age." Archived July 30, 2019. https://wayback.archive-it.org/12503/20190730165131/https://thehaguedeclaration.com/the-hague-declaration-on-knowledge-discovery-in-the-digital-age/.

† Hansen, David R. "Copyright Reform Principles for Libraries, Archives, and Other Memory Institutions." *Berkeley Technology Law Journal* 29, no. 3 (2015): 1559–94. https://doi.org/10.15779/Z38ZZ91.

* Hansen, David R., Kathryn Hashimoto, Gwen Hinze, Pamela Samuelson, and Jennifer M. Urban. "Solving the Orphan Works Problem for the United States." *Columbia Journal of Law and the Arts* 37, no. 1 (2013): 1–55. https://doi.org/10.7916/jla.v37i1.2145.

* Hargreaves, Ian. *Digital Opportunity: Review of Intellectual Property and Growth: An Independent Report.* Newport, South Wales, UK: Intellectual Property Office, 2011. https://www.gov.uk/government/publications/digital-opportunity-review-of-intellectual-property-and-growth.

* Hargreaves, Ian, Lucie Guibault, Christian Handke, Peggy Valcke, and Bertin Martens. *Standardisation in the Area of Innovation and Technological Development, Notably in the Field of Text and Data Mining: Report from the Expert Group*. Luxembourg: Publications Office of the European Union, 2014. https://doi.org/10.2777/71122.

‡ HathiTrust Research Center. "Digging Deeper, Reaching Further: Libraries Empowering Users to Mine the HathiTrust Digital Library Resources." Accessed December 3, 2018. https://teach.htrc.illinois.edu/.

‡ HathiTrust Research Center Task Force for Non-consumptive Research Use Policy. "Non-consumptive Use Research Policy." February 20, 2017. https://www.hathitrust.org/htrc_ncup.

‡ Hearst, Marti. "What Is Text Mining?" Unpublished paper, October 17, 2003. http://people.ischool.berkeley.edu/~hearst/text-mining.html.

‡ Hosoi, Mihoko. "CDL Model License Revised." California Digital Library, January 25, 2017. https://www.cdlib.org/cdlinfo/2017/01/25/cdl-model-license-revised/.

* Hrynaszkiewicz, Iain, and Matthew J. Cockerill. "Open by Default: A Proposed Copyright License and Waiver Agreement for Open Access Research and Data in Peer-Reviewed Journals." *BMC Research Notes* 5 (2012), article 494. https://doi.org/10.1186/1756-0500-5-494.

‡ Hsieh, Hsiu-Fang, and Sarah E. Shannon. «Three Approaches to Qualitative Content Analysis.» *Qualitative Health Research* 15, no. 9 (2005): 1277–88. https://doi.org/10.1177/1049732305276687.

‡ International Federation of Library Associations. *IFLA Statement on Text and Data Mining.* The Hague, Netherlands: International Federation of Library Associations, 2013. https://www.ifla.org/publications/node/8225.

† Jenkins, Jennifer. "In Ambiguous Battle: The Promise (and Pathos) of Public Domain Day, 2014." *Duke Law and Technology Review* 12, no. 1 (2013): 1–24. https://scholarship.law.duke.edu/dltr/vol12/iss1/1.

* Jockers, Matthew L., Matthew Sag, and Jason Schultz. Brief of Digital Humanities and Law Scholars as Amici Curiae in Authors Guild v. Google (August 3, 2012). http://dx.doi.org/10.2139/ssrn.2102542.

† ———. "Digital Archives: Don't Let Copyright Block Data Mining." *Nature* 490, no. 7418 (October 4, 2012): 29–30. https://doi.org/10.1038/490029a.

* Jockers, Matthew L., and Ted Underwood. "Text-Mining the Humanities." In *A New Companion to Digital Humanities.* Edited by Susan Schreibman, Raymond G. Siemens, and John Unsworth, 291–306. Malden, MA: John Wiley

and Sons, 2015. https://doi.org/10.1002/9781118680605.ch20.

* Jovic, Alan, Karla Brkic, and Nikola Bogunovic. "An Overview of Free Software Tools for General Data Mining." In *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO): Proceedings.* Edited by Petar Biljanovic, Zeljko Butkovic, Karolj Skala, Stjepan Golubic, Marina Cicin-Sain, Vlado Sruk, Slobodan Ribaric, et al., 1112–17. Rijeka, Croatia: Croatian Society for Information and Communication Technology, Electronics and Microelectronics—MIPRO, 2014. https://doi.org/10.1109/MIPRO.2014.6859735.

* Kansa, Eric. "Openness and Archaeology's Information Ecosystem." *World Archaeology* 44, no. 4 (December 2012): 498–520. https://doi.org/10.1080/00438243.2012.737575.

* Kaufman, Peter B., and Jeff Ubois. "Good Terms—Improving Commercial-Noncommercial Partnerships for Mass Digitization: A Report Prepared by Intelligent Television for RLG Programs, OCLC Programs and Research." *D-Lib Magazine* 13, no. 11/12 (November 2007). https://doi.org/10.1045/november2007-kaufman.

‡ Kroll, Susan, and Rick Forsman. *A Slice of Research Life: Information Support for Research in the United States.* Dublin, OH: OCLC Research, June 2010. https://www.oclc.org/content/dam/research/publications/library/2010/2010-15.pdf.

* Lammey, Rachael. "CrossRef's Text and Data Mining Services." *Learned Publishing* 27, no. 4 (October 2014): 245–50. https://doi.org/10.1087/20140402.

‡ Legal Information Institute. "17 U.S. Code §512. Limitations on Liability Relating to Material Online." Cornell Law School. Accessed September 20, 2019. https://www.law.cornell.edu/uscode/text/17/512.

‡ Legal Information Institute. "18 U.S. Code §1030. Fraud and Related Activity in Connection with Computers." Cornell Law School. Accessed September 20, 2019. https://www.law.cornell.edu/uscode/text/18/1030.

* Li, Yanpeng, Xiaohua Hu, Hongfei Lin, and Zhiahi Yang. "A Framework for Semisupervised Feature Generation and Its Applications in Biomedical Literature Mining." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 8, no. 2 (March 2011): 294–307. https://doi.org/10.1109/TCBB.2010.99.

‡ Ligue des Bibliothèques Européennes de Recherche (LIBER). "Text and Data Mining." Accessed September 14, 2019. https://libereurope.eu/text-data-mining/.

‡ Library of Congress. *Collections as Data: Stewardship and Use Models to Enhance Access*, executive summary. Accessed September 20, 2019. http://digitalpreservation.gov/meetings/dcs16/AsDataExecutiveSummary_final.pdf.

‡ Lipmanowicz, Henri, and Keith McCandless. Liberating Structures home page. Accessed December 3, 2018. http://www.liberatingstructures.com/.

* Lowry, Charles B., and Julia C. Blixrud. "E-Book Licensing and Research Libraries—Negotiating Principles and Price in an Emerging Market." *Research Library Issues,* no. 280 (September 2012): 11–19. https://doi.org/10.29242/rli.280.3.

† Matulionyte, Rita. "10 Years for Google Books and Europeana: Copyright Law Lessons That the EU Could Learn from the USA." *International Journal of Law and Information Technology* 24, no. 1 (Spring 2016): 44–71. https://doi.org/10.1093/ijlit/eav018.

* McDonald, Diane, and Ursula Kelly. *Value and Benefits of Text Mining.* London: Jisc, March 2012, upd. August 5, 2019. https://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining.

* Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, et al. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331, no. 6014 (January 2011): 176–82. https://doi.org/10.1126/science.1199644.

* Miller, Hillary. "Securing Text and Data Mining Rights for Researchers in Academic Libraries." Master's thesis, University of North Carolina, 2015. https://cdr.lib.unc.edu/record/uuid:704c0c1e-e103-4242-85d7-d3abf5b25835.

* Molloy, Jennifer, Maximilian Haeussler, Peter Murray-Rust, and Charles Oppenheim. "Responsible Content Mining." In *Working with Text: Tools, Techniques, and Approaches for Text Mining.* Edited by Emma L. Tonkin and Gregory J. L. Tourte, 89–109. Cambridge, MA: Chandos, 2016. https://doi.org/10.1016/B978-1-84334-749-1.00004-4.

* Murdock, Jaimie, Jacob Jett, Tim Cole, Yu Ma, J. Stephen Downie, and Beth Plale. "Towards Publishing Secure Capsule-Based Analysis." In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 261–64. Piscataway, NJ: IEEE, 2017. https://doi.org/10.1109/JCDL.2017.7991585.

* Myška, Matěj. "Text and Data Mining of Grey Literature for the Purpose of Scientific Research." *Grey Journal* 13 (January 2, 2017): 32–37.

‡ NASIG. *NASIG Core Competencies for Scholarly Communications Librarians.* West Seneca, NY: NASIG, August 11, 2017. www.nasig.org/Competencies-Scholarly-Communication.

† Neugebauer, Tomasz, and Annie Murray. "The Critical Role of Institutional Services in Open Access Advocacy." *International Journal of Digital Curation* 8, no. 1 (2013): 84–106. https://doi.org/10.2218/ijdc.v8i1.238.

* Neylon, Cameron. "Best Practice in Enabling Content Mining." *PLOS Opens* (blog), March 9, 2014. Accessed July 31, 2017. http://blogs.plos.org/opens/2014/03/09/best-practice-enabling-content-mining/ (page discontinued).

* Okerson, Ann. "Text and Data Mining—A Librarian Overview." Paper presented at the International Federated Library Association's World Library and Information Congress, Singapore, August 2013. http://library.ifla.org/252/.

* Orcutt, Darby. "Library Support for Text and Data Mining." *Online Searcher* 39, no. 3 (June 5, 2015): 27–30.

‡ Peng, Roger D. "Reproducible Research in Computational Science." *Science* 334, no. 6060 (December 2, 2011): 1226–27. https://doi.org/10.1126/science.1213847.

‡ Penn State University. "Text Mining: Web-Based Resources." Accessed September 14, 2019. https://guides.libraries.psu.edu/textmining/web.

‡ Piper, Andrew. "Where's the Data? Notes from an International Forum on Limited Use Text Mining." .txtLAB, April 10, 2018. https://txtlab.org/2018/04/wheres-the-data-notes-from-an-international-forum-on-limited-use-text-mining/.

* Plale, Beth. "Big Data Opportunities and Challenges for IR, Text Mining and NLP." In *Unstructure NLP '13: Proceedings of the 2013 International Workshop on Mining Unstructured Big Data Using Natural Language Processing*. New York: ACM, 2013. https://doi.org/10.1145/2513549.2514739.

* Rathemacher, Andrée J. "Developing Issues in Licensing: Text Mining, MOOCs, and More." *Serials Review* 39, no. 3 (September 1, 2013): 205–10. https://doi.org/10.1080/00987913.2013.10766397.

* Reichman, Jerome H., and Ruth L. Okediji. "When Copyright Law and Science Collide: Empowering Digitally Integrated Research Methods on a Global Scale." *Minnesota Law Review* 96, no. 4 (April 2012): 1362–1480. http://www.minnesotalawreview.org/wp-content/uploads/2012/08/ReichmanOkediji_MLR1362.pdf.

* Reichman, Jerome H., and Paul F. Uhlir. "A Contractually Reconstructed Research Commons for Scientific Data in a Highly Protectionist Intellectual Property Environment." *Law and Contemporary Problems* 66, no. 1/2 (2003): 315–462. https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=1283&context=lcp.

* Reilly, Bernard F. "When Machines Do Research: Automated Analysis of News and Other Primary Source Texts." Special issue, "Redefining Relevance: Exceeding User Expectations in a Digital Age." *Journal of Library Administration* 49, no. 5 (July 2009): 507–17. https://doi.org/10.1080/01930820903090888.

‡ ———. "When Machines Do Research, Part 2: Text-Mining and Libraries." *Charleston Advisor* 14, no. 2 (October 2012): 75–76. https://doi.org/10.5260/chara.14.2.75.

† Reilly, Susan. "Libraries at the Centre of the Debate on Copyright and Text and Data Mining: The LIBER Experience." Paper presented at the International Federated Library Association's World Library and Information Congress, Lyon, France, August 16–22, 2014. http://library.ifla.org/1007/.

† Reimer, Torsten. "Key Issue: Text Mining, Copyright and the Benefits and Barriers to Innovation." *Insights* 25, no. 2 (July 5, 2012). https://doi.org/10.1629/2048-7754.25.2.212.

* Ruan, Guangchen, Hui Zhang, Eric Wernert, and Beth Plale. "TextRWeb: Large-Scale Text Analytics with R on the Web." In *Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment*. New York: ACM, 2014. https://doi.org/10.1145/2616498.2616557.

* Sag, Matthew. "The Google Book Settlement and the Fair Use Counterfactual." *New York Law School Law Review* 55, no. 1 (2010/2011): 19–75. https://digitalcommons.nyls.edu/nyls_law_review/vol55/iss1/2/.

‡ ———. "The New Legal Landscape for Text Mining and Machine Learning." Unpublished paper, February 9, 2019.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3331606.

* ———. "Orphan Works as Grist for the Data Mill." *Berkeley Technology Law Journal* 27, no. 3 (2012), article 9. https://doi.org/10.15779/Z387M5B.

* Samuelson, Pamela. "The Quest for a Sound Conception of Copyright's Derivative Work Right." *Georgetown Law Journal* 101, no. 6 (2013): 1505–64. https://dx.doi.org/10.2139/ssrn.2138479.

‡ Sandvig v. Sessions, 315 F. Supp. 3d 1 (U.S.D.C. Dist. Columbia 2018).

* Sarkar, Indra N. "Biodiversity Informatics: Organizing and Linking Information across the Spectrum of Life." *Briefings in Bioinformatics* 8, no. 5 (September 2007): 347–57. https://doi.org/10.1093/bib/bbm037.

‡ Schwarcz, Andras. "Text and Data Mining: A New Service for Libraries?" *European Parliamentary Research Service Blog*, October 20, 2017. https://epthinktank.eu/2017/10/20/text-and-data-mining-a-new-service-for-libraries/.

‡ Shorish, Yasmeen. "Data Data Everywhere …but Do We Want to Drink?" *ACRL TechConnect* (blog), July 16, 2015. https://acrl.ala.org/techconnect/post/data-data-everywherebut-do-we-want-to-drink/.

‡ Sinclair, Stéfan, and Geoffrey Rockwell. Voyant Tools home page. Accessed December 3, 2018. https://voyant-tools.org/.

* Smith, Jane, and Eric Hartnett. "The Licensing Lifecycle: From Negotiation to Compliance." *Serials Librarian* 68, no. 1–4 (January 2015): 205–14. https://doi.org/10.1080/0361526X.2015.1017707.

‡ SPARC. "Open Access." Accessed September 20, 2019. https://sparcopen.org/open-access/.

* Stewart, Neil, Jane Secker, Chris Morrison, and Laurence Horton. "Liberating Data: How Libraries and Librarians Can Help Researchers with Text and Data Mining." *London School of Economics Impact Blog*, July 12, 2016. http://blogs.lse.ac.uk/impactofsocialsciences/2016/07/12/how-libraries-and-librarians-can-help-with-text-and-data-mining/.

† Surden, Harry. "Technological Cost as Law in Intellectual Property." *Harvard Journal of Law and Technology* 27, no. 1 (2013): 135–202. https://papers.ssrn.com/abstract=2383529.

* Terras, Melissa M. "The Potential and Problems in Using High Performance Computing in the Arts and Humanities: The Researching e-Science Analysis of Census Holdings (ReACH) Project." *Digital Humanities Quarterly* 3, no. 4 (2009). http://www.digitalhumanities.org/dhq/vol/3/4/000070/000070.html.

‡ Text Mining with Limited Access Data 2018 National Forum. "Data Mining Research Using In-Copyright and Limited-Access Text Datasets." Accessed September 14, 2019. https://publish.illinois.edu/limitedaccess-tdm/.

* Truyens, Maarten, and Patrick Van Eecke. "Legal Aspects of Text Mining." *Computer Law and Security Review* 30, no. 2 (April 2014): 153–70. https://doi.org/10.1016/j.clsr.2014.01.009.

* Tsai, Hsu-Hao. "Global Data Mining: An Empirical Study of Current Trends, Future Forecasts and Technology Diffusions." *Expert Systems with Applications* 39, no. 9 (July 2012): 8172–81. https://doi.org/10.1016/j.eswa.2012.01.150.

* United Kingdom Intellectual Property Office. *Exceptions to Copyright: Research*. Newport, UK: Intellectual Property Office, March 2014. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf.

‡ University of Cambridge. "Text and Data Mining: Home." Accessed September 14, 2019. https://libguides.cam.ac.uk/tdm/home.

‡ University of Chicago. "Text and Data Mining." Accessed September 14, 2019. http://guides.lib.uchicago.edu/textmining.

* Urban, Jennifer M. "How Fair Use Can Help Solve the Orphan Works Problem." *Berkeley Technology Law Journal* 27, no. 3 (2012): 1379–492. http://btlj.org/data/articles2015/vol27/27_3_S/27-berkeley-tech-l-j-1379-1430.pdf.

* Van Noorden, Richard. "Text-Mining Spat Heats Up." *Nature News* 495, no. 7441 (March 21, 2013): 295. https://doi.org/10.1038/495295a.

* ———. "Trouble at the Text Mine." *Nature News* 483, no. 7388 (March 8, 2012): 134. https://doi.org/10.1038/483134a.

‡ Westera, Wim. *The Digital Turn: How the Internet Transforms Our Existence*. Bloomington, IN: AuthorHouse, 2012.

‡ Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific*

Data 3 (March 15, 2016), article 160018. https://doi.org/10.1038/sdata.2016.18.

* Williams, Leslie A., Lynne M. Fox, Christophe Roeder, and Lawrence Hunter. "Negotiating a Text Mining License for Faculty Researchers." *Information Technology and Libraries* 33, no. 3 (September 2014): 5–21. https://doi.org/10.6017/ital.v33i3.5485.

* Williford, Christa, and Charles Henry. *One Culture: Computationally Intensive Research in the Humanities and Social Sciences: A Report on the Experiences of First Respondents to the Digging into Data Challenge.* CLIR Publication 151. Arlington, VA: Council on Library and Information Resources, June 2012. https://www.clir.org/pubs/reports/pub151/.

* Yang, Peilin, Mianwei Zhou, Yi Chang, Chengxiang Zhai, and Hui Fang. "Towards Privacy-Preserving Evaluation for Information Retrieval Models Over Industry Data Sets." In *Information Retrieval Technology: 13th Asia Information Retrieval Societies Conference, AIRS 2017, Jeju Island, South Korea, November 22–24, 2017, Proceedings*. Edited by Won-Kyung Sung, Hanmin Jung, Shuo Xu, Krisana Chinnasarn, Kazutoshi Sumiya, Jeonghoon Lee, Zhicheng Dou, et al., 210–21. Cham, Switzerland: Springer, 2017. https://doi.org/10.1007/978-3-319-70145-5_16.

* Zeng, Jiaan, Guangchen Ruan, Alexander Crowell, Atul Prakash, and Beth Plale. "Cloud Computing Data Capsules for Non-consumptive Use of Texts." In *ScienceCloud '14: Proceedings of the 5th ACM Workshop on Scientific Cloud Computing*, 9–16. New York: ACM, 2014. https://doi.org/10.1145/2608029.2608031.

PROJECT INVESTIGATORS AND LOCAL ADVISORY BOARD

Project Investigators

Bertram Ludäscher (Principal Investigator) is a professor at the University of Illinois School of Information Sciences, where he directs the Center for Informatics Research in Science and Scholarship (CIRSS). He also holds affiliate appointments with the National Center for Supercomputing Applications (NCSA) and the Department of Computer Science. He works in data and knowledge management, focusing on the modeling, design, and optimization of scientific workflows, provenance, data integration, and knowledge representation. He is a founder of the open source Kepler scientific workflow system project and a member of the DataONE leadership team. Ludäscher leads the NSF-funded Whole Tale project, which aims to transform the practice of knowledge discovery and dissemination into one where data and code are united with research articles to create "living publications," or tales. In other projects he develops workflow and provenance technologies for quality control and data curation of biodiversity data and for taxonomy alignment. Before coming to Illinois, Ludäscher was a professor at the Department of Computer Science and the Genome Center at the University of California, Davis. Prior to joining UC Davis, he worked at the San Diego Supercomputer Center, UCSD, as a research scientist. He received his MS in computer science from the University of Karlsruhe and his PhD in computer science/databases from the University of Freiburg, Germany.

Beth Sandore Namachchivaya (Co-principal Investigator) is the University Librarian at the University of Waterloo. She has significant leadership experience specific to academic libraries and expertise in research and information technology and has led the development of information discovery systems and digital preservation programs. Her research areas include discovery and access, new forms of scholarship, and the development of sustainable digital curation practices. Prior to her appointment at Waterloo, Namachchivaya held appointments as Associate Dean of Libraries, Associate University Librarian, and Professor at the University of Illinois at Urbana-Champaign Library in technical services and information technology and most recently led the establishment of a comprehensive research and scholarly communication program. Namachchivaya has held prior appointments at Northwestern University; the University of California, Berkeley; the California Digital Library; and the National Agricultural Library. She was selected as a fellow of the National Center for Supercomputing Applications (NCSA), participated in the Association of Research Libraries' Research Library Leadership Fellows program, and is currently affiliated with the Center for Informatics Research in Science and Scholarship (CIRSS) at the University of Illinois School of Information Sciences. Namachchivaya is professionally active as the incoming chair of the Tri-university Group (TUG) Library consortium (UWaterloo, Wilfrid Laurier, and Guelph Universities), the OCUL Collaborative Futures Directors' Committee, and the Canadian Association of Research Libraries (CARL) Portage Steering Committee.

Megan Senseney (Co-principal Investigator) is Head of the Office of Digital Innovation and Stewardship at the University of Arizona Libraries. She works collaboratively on a number of projects that sit at the intersection of digital humanities and data curation with a research interest in qualitative studies of complex socio-technical research environments. In particular, she is interested in understanding what happens when people's needs, technical infrastructures, and policies are at odds in the creation of new knowledge. Previously, she held appointments as a research scientist and project coordinator at the University of Illinois School of Information Sciences, where she managed a number of projects funded by the Mellon Foundation, IMLS, and NEH. Significant projects includes Publishing without Walls, a library-based digital scholarly publishing initiative; Workset Creation for Scholarly Analysis, a project geared toward facilitating the creation of text datasets for computational analysis; and Digital Humanities Data Curation, a series of NEH advanced institutes designed to help scholars steward their digital resources through the research life cycle. For the past four years she has also been an instructor and workshop organizer for the DHOxSS Humanities Data Curation strand.

Eleanor Dickson Koehl (Investigator) is the Program Manager for Research Facilitation in the Office of Advanced Research Computing at UCLA. Previously she was Associate Director for Outreach and Education Services for the HathiTrust Research Center and Digital Scholarship Librarian at HathiTrust. She led training and outreach for the HathiTrust Research Center, which facilitates computational text analysis of the HathiTrust Digital Library. She was one of the key personnel on an IMLS-funded project called Digging Deeper, Reaching Further, which developed a train-the-trainer curriculum for librarians to develop skills to support TDM. Her research interests include research behaviors in digital scholarship, particularly for text analysis, and scholarly needs for library-supported digital humanities and computational social science research.

Local Advisory Board

- **Scott Althaus,** Merriam Professor of Political Science, Professor of Communication, and Director of the Cline Center for Advanced Social Research at the University of Illinois at Urbana-Champaign
- **Sara Benson,** Copyright Librarian and Assistant Professor, University Library, University of Illinois at Urbana-Champaign
- **Melissa Cragin,** Executive Director, Midwest Big Data Hub, National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign
- **Jana Diesner,** Associate Professor and PhD Program Director, School of Information Sciences, University of Illinois at Urbana-Champaign
- **J. Stephen Downie,** Professor and Associate Dean for Research at the University of Illinois School of Information Sciences and the Illinois Co-director of the HathiTrust Research Center
- **Heidi Imker,** Director of Research Data Services, Associate Professor, and Associate Dean and Associate University Librarian for Research, University Library, University of Illinois at Urbana-Champaign
- **Victoria Stodden,** Associate Professor, School of Information Sciences, University of Illinois at Urbana-Champaign
- **Tom Teper,** Associate Professor, Associate Dean, and Associate University Librarian for Collections and Technical Services, University Library, University of Illinois at Urbana-Champaign
- **Dan Tracy,** Head of Scholarly Communication and Publishing and Assistant Professor, University Library, University of Illinois at Urbana-Champaign
- **Ted Underwood,** Professor of Information Science and English, School of Information Sciences and Department of English, University of Illinois at Urbana-Champaign