IMLS National Forum
(LG-73-17-0070-17)

Data Mining Research Using In-copyright and Limited-access Text Datasets
Participant Forum Statements and SWOT Analyses

Prepared by: Scott Althaus, Christine L. Borgman, Brandon Butler, Marc Cormier, Krista Cox, J. Stephen Downie, Patricia Feeney, Lucie Guibault, Wolfram Horstmann, Clifford Lynch, Peter Murray-Rust, Darby Orcutt, Thomas Padilla, Michelle Paolillo, Andrew Piper, Matthew Sag, Rachael Samberg, Jean Shipman, George Strawn, Paul F. Uhlir, Günter Waibel, Kate Wittenberg, Glen Worthey

Compiled by: Megan Senseney and Eleanor Dickson
FORUM STATEMENT

Background on the Cline Center for Advanced Social Research

The Cline Center for Advanced Social Research at the University of Illinois at Urbana-Champaign aims to equip and empower a new generation of social scientists, humanists, and data scientists to take up key challenges that threaten human flourishing in the 21st century—climate change, civil unrest, sustainability, inequality, security, and public health, to name a few—by applying advanced computational techniques at extreme scales to discover innovative solutions hidden in unstructured textual data. Extreme-scale unstructured data resources like the 500 million Tweets produced daily, or the millions of news articles published in the New York Times between 1945 and today, offer immense opportunities to better understand important societal needs such as effective environmental decision making, managing food production resources and policies, studying the effects of climate change on societal instability, and improving diagnosis and treatment of illness. Harnessing the potential of unstructured data is one of the grand challenges facing data scientists, social scientists and humanists today. The Cline Center serves to (1) empower non-STEM researchers with research tools and textual data resources to carry the data science revolution into the social sciences and humanities without requiring those researchers to develop advanced computer science skills, (2) connect computational expertise in the data sciences with subject-matter expertise in the social sciences and humanities in ways that address pressing societal problems around the world, and (3) provide students with practical experience in data science research with real-world impact. Its many activities encompass several research initiatives, including monitoring civil unrest globally; tracking the over-time development of economic, political, and social indicators for 165 countries; and developing new algorithms and analysis tools for studying the nearly 100 million articles in the Cline Center’s Global News Archive.

The Cline Center’s experience of nearly 12 years of text-analytics research using limited-access news datasets in an academic research context highlights a number of key vulnerabilities and ambiguities that hinder further progress in bringing text-analytic opportunities involving limited-access data to a broader range of academic researchers.

Vulnerabilities

- As a general rule, the academic research community tends to be unaware of or uninterested in the ethical and legal challenges involved in working with in-copyright or licensed textual data. The operant assumption seems to be “if I can access it, I can use it for anything I want to.” We might call this “gray hat” research because it operates in a space where legal and ethical compliance is uncertain or ignored. A much smaller part of the academic community seems willing to engage in “black hat” research using limited-access textual resources, where a researcher knowingly uses data in ways that are out of compliance with copyright law or terms of use, but will continue to do so unless directly caught, confronted, or sanctioned by responsible authorities.
- The disciplinary mindsets encouraged by “open source” models of science and innovation, the academic need for transparency in order to properly validate methods and models, and the emerging norms of information openness that have taken root among younger generations and
hacker subcultures (following the “WikiLeaks” model) combine to create a “perfect storm” of significant risk for gatekeepers of limited-access text datasets. The risk is that their users will feel perfectly free to access, copy, store, or publicly expose sensitive data in ways that violate terms of use or copyright law.

- Student status covers a multitude of sins. Even serious breaches of copyright law or terms of use committed by undergraduate or graduate students are likely to be dismissed by institutional authorities or addressed with minimal consequences for those who violate rules in any but the most flagrant extremes. Lax cultures of accountability encourage risk-taking among data users and potentially undermine the willingness of data gatekeepers to innovate or allow limited-access data to be exposed to larger research communities who could benefit from such access.

- Unless exceptional efforts are undertaken to build cyberinfrastructure that is specifically designed to mitigate a wide range of security threats, the typical computational infrastructure used to store and work with limited-access text data provides few opportunities to clearly observe what is actually being done on systems, and tends to create divisions of labor in which single individuals become responsible for controlling access to sensitive data without themselves being held to any oversight or accountability process. This creates a wide range of security vulnerabilities and risks associated with sudden loss of institutional memory following staff departures that are difficult to address without heavy investments of institutional resources.

- As a general rule, academic institutions have few incentives to invest the staffing time, computational resources, or legal expertise required to define clear boundaries for text analytics research using limited-access data and to provide the resources needed to support “white hat” research using limited-access text. As a result, even “white hat” researchers who want to properly adhere to copyright and licensing restrictions often have difficulty accessing license agreements that govern acceptable use, and difficulty interpreting terms of use for licenses that do not explicitly consider the needs of text-mining applications.

Ambiguities

- The chains of authority within academic organizations that govern use of limited-access datasets were unlikely to have been designed with this purpose in mind. As a result, these important sites of institutional oversight span numerous authoritative communities with conflicting goals, incentives, and perspectives on the key issues. In addition, these authoritative communities are often distributed within a university’s organization chart in ways that make it difficult to know who has ultimate authority to make a particular decision. A central tension is between university risk managers, typically lawyers tasked with minimizing institutional exposure to potentially adverse consequences, and university researchers, who probably face limited personal exposure to adverse consequences so long as they are operating within their proper institutional roles, but whose actions could potentially place the large institution at risk of adverse consequences.

- The legal boundaries defining what is allowable for text-analytics researchers working with limited-access data are constantly moving, defined by an evolving case law that few university officials are aware of (or can keep up with). In the absence of clear boundaries for what is permissible, the easiest institutional response is “no”.

---

1 However, even individual researchers are targeted for potentially small-scale copyright violations, as a recent case illustrates: [http://www.niemanlab.org/2016/06/1884-to-quote-300-words-from-the-new-york-times-in-a-book-two-authors-try-to-stand-up-for-fair-use/](http://www.niemanlab.org/2016/06/1884-to-quote-300-words-from-the-new-york-times-in-a-book-two-authors-try-to-stand-up-for-fair-use/)
• Cross-national research collaborations making use of limited-access text have special difficulty ensuring compliance with a potentially wide range of country-specific copyright and licensing laws. If the limited-access data are stored and analyzed in the United States by researchers who are physically visiting the US from the European Union, which set of copyright laws apply: those of the US or the EU? If the EU researchers are provided virtual access to the limited access data stored on US servers while the researchers remain in the EU, which set of laws apply?

**Strengths**

- US copyright law provides clear exceptions for “fair use” of in-copyright text in academic research and other non-commercial applications.
- Methods for large-scale algorithmic text analysis have well known strengths and weaknesses.
- Computational power required for large-scale text analysis is more readily available to academic communities today than in the past.

**Weaknesses**

- Researchers are often unaware that in-copyright data they can publicly access (such as through a web site) might have terms of use or licensing that prohibits redistribution or certain kinds of research uses.
- Licensing or terms-of-use restrictions governing university licensing of in-copyright text are often unavailable to researchers, even upon request, so the boundaries of appropriate uses are often difficult to determine.
- No central authority or source for researchers to contact for clarification on what is allowed or not allowed for research with particular in-copyright textual corpora. Instead, distributed set of potential authorities located in different parts of the institutional organizational chart make it difficult to get any clear or final answer on what is permitted.
### Opportunities

- Moving from analysis of random samples of text content by human analysts to entire populations of text content by algorithmic analysis.
- More clearly defining boundaries of “white”, “gray”, and “black” zones of what is allowable in research on particular corpora would help individual researchers make more informed decisions on which research projects to pursue.
- Developing a clear set of norms/guidelines defining appropriate uses of in-copyrighted textual data would help journal editors and larger academic communities evaluate the quality and transparency of text analytic research products submitted for peer review (the text-analytic equivalent of an Institutional Review Board for human subjects research, where it becomes clear when a research project has followed all appropriate guidelines).

### Threats

- Replication and transparency initiatives in many social sciences now require public access to replication data, but providing public access to in-copyright replication data isn’t possible.
- Limitations on sharing in-copyright text makes it difficult to test whether new algorithms offer clear improvements over old algorithms.
- It is unclear which country’s copyright standards might apply to in-copyright text mining conducted by cross-national research teams.
- Guidelines for “fair use” of in-copyright texts are constantly evolving, and licensing or terms-of-use restrictions might not be readily discoverable by individual researchers.
- Risks of violating copyright or licensing terms fall harder upon individual researchers more than their institutions, which may create disincentives for certain kinds of higher-risk research projects.
- Lack of clarity on “what is allowed” may lead researchers to simply do whatever they want to do, which could raise the individual and institutional risks for conducting such research.
- The more impediments are piled up against academic researchers working with in-copyright text, the more likely that innovative discoveries using text mining methods will occur outside the academy, and university researchers will lose their privileged places as leaders of innovation and discovery in their fields.
FORUM STATEMENT

Text Data Mining from the Author’s Perspective: Whose Text, Whose Mining, and to Whose Benefit?

Researchers have sought technical access to proprietary databases of published materials since the earliest days of online databases in the latter 1970s, yet publishers continue to write university contracts based only on human readership. By the time of Google Books and the associated author lawsuits, ca 2005, we learned that publishers wished to restrict “non-consumptive use” of scholarly content (Duguid, 2007; Leetaru, 2008; Nunberg, 2011). Throughout this period, the move toward open access to journal articles accelerated, with arXiv launching in 1991 (Ginsparg, 2011) and PubMed Central in 2000 (“PMC Overview,” 2018). Numerous other discipline-specific preprint servers, institutional repositories, and commercial services designed to distribute or redistribute open access versions of scholarly publications have been launched since. Concurrently, open access to publications became mandatory or highly recommended by many funding agencies and universities, in the U.S. and abroad.

Text data mining has become an essential part of scholarship in a world of “big data” where integrating disparate content can lead to new insights. More than 30 years ago, Don Swanson (1986) demonstrated the power of data integration to reveal “undiscovered public knowledge.” Swanson’s labor-intensive manual methods have been supplanted by machine searching, yet user interfaces and publisher contracts continue to be predicated on assumptions of human-computer interaction. As The Hague Declaration on Knowledge Discovery in the Digital Age states, “the right to read is the right to mine” (LIBER, 2015). The Public Library of Science, acting on this principle of the Hague Declaration, provides mining access to its corpus of more than 200,000 articles from all of its journals (Public Library of Science, 2018). The FAIR principles (Findable, Accessible, Interoperable, Reusable) for data sharing are similarly based on machine access to scholarly content (Wilkinson et al., 2016).

Given the many technical, social, and policy shifts in access to scholarly content since the early days of text data mining, it is time to expand the conversation about text data mining from concerns of the researcher wishing to mine data to include concerns of researcher-authors about how their data are mined, by whom, for what purposes, and to whose benefits.

This snippet from a recent Data Science Newsletter reflects some of the concerns expressed by authors (Noren & Stenger, 2018):
Elsevier may be shifting its business model to become a data provider to scientists which is different than being a journal publisher. Elsevier execs were at Harvard’s Data Science Initiative, where they are building intellectual partnerships to explore how this new model could work. Critics are warning scientists to heed the internet-age adage: if you aren’t the customer, you are the product. With Elsevier (and Facebook), the user will be both the customer and the product, to the tune of a 30% profit margin arguably due to the fact that the writing, reviewing, and much of the editorial work are done by volunteers. As we noted last week, Germany is having none of that anymore, demanding that all of the articles published by scientists funded by the German state be made freely available to anyone in the world. That’s an ongoing battle (Bert, 2017; Schiermeier, 2018; Smith, 2018).
Proponents of open access publishing tend to express one of two goals: (1) democratizing access to knowledge or (2) limiting the role of big publishers in controlling access to scholarly content, and in their ability to charge high fees to universities and readers (Borgman, 2007; Harnad, 1991, 1999, 2005; Suber, 2012; Willinsky, 2006). These competing views are conflated in recent developments, with proprietary publishers charging several thousand dollars (or euros) to make a single article open access, and with large publishers such as Elsevier purchasing independent, community-based open access venues such as SSRN and Bepress. Open access is not turning out to be the information commons that was envisioned by its pioneers (Benkler, 2004; Hess & Ostrom, 2007; Kranich, 2004; Lessig, 2001; O’Sullivan, 2008; Reichman, Dedeurwaerdere, & Uhlir, 2009; Reichman, Uhlir, & Dedeurwaerdere, 2016).

The rise of bibliometrics and “altmetrics” to assess scholarly productivity in the academy also has authors worried about who can mine their content and for what uses. Despite a long history of scholarship that demonstrates the meaninglessness of most of these numbers in assessing actual impact or long-term value of an individual’s contributions, the uses of these metrics continue to proliferate (Borgman, 1990, 2016; Cronin, 2005; Cronin & Atkins, 2000; Cronin & Sugimoto, 2014, 2015). Calculations of citations, H-index, and other indicators vary widely between common sources such as Web of Science, Google Scholar, and Scopus, due to differences in editorial coverage, algorithms, and methods used by those attempting to mine these databases.

Whether mining a corpus for bibliometrics, textual content, images, or numeric data, the bibliographic descriptions are essential metadata. Original articles typically provide accurate bibliographic descriptions, and may also include “please cite as” instructions. However, references to published articles, which are essential for bibliometrics or for integrating content across databases, are inherently dirty data due to the vagaries of how authors create reference lists. A bibliography in a journal article is far from the “necessary and sufficient” set of citations that might be assumed by bibliometric evaluations. Rather, it is often an idiosyncratic list of familiar sources, compiled based on what is handy when the publication is submitted. Too few authors are bibliographic purists who verify middle initials, dates, DOIs, and page, volume, and issue numbers (Borgman, 2015, 2016). Complicating matters further is the lack of agreement on bibliographic styles. At last count, Zotero offered about 9,000 journal styles for referencing, representing about 1750 unique bibliographic styles (“Zotero Style Repository,” 2018).

As universities automate academic personnel processes, faculty dossiers also become rich sources of bibliographic data. These records tend to be more accurate than citation lists because authors have a vested interest in establishing claims to their oeuvre. Faculty are becoming concerned about who has access to these data in machine-readable form, and how the data can be mined for making decisions about their careers, their departments, and their fields. Privacy issues abound. Individuals can give informed consent for specific uses of specific data, such as a dossier for a personnel action. When universities employ these data for other purposes or share them with external parties such as the publishers who own some of these dossier systems, privacy and academic freedom issues also arise. Following an extensive analysis of data governance issues at UCLA, for example, operational structures were developed to address the following concerns (“Data Governance Task Force: Final report and recommendations,” 2016):

- When data are used to make decisions about people
- When data are collected about people without their knowledge or consent
• When data about people are used in unexpected ways without subjects’ knowledge or consent (e.g., new applications of data or systems; mining, analysis, and aggregation)
• When data are used for evaluation purposes
• When data are shared with external entities, whether with research partners or through service contracts with the private sector

A related concern is the ability of publishers to surveil uses of scholarly materials. Ownership of intellectual property carries a large set of rights and responsibilities, some which are associated with privacy protection and intrusion. Corporate owners of scholarly publishing, mass media, and social media content deploy “digital rights management” (DRM) technologies to track uses and users in minute detail. These technologies have eroded traditional protections of privacy and intellectual freedom in libraries and other domains (Cohen, 1996; Lynch, 2017). Universities have special responsibilities for managing their intellectual property in ways that protect the privacy of their communities and minimize harm (Borgman, 2018). The Hague Declaration also includes a principle that “providers of content should respect the intellectual privacy of individual readers” (LIBER, 2015).

Scholarly publishers are becoming data services vendors, entering new markets by acquiring companies in multiple sectors of the information economy. In doing so, they follow the successful business models of Alphabet/Google, Facebook, and Amazon in aggregating vast amounts of data about people’s lives. To the consumer, they promote the advantages of improving user experience with intelligent adaptation. To their business clients and investors, they promote the advantages of predictive analytics that can be deployed to strategic advantage. In the academic community, predictive analytics are being used to assess the performance of students and faculty, departments, universities, journals, research programs, and much more. The concentration of data by a few large players gives them a “god’s eye view” of their domains, with minimal oversight or regulation (“The world’s most valuable resource is no longer oil, but data,” 2017). Only gradually are scholarly authors coming to realize that if you are not at the table, you are on the menu. Authors must have a seat at the table in this national (and international) forum on Data Mining with Limited Access Text. Privacy, autonomy, and academic freedom are at stake. Let the first question be “Whose Text, Whose Mining, and to Whose Benefit?”

Acknowledgement:

Thanks to Michael Scroggins, UCLA, for comments and discussion on an earlier draft.

References


### Strengths
- Authors can release their work in open access venues
- Authors can position their work for multiple audiences

### Weaknesses
- Authors need to make informed choices about where to place their work based on their IP rights and potential for surveillance
- Authors face large risks in unanticipated reuses of their data, which may be unacknowledged and unattributed, e.g., finding their work in some collection, in whole or part, attributed to them without notification

### Opportunities
- Authors can position their work more effectively than in prior generations, if they choose to do so

### Threats
- Privacy
- Surveillance
- Academic freedom
- Decisions made about authors without their knowledge, which may have profound effects on careers, and status of their academic departments, e.g., promotions, job offers, awards, etc
FORUM STATEMENT

As part of the legal community, my primary lens for looking at these activities is the fair use doctrine. Developments in the doctrine are what make these practices possible despite copyright law’s otherwise draconian consequences for unlicensed uses. In thinking through what implementation of a non-consumptive use policy might look like for HathiTrust, it was really important to understand the contours of the latest caselaw, and in particular the Google Books decision from the 2nd Circuit. One fascinating question that I still wrestle with in thinking about this line of law is: if my purpose is transformative, to what extent am I responsible for the substitutive/superseding uses of others?

Fair use favors transformative uses because they add to the overall culture by repurposing existing works to create cultural value. In a way, the concept of transformative use is just the inverse of the much older concept of “merely superseding” uses. Since the widely-cited 19th century case, Folsom v. Marsh, courts examining fair uses have asked whether a new use is “merely superseding,” i.e., is it a use that seeks to usurp/substitute/replace the original in the market. Simple piracy is the classic example of an unfair use: selling copies of a work to customers who would otherwise obtain copies from the copyright holder (or her licensee). So, providing the public with substitutes is typically unfair, but what if your purpose is non-substitutional, but you provide the public with enough of the original that they can (contrary to your purpose) use your activity in a superseding way? As I read the Google Books opinion, this could be fatal. It is essential in Leval’s eyes that the “snippets” shown in Google Books search results are not enough to substitute for reading the entire book.

Other cases have similar lines of logic. The image search cases rely on the relatively low quality of the thumbnails created by search engines. The Bill Graham Archives case, about reproducing rock concert posters, relied heavily on the fact that the reproductions in the allegedly infringing book were too small and low-resolution to serve as decorations for college dorm rooms.

On the other hand, we have the general rule of secondary liability in copyright law, which follows the common-law rule that no one is liable for the acts of another unless they are in a special relationship with that person where either they benefit from or can control the behavior of the third party, and, (in some cases) they have knowledge of that person’s bad actions. That rule would typically exonerate a good-faith fair user whose use was for a non-superseding purpose, but was then converted into an infringing use by a third party. Imagine, for example, a scholarly book that included a reproduction of a fine art painting, where the criticisms in the book simply wouldn’t make sense without the painting as reference. If a scofflaw used the reproduction as the basis for a coffee mug sold online to fans of the artist, that scofflaw would be infringing, but the scholar and her publisher would not be secondarily liable under ordinary principles. But could their use be determined, retroactively as it were, to be unfair because of its (now
realized) potential for abuse by third parties? In that case, they wouldn’t be secondarily liable, but they would be directly liable for their own use, which is not fair because of its potential for abuse.

All of this goes to the questions of how much material can be released from the search corpus, and in what circumstances. If researchers have unique, transformative purposes that are not served by the ordinary market for access to these works, perhaps the answer could be “quite a lot.” Limiting the audience may be sufficient to avoid a substitution effect, even if the amount disclosed is not limited. Anyway, these are the moving parts of the doctrine that I’m thinking about today.
### Strengths
- Transformative use framework/HathiTrust/Google Books
- 17 USC 504© liability limitation for reproduction by nonprofit educational institutions
- eBay v. MercExchange re injunctions
- State sovereign immunity (for publics)
- Fee shifting to winning defendants (Fogerty v. Fantasy/Kirtsaeng v. Wiley II)

### Weaknesses
- Google Books logic bars making data fully available for reproducibility or closer reading/consultation
- 504(c) limit does not apply to distribution, display
- Private schools lack sovereign immunity
- Only employees of non-profit ed inst’ns are shielded by 504(c)—service providers, contractors, etc. not covered

### Opportunities
- The lower legal risk associated with public and non-profit institutions gives them an advantage, making them attractive partners/leaders in datamining cooperatives
- HathiTrust has such a large corpus, it will also be a natural leader and have power to attract collaborators
- The need to collocate collections for TDM could be a lever to help bring content together for other purposes, and perhaps to convince partners to share library views about access to info, or at least to go along with library practice as partners.

### Threats
- One alarming threat is the notion of a TDM “license,” something publishers are already selling, and that is increasingly popular in other countries, which lack fair use. TDM is a fair use; you don’t need a “license” to do it, though you of course do need access to the content. Clarity is needed here so that researchers understand what they’re paying for.
- I don’t expect any backsliding on the core principle of non-consumptive use being fair, but if the concept of “transformative use” were to come into disfavor, it could lead, in turn, to a shrinking of the zone of permissible TDM. Watch the TVEyes case!
- TDM does require access to content, and some kinds of content are, notoriously, locked up in proprietary databases. Fair use is worthless without access to the material.
FORUM STATEMENT

As a reference and archive publisher and periodical aggregator, Gale has a long-standing tradition and responsibility to pursue and preserve copyright integrity for the works we create as well as the content we source through myriad partners in the academic and cultural preservation communities.

As a for-profit enterprise serving the public, school, and academic library markets, we have seen a dramatic increase in demand for data, but not the formatted, packaged kind that traditional publishers like Gale have been delivering for decades. In these cases, there are no interfaces needed, no narrative introductions to reprinted content – just the raw information, the data, for use in ways that were not imagined or envisioned in 1974 when our first serials began filling library bookshelves and our microfilm program struggled to meet demand.

Through decades of publishing, we have created and amassed terabytes of narrative and fielded data in the pursuit of our publishing programs. With the advent of Digital Humanities and natural language processing, we see a unique opportunity to amplify the discovery resident in this content through the application of data science and analytics. Much like the winding evolution from print to digital delivery, the contract and copyright landscape in traditional institutions has had to adapt to this new frontier in academic research and, as a byproduct, brings one of the biggest hurdles to access to audiences everywhere.

In our capacity as a publisher, we’ve seen significant challenges to the sanctity of copyright and contract through The New York Times vs. Tasini, Jerry Greenberg vs. National Geographic Society, and other challenges to copyright law and access rights that have emerged on the back of new methods of distribution and evolving technologies. We have watched closely as the legal rulings unfolded, not only to understand how the decisions impacted both parties, but publishing at large.

As a publisher that provides raw data to institutions based on their digital archive purchases, Gale is squarely positioned in the view that, provided an institution abide by the license terms of use, data mining is allowed for the sole purpose of supporting non-commercial scholarly research at the funding institution. The scholarly impact of data mining across the humanities and scientific fields of study are both obvious and evolving. It is equally clear that both areas would benefit from increased access to the scientific and historical record, to amplify learning and outcomes through new and emerging research techniques. What is diminished by denying access to these data can be quantified in basic terms: new learning.

Many of our source libraries agree with the need to adapt to and enable the new and emerging forms of inquiry seen today. As a publisher, we are eager to support this provision wherever and however we can but there are, at times, circumstances that prevent this. As much as the digital humanities today can defy common definition, the value of the practice can be misconstrued by source libraries as a means to disaggregate content and dilute the terms of archival license agreements.

In many cases where access to content is limited, the underlying policy from source institutions is at times grounded in the overarching desire to retain control of content where they are not at all clear what the
use and access to it is in a text data mining context. The trepidation can often be traced back to the fluid nature of datasets. How these datasets are created, versioned, accessed, and distributed through the process of data mining a “parent” dataset raises concerns about data security, academic citation, and its relationship to its primary source.

Gale can speak to numerous publisher relations where access to content for the purpose of data mining is restricted based on a blind spot – the source institution has little to no visibility into how its content is being used in the context of data mining in support of scholarly research and publishing. In some cases, this lack of transparency or understanding is a critical factor in deciding whether or not data mining will be allowed.

In addition, as a periodical aggregator, Gale provides access to millions of articles from peer-reviewed and consumer-oriented publications that are under copyright. The possibilities of using this content to understand contemporary history, thinking, and culture through the lens of data science are many. We endeavor to work with the publishing and scholarly communities to define what is “in bounds” for use in digital scholarship and to broaden access to data to drive research and discovery in ways that benefit the academic and publishing communities at large.
**STRENGTHS**

- Large repository of OCR text suitable for data mining and natural language processing.
- Long-standing relationship with publishers and content partners who endeavor to support digital scholarship and research.
- Technologies that support data mining and natural language processing.
- Ongoing publishing into areas of history and social issues that resonate in contemporary society and are of deep interest to digital humanists.
- Supports established DH projects, teams, and community at large.

**WEAKNESSES**

- Rights restrictions for some content sets not only limit the amount of data available for large-scale analysis, they also can impact the scope of the data itself, possibly weakening the comprehensive nature of some collections in primary source form.

**OPPORTUNITIES**

- Bringing humanities datasets into new areas of research through the provision of data mineable content sets.
- Providing new opportunities for source libraries to support digital scholarship and to drive new insights through distant reading.
- Establishment of universal guidelines to determine access rights that support the needs of both publishers and researchers.

**THREATS**

- Changes/amendments that further broaden the time-based definition and duration of copyright.
- Lack of data security at academic institutions for publisher-provided text datasets.
FORUM STATEMENT

Text and data mining (TDM) is a valuable tool for research, allowing for the discovery of new knowledge, leading to advancement in scholarly research, facilitating scientific and medical breakthroughs, promoting innovation and more. It allows a researcher to filter through information and crawl through vast amounts of text and data that would not otherwise be possible. TDM also facilitates cross-disciplinary studies to find novel links.

While TDM is an important tool for all different types of research and fields of study, its use may be restricted for a number of reasons. Much of the information that a researcher wants to access for TDM purposes may be behind a paywall or otherwise subject to licensing restrictions. For works that are in copyright, individuals may be hesitant to engage in TDM without first seeking permissions.

Fair use can provide a solution, allowing for works to be used in a transformative manner. Although U.S. copyright law does not have a specific limitation or exception to explicitly allow for TDM, fair use has accommodated this innovation. TDM provides a significant benefit to the public and the creation of databases that ingest copyrighted works for search are highly transformative uses that offer a new purpose, rather than substituting for the original.

Even though fair use allows for TDM, some researchers may be hesitant to rely on this right and instead prefer to rely on licenses or material that is not behind a paywall. Unfortunately, the permissions culture can result in over-reliance on licenses, even where they are unnecessary, resulting in higher costs and greater expenditure of resources. Relying solely on licenses might also significantly reduce the scope and breadth of a particular corpus that the researcher is engaging his TDM efforts, thus altering the results. An individual might use TDM solely on abstracts of scholarly articles, which are generally available more widely, rather than the full text of articles, which can greatly alter or skew research results.
<table>
<thead>
<tr>
<th><strong>STRENGTHS</strong></th>
<th><strong>WEAKNESSES</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Discovery of new knowledge from existing knowledge and ability to quickly filter through texts and organize information</td>
<td>• Permissions culture and overreliance on licenses; belief that it is necessary to obtain a license to engage in TDM.</td>
</tr>
<tr>
<td>• Ability to make new links and connections across broad materials, not possible without the aid of computer programs</td>
<td>• Born digital content is often being made available on licensing terms only, leading to walls and limitations, where TDM may only be able to mine limited sets of data</td>
</tr>
<tr>
<td>• Ability to detect patterns/trends and conduct information analysis across greater number of sources</td>
<td>• Licenses may be silent on issue of TDM (and fair use) and some may be hesitant to allow TDM unless it is explicitly contracted for</td>
</tr>
<tr>
<td>• Use of TDM across materials in different disciplines, encouraging the growth and expansion of cross-disciplinary studies</td>
<td></td>
</tr>
<tr>
<td>• Enhanced ability to analyze vast numbers of documents in different languages and extract data</td>
<td></td>
</tr>
<tr>
<td>• Ability to more effectively analyze (unintentional) biases</td>
<td></td>
</tr>
<tr>
<td><strong>OPPORTUNITIES</strong></td>
<td><strong>THREATS</strong></td>
</tr>
<tr>
<td>• Libraries can provide access to new types of research through TDM</td>
<td>• Fear of publishers/rights holders initiating litigation for unlicensed TDM.</td>
</tr>
<tr>
<td>• Researchers in different fields looking to do this type of research to find more relevant information and novel links</td>
<td>• Publishers/rights holders ability to restrict TDM via licensing language for content that is only available digitally (and licensed rather than sold). Elsevier, for example, has “adopted a license-based approach” for text mining for non-commercial research. If someone who is not a subscriber, “Requests are considered on a case-by-case basis.”</td>
</tr>
<tr>
<td>• Better use of resources/time to identify relevant works and extract the relevant information from the works</td>
<td>• Some researchers do not admit to TDM or are unwilling to share their projects because of fear of being sued, leading to difficulty in reproducibility</td>
</tr>
<tr>
<td>• Confirmation from Google Books and HathiTrust cases that digitization for nonconsumptive purposes is a fair use</td>
<td></td>
</tr>
</tbody>
</table>
FORUM STATEMENT

As a forum participant, I would say that I represent two communities: 1) a researcher who uses text data mining methods; and b) a content provider.

As a researcher, I am a professor and associate dean for research at the School of Information Sciences at the University Illinois. My research areas include digital humanities, digital libraries, informetrics and music information retrieval (MIR). In each of these areas I have used a large number of data mining and text analysis tools across a wide range of data coming from such things as user-generated questions, user-generated reviews, books, articles, jazz performance metadata, music metadata, lyric data, and so on. Most of the data used in my research has some from a mix of open and restricted sources. Acquiring the restricted data has always required many hours of negotiation, usually some considerable monetary cost and much effort in keeping the data secure to the satisfaction of the data providers. In recent years, we have received more and more “push-back” from reviewers and colleagues for not being willing/able to provide access to our restricted data sets. Sometimes the negative comments are expressed in terms of the need for reproducibility; sometimes, they are expressed in terms of collegiately in the presence resource scarcity (i.e., good colleagues should share rare/expensive/restricted data).

My first exposure to the problems associated with copyright restricted data happened when I began working with music files in the 1990’s as part of my MIR evaluation research. In the early days of MIR, it was difficult to build up standardized test collections of music in audio form because of all the copyright issues and practices that surrounding the music industry. Beyond the cost of purchasing enough recordings (we would to buy random CDs by the pound in big cardboard shipping boxes to save money), it was very difficult (read IMPOSSIBLE) to locate all the rights holders to get permission to share the collections with fellow researchers for non-commercial research uses. This lack of shareability/access prevented the nascent MIR research community from making any kind of advances that required testing and evaluation against a common dataset. However, at the same time, the text retrieval research community was making great advances with its research (think Google) as they were able to create large collections of textual data that they were able to share with researchers under the auspices of the TREC (Text REtrieval Conference).\footnote{\textsuperscript{1}}

Each year, TREC participants would create new datasets that could be shared, create some new standardized queries, and then let IR researchers build systems that answer the queries against the common datasets. The system responses would then evaluated by a common set of evaluators who would provide a set of scores for each systems showing how well each system did against each query. Many (most?) of the search engine algorithms we use every day were proven and then improved via the
annual TREC evaluations. This is a perfect example how more open data provides real impact on our everyday lives.

Because the MIR community could not get into a position where it could create a set of shareable datasets, I help to lead an effort to replicate the benefits of the TREC conference approach, but at the same time not violate copyright laws. In the end, we achieved this noble goal by inverting part of the TREC paradigm; that is, instead of sending data out to the researchers for them to run their algorithms, we built a model wherein the researchers sent us their algorithms to run again the data securely stored on the servers here at Illinois. In 2005, Illinois hosted the first Music Information Retrieval Evaluation eXchange (MIREX)\(^2\) using this new algorithm-to-data model. MIREX has been held every year since 2005. It has evaluated 3,338 individual algorithm runs and host some 55 different MIR datasets. Google Scholar lists some 4,120 MIR papers that cite MIREX. MIREX has become to the MIR world what TREC is to the text IR world: the common ground where real comparisons can be made. In this way, I like to think of MIREX as a success story in providing new kinds of productive access to otherwise restricted data.

As content provider, I am currently the Illinois co-Director of the HathiTrust Research Center (HTRC).\(^3\) The HTRC is a unique collaboration between the University of Illinois and Indiana University designed to provide non-consumptive research access to the 16.2 million volumes (5.7 billion pages) found in the HathiTrust Digital Library (HTDL).\(^4\) Roughly 37% of the volumes are in the public domain (6.1 million). The remaining 63% (10 million) are under copyright restrictions and cannot be shared with researchers. As you can see, these are non-trivial numbers that fall well within the domain of “big data.”

There are many challenges associated with providing non-consumptive access to text collections at this scale. These include, for example, the problem of simply finding the right items to include in an analysis (think needle in a very big haystack, and you are not allowed to actually look into the haystack). Moving and then crunching the data can, without exaggeration, take months and many thousands of hours of computing time (with failure at the end of the process being a much too common occurrence). Getting the data, which often contains many different kinds of OCR errors, into a consistent form for use by various non-consumptive tools can challenge even the most sophisticated user (again, made more challenging by not being able to see the copyright-restricted texts).

In helping to run the HTRC, I have found it useful to draw upon my experience in building up MIREX. One of the key lessons that has transferred over from MIREX to HTRC includes minimizing the amount of interference the system has to impose (for security) on the researcher (i.e., getting out of the way as much as possible). At HTRC we have released an Extracted Features dataset that contains page-level information (e.g., token counts, parts-of-speech, etc.) for each of the 5.6 billion pages. This data is NOT covered by copyright and thus can be freely shared with researchers who can do with it as they please. They second way that HTRC has aimed to “get out of the way” has been through its Data Capsules

\(^3\) See [https://analytics.hathitrust.org](https://analytics.hathitrust.org)
\(^4\) See [https://hathitrust.org](https://hathitrust.org)
technology which provide sophisticated users with their own special virtual machine that allows them a great deal of coding freedom while still protecting the copyright-restricted data.

<table>
<thead>
<tr>
<th>STRENGTHS</th>
<th>WEAKNESSES</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Early non-consumptive access systems like HathiTrust Research Center, Portico, etc. generating lots of interest</td>
<td>• No real consensus yet on what needs to be done</td>
</tr>
<tr>
<td>• Early non-consumptive systems showing potential and some real results are being generated using new systems</td>
<td>• No deep understanding of all the possible collections and their characteristics</td>
</tr>
<tr>
<td>• Funders, collection holders, and institutions interested and willing to support early efforts</td>
<td>• No comprehensive understanding of the many different users and their use cases</td>
</tr>
<tr>
<td>• Early adopters have been open-hearted and collaborative</td>
<td>• No clarity on the legal boundaries</td>
</tr>
<tr>
<td>• New data becoming available all the time</td>
<td>• A general hesitancy to expand boundaries based on probably faulty understandings of options and risks</td>
</tr>
<tr>
<td>• New tools coming online</td>
<td>• Data currently siloed in locations</td>
</tr>
<tr>
<td>• Lots of goodwill presently (but has short shelf life)</td>
<td>• No real progress yet on standardizing data formats for cross-collection analyses</td>
</tr>
<tr>
<td></td>
<td>• Current systems too intrusive</td>
</tr>
<tr>
<td></td>
<td>• Inconsistent/broken metadata</td>
</tr>
<tr>
<td></td>
<td>• Many tools assume access to raw data</td>
</tr>
</tbody>
</table>
### Opportunities

- Growing awareness of the availability of data that was previously not open for analysis has created new communities of potential users that would like to learn more about the data
- New research questions are being formulated inspired by the newly available data
- New teaching opportunities being inspired by access to new data sources
- Chance now, in early days, to get cross-collection systems in place before things ossify
- New collections coming to light build new interest (in a virtuous cycle)

### Threats

- Terminal disappointment created by the mismatch between user expectations and what can actually be delivered with regard to access to restricted data
- Lack of financial or institutional support for provisioning the data and the requisite compute resources leave the potential unrealized
- Publishers or other commercial players will introduce partial, collection-specific, solutions that will crowd out further support for non-profit, cross-collection, efforts
- Goodwill evaporates if no progress seen by users and collection owners
FORUM STATEMENT

I am writing most closely from the perspective of a professional organization, although Crossref doesn’t fall neatly into this category and may also be considered a content provider. Crossref is a not for profit membership organization with scholarly publishers as voting members. Our members register metadata about their content (journal articles, books, conference papers, datasets, preprints, standards, reports) with us, we assign persistent identifiers and collect metadata about scholarly works but do not collect content itself.

One of the headaches for researchers with implementing TDM is identifying and accessing data that is mine-able. Publishers all have different licenses defining what can be done with their content. Approaching each publisher individually takes lots of emails and poking around on websites. Publishers also have historically had different methods of providing data to mine.

We collect the following metadata related to text and data mining and open content:

- Text and data mining license URLs and applicable dates
- URLs specified for data mining retrieval
- ‘free_to_read’ tag to identify open content.

Our metadata for license info is consistent with the NISO ALI recommendation. Our members are comprised of open and closed access publishers and the content registered with us is cross-disciplinary and international. Our goal is to provide this data comprehensively. We don’t provide content discovery tools but could potentially be used to surface content not consistently considered by researchers for text and data mining.

Some of our publisher members provide this data regularly but our data collection is far from comprehensive. Large publishers (Wiley, Elsevier, Springer) provide license data for content, smaller members may not. Many open access publishers do not. This is due to technical hurdles for providing data, or because they are unaware of the value of providing that data in a central location.

At the moment it is not possible to determine from Crossref metadata what percentage of content is ‘open’, although we do provide members with the means to provide this data. We have a neutral stance on member business practices but want to make all data available so that researchers can use our freely available metadata to get the information they need.

We began collecting this data in 2014 and have worked to encourage members to provide it, but haven’t focused specifically on collecting the ‘free to read’ metadata tag. Our plan is to educate our members on why providing comprehensive metadata is valuable.
Another challenge is that we find that members often assume they are sending us data that they aren’t - they work with third parties (hosting platforms, submission systems, typesetters) and often make assumptions about how their metadata moves from one provider to another. Text and data mining information is often omitted, in conversations with members we’ve found that they assume their vendor is providing “all data” to Crossref when in fact they are not.

We’ve also found that data quality is an issue, we don’t normalize or validate license URLs beyond requiring that they be a URL so we have licenses like http://creativecommons.org/LICENSEs/by/3.0/ (instead of http://creativecommons.org/license/by/3.0/) and obvious typos like http://eativecommons.org/licenses/by/2.0.

We are in the process of creating new reports for our members to allow them to view what data they are sending us, this will include license and free to read information. This information is available via our REST API but does require some interrogation. We’re also going to be undertaking some general metadata quality initiatives to reach out to members individually to address data quality issues.

Our hope is that making this data easily available to both members and (eventually) the public will help our members initiate conversations with vendors that result in sending us comprehensive TDM metadata.
<table>
<thead>
<tr>
<th><strong>STRENGTHS</strong></th>
<th><strong>WEAKNESSES</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• collect metadata specific to text and data mining licenses</td>
<td>• license metadata still requires researchers to review individual licenses</td>
</tr>
<tr>
<td>• provide a variety of channels for members to send us this data</td>
<td>• members are not comprehensively providing the data or keeping it up to date</td>
</tr>
<tr>
<td>• distribute metadata freely through a REST API</td>
<td>• method doesn’t address issue of licenses for specific institutions</td>
</tr>
<tr>
<td></td>
<td>• a wide range of licenses are available, maintaining data quality for licenses is a challenge</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>OPPORTUNITIES</strong></th>
<th><strong>THREATS</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• encourage members to provide more TDM-specific data by making the process easier and more understandable</td>
<td>• members may fail to see the value of providing TDM metadata to Crossref</td>
</tr>
<tr>
<td>• make the process to retrieve data more transparent</td>
<td>• members may not keep license and free to read data up to date</td>
</tr>
<tr>
<td>• provide more information to researchers about what data we have and do not have</td>
<td></td>
</tr>
</tbody>
</table>
FORUM STATEMENT

I identify most with the legal expert community, being someone who has never used text mining methods for her own research, but who has studied the obstacles to text mining from a legal perspective. I am convinced of the importance of text mining methods in science, as well as in countless other sectors of economic activity, as a means to generate new knowledge and insight and to significantly reduce the time and money involved in carrying out research. My research has focused primarily on European Union and European national intellectual property law, with an emphasis on Dutch, French and German law.

On the basis of my research findings, it is fair to say that the current European legal framework creates definite barriers to text mining. The broad interpretation given to the exclusive rights granted under the European copyright and sui generis database right regimes is such that rights owners enjoy extensive protection on collections of texts that could be used for mining. In Europe, text mining activities are in principle subject to the prior authorization of the rights owner, unless a statutory exception applies. Whether current European copyright law provides for an exception permitting TDM activities is altogether uncertain, fragmented and restrictive: 1) uncertain, because the exceptions listed either the European acquis communautaire (mainly Directive 2001/29/EC) or the Member States law do not clearly qualify TDM as non-infringing acts; 2) fragmented, because of a lack of effective harmonization of rules between the Member States, where the law of some countries do contain certain exceptions for research purposes, but not others; and 3) restrictive, because even where the exceptions would allow the use of works for purposes of scientific research, they are generally too narrow to cater for the different acts of reproduction involved in TDM or to allow the communication to the public of parts of the results.

Moreover, exceptions for the benefit of research are usually limited to non-commercial activities in the academic environment. The United Kingdom is the exception to rule, as it is the only Member State to have adopted in 2014 an exception expressly permitting acts of reproduction for non-commercial TDM purposes. It is unclear whether this isolated move on the part of the UK government has had a tangible impact in reducing the legal uncertainty in the research community.

In practice, a number of (mainly large) publishers are taking advantage of the legal uncertainty: on the premise that they are entitled to control this type of use, they are requesting users to obtain (add-on) licenses for the use of their databases for TDM purposes, above and beyond the licenses required to access the database in the first place. License terms are often accompanied by technical restrictions, like APIs that limit TDM to specific data sets or query algorithm, e.g. Elsevier. Many researchers see this practice as an unnecessary and illegitimate burden on their research activities. Fortunately, some publishers do allow TDM activities to take place with little to no restriction, with the understanding that all content published under open access conditions is in any case freely minable. The lawfulness of mining activities using other sources than the databases of publishers is less clear. The obligation to obtain express permission before carrying out TDM activities creates such prohibitively high transaction costs.
that researchers and their institutions tend to look for alternative routes to the mine(s) or to refrain from mining altogether. As a result, Europe lags behind other parts of the world in terms of TDM output.

TDM and copyright law has become a political hot topic in the European Union. In September 2016, the European Commission tabled a Proposal for a Directive on Copyright in the Digital Single Market which contains a provision on TDM. The negotiation process is slow and extremely polarized; the fate of this exception is also tied to that of other, even more controversial, provisions in the Proposal (e.g. right for newspaper publishers and internet filtering obligation). Anything can happen on the way to final adoption, but as things currently stand, it is to be expected that the TDM exception will be restricted to activities carried out by 'research organizations' and strictly for 'scientific research purposes' with respect to works to which they have lawful access. Moreover, the exception would only apply to the right of reproduction, not the right of communication to the public, and require the deletion of the copies of works after a certain period of time. Whether such a provision would be an improvement over the current situation is doubtful.

In all this discussion, the one question that only academics have dared ask so far is whether acts of non-expressive reproduction of works like in the case of TDM should fall within the scope of copyright protection in the first place. If one took a normative approach to copyright law, there would be a strong argument for not counting TDM as a copyright relevant act at all.
<table>
<thead>
<tr>
<th>STRENGTHS</th>
<th>WEAKNESSES</th>
</tr>
</thead>
<tbody>
<tr>
<td>• TDM is one of THE research methods of the future; it takes advantage of high degree of digitization and computational capacity to generate new knowledge;</td>
<td>• TDM is only available for researchers with sufficient <strong>computer skills</strong>;</td>
</tr>
<tr>
<td>• TDM gives meaning to the vast amount of cumulative scientific literature published annually (some of which never get read by humans);</td>
<td>• Dependent on <strong>finding and accessing data</strong>;</td>
</tr>
<tr>
<td>• Allows for the generation of new knowledge;</td>
<td>• If data is accessible, it is often in different <strong>formats, poor organization and varying quality</strong>;</td>
</tr>
<tr>
<td>• Reduces time and costs of conducting analytical research</td>
<td>• Verification, replicability and reliability of results requires availability of the original datasets;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OPPORTUNITIES</th>
<th>THREATS</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Potential of TDM greater than only in science, i.e. also investigative journalism, innovative start-ups, education (student learning and performance analysis), etc.</td>
<td>• Uncertainty about the lawfulness of TDM in Europe;</td>
</tr>
<tr>
<td>• TDM should be allowed for all types of purposes and activities, whether commercial or not;</td>
<td>• Legal uncertainty likely to create chilling effect on TDM;</td>
</tr>
<tr>
<td>• TDM should be possible for all types of digitized works, and not only literary works;</td>
<td>• TDM activities likely to migrate to countries where TDM is clearly allowed or openly tolerated;</td>
</tr>
<tr>
<td>• TDM should be equally possible on a cross-border basis;</td>
<td>• Legal uncertainty likely to stifle innovation;</td>
</tr>
</tbody>
</table>
FORUM STATEMENT

Supporting and enabling research are main tasks of research libraries. In the last decades, libraries provided the best possible physical or digital literature that could be acquired from publishers to forward it to the researchers. Today, libraries increasingly engage in publishing outputs of (local) researchers as open access literature, generating and curating data, produce digital editions, or develop authority files and ontologies. Engaging in this workflow means to support the complete data lifecycle. Libraries will have to gain insight and influence on digital output, the data, transformed into articles and books, but also the complete scientific workflow, and also the software that can be provided as a service to offer the required technologies to research. In text-based research, libraries already have a long-standing tradition as partners and service providers for research. However, the recent methodological process in algorithmic and large-scale text-analysis (‘text-mining’) requires a new understanding how libraries can support text-based research. The following activities show examples from the State and University Library in Göttingen (SUB), Germany.

Text-mining researchers often start by collecting resources and preparing a text corpus, that gives the bases for algorithmic analysis. SUB manages or participate in several national-scale resources that can be made applicable for resource collection and corpus construction: (1) the digital journal archive http://www.digizeitschriften.de (similar to JSTOR but mostly German content, c. 700 journals and 8M pages), (2) the collection of the digitalization center https://gdz.sub.uni-goettingen.de as outputs of nation-wide imaging (c. 15M pages) or (3) National Licenses of publisher’s content (see below).

Only the last case of National Licenses will be taken into account in more detail as it elucidates challenges and novel opportunities particularly well. Relevant changes in the German IPR regulations apply from March 2018: Section 60d UrhG allows researchers to automatically and systematically process multiple works for non-commercial use. Thus, text-mining projects would actually be possible at a much larger scale but in very practical terms, there are very simple (and essential) obstacles that have to be overcome.

- Discovery and set-selection: Based on a basic metadata set, nationally licensed materials can be found through the library catalog and in some cases through federated search1. These

---


2 For example http://finden.nationallizenzen.de – a single index for resources licensed for a consortium.
resources are usually accessible via the publisher’s platform only, except a small part where we provide the hosting interface\(^3\).

- OCR Missing or hidden: In some cases, they are a text database but also sometimes merely a set of image scans with a hidden text layer based on OCR.
- No batch download: in other cases full-text is available but the materials are not available for machine based or scripted downloads.

It is obvious that these nationally licensed resources are designed for human consumption and not for text-mining. In fact, they occasionally restrict the possible machine-based usage. Beside functional and technical obstacles, we also have to gather more experience which concrete use cases are legally possible and which are not: the new regulation gives us the right to mine all corpora we have access to and publishing the results, but we cannot publish the original text.

We will use these challenges as an opportunity to improve text-mining related services provided by the library:

- provide more detailed metadata on the resources
- use discovery systems that simplify set construction
- offer toolsets for corpus construction and analysis: many scalable open source tools and are available
- support collection, conversion, transformation, annotation as a library service
- offer virtual storage and computing resources (up to high-performance and high-throughput computing)
- as a long-term goal, develop an API to all the content we can provide combined with the available text mining toolkit

These services will fill the gap to other activities regarding text-based research, particularly the work on tools for digital editions in TextGrid (https://textgrid.de), the terminology services (see e.g. https://www.cerl.org), and digital infrastructure for the humanities DARIAH (https://de.dariah.eu).

\(^3\) https://nl.sub.uni-goettingen.de
<table>
<thead>
<tr>
<th><strong>STRENGTHS</strong></th>
<th><strong>WEAKNESSES</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>- combined access to heterogeneous content (incl. increasing number of nation-wide access to specific resources)</td>
<td>- gap between text mining based research and text mining service applications</td>
</tr>
<tr>
<td>- locally controlled repositories, created/maintained by libraries</td>
<td>- regular license contracts do not cover text mining</td>
</tr>
<tr>
<td>- libraries have become content creator by supporting digital editions and other textual collections</td>
<td>- missing connection between data, software and research articles for reproducible research (also lack of awareness in researchers)</td>
</tr>
<tr>
<td>- availability of high performance computer</td>
<td>- missing normalization of texts incl. descriptive metadata</td>
</tr>
<tr>
<td></td>
<td>- licenses that restrict sharing of text mining output undermine the ideals and even the practical motivations of academia</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>OPPORTUNITIES</strong></th>
<th><strong>THREATS</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>- modification of legal parameters towards text and data mining for all licensed resources</td>
<td>- distribution of text databases:</td>
</tr>
<tr>
<td>- special clauses for educational sector</td>
<td>- licensed resources have to be available via API (with support for Corpus Query Language, or something newer like GraphQL or something very different in some years)</td>
</tr>
<tr>
<td>- tools and workflows are available</td>
<td>- IP based access to images scans together with a full-text index is not sufficient</td>
</tr>
<tr>
<td>- tradition of text-based services in libraries as a basis</td>
<td>- content user should be able to use text databases for further annotation/metadata enrichment and set up own APIs to prepared data</td>
</tr>
<tr>
<td>- increasing significance of full-text search and transformations based on texts, like visualizations of content words, bag-of-words models or NER-enriched material</td>
<td>- copyright issues still present (at least in Germany; e.g. »internet as a corpus«)</td>
</tr>
<tr>
<td></td>
<td>- restrictive licenses weaken subsequent claims to accessing scientific corpuses under open terms</td>
</tr>
</tbody>
</table>
FORUM STATEMENT

I believe that we need to move away from the model of massive downloads and custom normalization for most text mining activities; the overhead from this is way too high. We need some simple APIs that will support, say, 80% of the text mining requirements and that are uniform across various publisher corpora. The publishers should be more comfortable with this because they don’t have to relinquish control of their corpora. It’s important that we get this into the basic functionality of licensing access to journals, and not let it become an expensive extra-cost add-on.

It’s clear to me that computational approaches to the scholarly literature are going to be absolutely essential to dealing with the explosion of the scholarly record. It’s out of control, and there’s no other alternative.
<table>
<thead>
<tr>
<th><strong>Strengths</strong></th>
<th><strong>Weaknesses</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• This has a very high potential payoff.</td>
<td>• The legal ambiguity is immense and needs to be reduced.</td>
</tr>
<tr>
<td>• Not only is it important in its own right, but given the absurd expansion of scholarly publishing, it offers just about the only possibility of managing this massive over-publishing behavior.</td>
<td>• Approaches that rely on fair use are not attractive; they assume a lawsuit and then an (expensive) defense to establish precedent. It would be much better to get contractual clarity when we can.</td>
</tr>
<tr>
<td>• It could add value to licensing access to journals.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Opportunities</strong></th>
<th><strong>Threats</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• We need to define standard APIs to allow many users to routinely mine journal content. This should become part of a standard site license agreement.</td>
<td>• Text mining could become an extra cost item as part of a journal licensing.</td>
</tr>
<tr>
<td>• Differential mining (eg comparing public and private knowledge) could become a very powerful tool.</td>
<td>• Commercial sector (esp. Pharma, life sciences, biotech, etc) could shape the arrangements for everybody; they have so much more money than the humanities, for example, that this would be very damaging.</td>
</tr>
<tr>
<td>• This is essentially the only way we can manage insane levels of literature growth</td>
<td>• The EU is much more organized than the US, and is having some conversations with the STM publishers. They could potentially establish the standards that the US would subsequently have to develop strategies to support or overturn. The US needs to develop strategies for collaborative leadership with the EU.</td>
</tr>
<tr>
<td>• Public access mandates in the US are very unclear about text mining. Re-inforcing and clarifying this could be very helpful, if government support can be found.</td>
<td></td>
</tr>
</tbody>
</table>
FORUM STATEMENT

I am a content miner of the scientific literature and I also develop Open tools though my non-profit company contentmine.org.

I have been mining scientific literature for 40 years as a means of doing research and have developed many novel tools, some new science. We now have 5000 new articles each day (3/minute!) so machines are essential. But relatively few people see the point of using machines to read the literature so we are stuck with approaches that are 50 years out of date. It is incredibly hard to get this point across:

"PMR: we spend 500,000,000,000 USD (half trillion) on publicly funded research but up to 80% is wasted because of the way we capture knowledge and share it. Closed knowledge means people and the planet die".

"Stop being dramatic; the real point of research and publication is to boost my career metrics".

I have been pushing this for 20 years - knowledge must be fully Open, en masse (not just for human eyes) and machines can help us extract, filter, combine, transform, repurpose... and still we publish PDFs (one of the most effective ways of destroying information).

I do science using this. Recently we read 5000 papers and extracted the individual images and turned them into a microbial evolutionary tree. I have the right to do this in UK but every day I have to worry about being sued for doing legal research or cut off by my University.

Therefore I have to advocate and be activist. I’ve been heavily involved in copyright reform in EU which was meant to lead to freedom, but is actually making things worse. Only “public interest research institutions” will be able to do “non-commercial” TDM from closed literature. Am I legally allowed to do this? I probably have to get sued in Europe to find out.

And because there are so few people doing it openly, TDM has become an underground activity. We don’t tell people what we are doing in case the publishers tell our Universities to close down our research. We work with home-grown tools which need investment for wider use, but such investment is only likely to come from organizations who wish to control and exploit published science.

At present I am developing Open tools without public funding. These ought to be valued by Universities and their libraries. But Libraries are scared of publishers and act as police to stop “illegal downloading” of publicly funded closed science. I gave an invited lecture at IFLA (Int Federation of Library Assocs) last autumn. It was very well attended but when I asked for people wanting to follow this up, zero interest.
Early career researchers understand the need and value of TDM and we have managed to create a small community of them. Some have already fallen foul of universities who support publishers rather than researchers. If we want to move forward, ECRs is the area we should concentrate on. So some proposals:

- Universities should make public statements valuing TDM and undertaking to support it. This included funding for tool development, and courses.
- Universities should challenge all publishers who cut off access due to legitimate TDM and assert breach of contract.
- All graduate courses (not just Compsci or LibSci) should include material on TDM - technology, resources.

Universities should challenge the publishers’ control and assert “The Right to Read is the Right to Mine”. If that means a public showdown, good.

And, if positive action comes out of this meeting ContentMine is interested in being part - tools, training, research projects, etc.

---

PMR>I published a blog post on TDM 4 years ago - here is Susan Reilly's summary: https://libereurope.eu/blog/2014/02/06/a-scientist-s-take-on-the-new-elsevier-tdm-policy/

SR>“Content Mining: Why you and I should NOT sign up for Elsevier’s TDM service” is the title of Peter Murray-Rust’s latest blog piece. In it he puts Elsevier’s newly launched text and data mining service and accompanying policy under the microscope. LIBER has republished an excerpt here.

PMR>In the last few days Elsevier has announced their policy on Text And Data Mining (TDM). I use the term “content mining” as I wish to mine every part of published content (images, audio, video) and not just text.

This post contains a lot of material (from Elsevier and my comments) so I’ll try to summarise. Note that Elsevier’s material seems inconsistent in places (common with this publisher). I have had to go behind Elsevier’s paywall to find one statement of agreement and rights and it is probable that I have not found everything. In essence:

- Elsevier asserts complete control over “its” content and requires both institutions and individuals to sign licences.
- Elsevier is the sole author and controller of the policy – there has been no Open discussion or agreement with scholarly bodies.
- Libraries have to – individually – sign agreements with Elsevier. There are no details of these policies or whether they entail additional institutional payment. It is also possible that Institutions may be asked to give up content-mining rights in return for lower overall prices. (Libraries have
universally and unilaterally given away all these rights over the last decade and support publishers to forbid machine access to content).

- Researchers have to register as a developer (I think) and ask permission of Elsevier for every project they wish to do. It is not clear whether permission is automatic or whether Elsevier exercise control over choice and scope of project (they certainly did when I “negotiated” with them).
- Researchers can only access content through an Elsevier-controlled portal. They have to register as a Developer and get an APIKey (conflicts with “sign a click-through licence”).
- Researchers can only mine text. Images are specifically prohibited. This is useless for me – as I and colleagues are mining chemical structure diagrams.
- There is no indication of how current the material will be. I shall be mining the literature an hour after it appears. Will the API provide that?
- The amount that can be republished is often useless (“200 characters”). I want to build corpora (impossible); vocabularies (essential to record precise words – impossible); chemical names (often > 200 characters so impossible). Figure captions (impossible).
- The researchers must commit to a CC-NC licence. This effectively kills downstream use (I shall use CC0). It also trains them into thinking CC-NC is a “good thing”. It isn’t.
- If a researcher has a LEGITIMATE collection of papers that they wish to mine (say on their hard disk) they are forbidden. They have to go to each publisher (if this awful protocol is promoted elsewhere) and find the API and mine the individual papers. Absurd.

This is licence-controlled TDM. The publishers tried very hard to get Europe (Neelie Kroes) to agree to licences for TDM (“Licences for Europe”). They failed.

They tried to stop the UK Hargreaves process exempting data analytics from copyright reform. They failed.

The leading library organizations and funders such as the British Library, JISC, LIBER, Wellcome Trust, RCUK are united in their opposition to licences. This is simply Licences under another head. The danger is that University libraries – who have signed these restrictive clauses for years will continue to sign them.

DON’T.
Don’t take my word for this. Ask the BL, or JISC or LIBER.

BUT DON’T SIGN ELSEVIER’S TDM.

And:
YOU DO NOT NEED ANY API.
APIs make it HARDER to mine. We are releasing technology that will work directly on PDFs. It’s Open Source and works. And others are doing the same. If every publisher came up with a similar process it would make the burden of mining huge. This is probably what some publishers hope.

*The full text of Peter’s blog piece, plus supporting analysis, is available here:*

http://blogs.ch.cam.ac.uk/pmr/2014/01/

<PMR>

The history is that the Hargreaves legislation allowing for TDM had recently been drafted and agreed by the UK governments. Elsevier created their own T&C, without reference to the community and inconsistent with the terms of Hargreaves. Every time they released it, I challenged it, and I think this resulted in 5 or 6 versions.

At a scholarly meeting about TDM in 2014, an Elsevier staff member (Gemma Hirsh) attended and gave an unscheduled presentation on Elsevier’s TDM policy. She made several statements which were factually untrue, essentially that the law didn’t apply to Elsevier and that they could stop researchers mining unless they used Elsevier's API. The "discussion" is recorded in detail at


Professor Charles Oppenheim, an expert in the area, stated that GH was absolutely wrong.

The T+C still contain a clause forbidding researchers to publish any mined data that would compete with any Elsevier product (not just the manuscript) (https://www.elsevier.com/__data/assets/pdf_file/0012/102234/TDM-sign-up-short-form.pdf)

>>You are not allowed to:

Utilize the TDM output to enhance institutional or subject repositories in a way that would compete with the value of the final peer review journal article, or have the potential to substitute and/or replicate any other existing Elsevier products, services and/or solutions.

I am a chemist; Elsevier has a major chemical database. I am therefore forbidden to publish any chemistry in mined material.
<table>
<thead>
<tr>
<th>STRENGTHS</th>
<th>WEAKNESSES</th>
</tr>
</thead>
<tbody>
<tr>
<td>● (some) governments recognize the critical value of TDM</td>
<td>● By default everything is copyright. This means that the miner has to explore legal aspects of every document they wish to mine. This is a killer.</td>
</tr>
<tr>
<td>● Many organizations in free and Open Knowledge Culture are passionate advocates and activists.</td>
<td>● Many “content owners” of scientific articles (companies and learned societies) regard content mining as under their legal and contractual control and by default forbid it, or impose impossible conditions.</td>
</tr>
<tr>
<td>● Some projects (e.g. EU H2020 FutureTDM) are committed to developing business and political cases.</td>
<td>● content owners (including learned societies) frighten miners either legally or through their subscribing institutions.</td>
</tr>
<tr>
<td>● Some projects (e.g. EU H2020 OpenMinted) are developing Open resources for TDM</td>
<td>● Universities and their libraries have done nothing effectively to support researchers. Many libraries are scared of publishers and act as “journal police”.</td>
</tr>
<tr>
<td>● TDM is widely used by mega-corporations (Google, IBM...) and some of the technology and impetus trickles down.</td>
<td>● Megapublishers restricting large parts of the scientific literature unless they control it.</td>
</tr>
<tr>
<td>● There is a real business need for businesses and citizens to have tools and services to mine scientific articles.</td>
<td>● Copyright maximalists (e.g. in EU) who want everything to be copyrighted for the benefit of industries</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OPPORTUNITIES</th>
<th>THREATS</th>
</tr>
</thead>
<tbody>
<tr>
<td>● Funding bodies who want their output to be used.</td>
<td>● Universities who do not support researchers</td>
</tr>
<tr>
<td>● (possibly) megacorporations who wish to develop new technology by collaborating with miners</td>
<td>● technology being developed only by megacorporations in Silicon valley so no one else can play.</td>
</tr>
<tr>
<td>● commercial investment. BUT the “non-commercial” legal restriction makes it very hard to make this happen.</td>
<td>● Hopefully some universities who have had enough of being trashed by publishers and want to develop new ways of doing research and teaching. But I see no sign yet.</td>
</tr>
</tbody>
</table>
FORUM STATEMENT

I have been advocating for many years within the research library community for a stronger position and role for libraries in providing access to limited-access datasets for mining research. One of my concerns has been that, if libraries are to maintain our central place as hubs or connectors between researchers and information resources, we need to be much more forward-thinking with regard to computational research. As I wrote last year, “If research libraries don’t get on board in a big way with ‘content as data,’ then we will be consigning ourselves to niche status within our user communities.” (http://www.against-the-grain.com/2017/10/v29-4-liaisonship-the-law-libraries-supporting-content-mining-research/)

We need to think more broadly, beyond “text and data mining” to “content mining,” a term inclusive of all possible information formats, including text, images, text as images, audio, video, and more. All of these formats are already being mined by researchers.

We need to focus on securing access to content, and worry less about quality of metadata and the current tools used to mine. Many of these issues will resolve themselves as artificial intelligence (AI) improves and more and easier to use tools are developed such that fewer researchers will require technical expertise or support.

Similarly, we need to strategically focus on the current and near-future needs of high-end researchers rather than those of the majority of our current users. As we work to create tenable and sustainable for access to limited-access datasets, we need to avoid fettering the true needs of the overall research community with the needs and desires of more routine or applied mining research and/or of pedagogy.

From a legal standpoint, we need to be active in current copyright conversations, of course, as well as pay careful attention to proposed legislation impacting computational research. Although I am not a legal expert, I am hopeful that we might discuss with those who indeed are and are a part of this forum possible avenues forward, especially under Fair Use and perhaps opened up in the wake of the Google Books legal actions. It seems to me that this latter especially might open up greater possibilities for computational research and computer “reading” of texts with less need for licensing.

One of the greatest business needs is to establish a standard model, fair for both libraries and content providers, for content that is typically subscribed rather than sold in perpetuity.

Perhaps the greatest expressed need across disciplines is for access to in-copyright newspaper content. While I understand present-day challenges to the newspaper industry, it is particularly vexing that virtually all of the “Fourth Estate’s” record of the last hundred years is closed to computational research. If newspapers are, as they themselves claim, essential to the U.S. practice of democracy, then it is in the public interest that this content should be readily available for mining by scholars.
IMLS National Forum
Data Mining Research Using In-copyright and Limited-access Text Datasets
Prepared by: Darby Orcutt, North Carolina State University

We need to firmly establish an industry norm whereby the library serves as a trusted intermediary between researcher and rights holder, and academic freedom is upheld by clear avenues of access to datasets that do not require prior disclosure or approval of any aspect of research or identity (beyond the fact of institutional affiliation). In other words, computer-assisted (or “distant”) reading should be supported in the same ways as human (or “close”) reading.

Given the complex relationships of today’s university, the many sorts of public-private (or institution-industry) partnerships that may exist, and even student entrepreneurship possibilities, we may, in fairness to rights holders and to avoid legal issues, need to define more specifically what constitutes “commercial” use of data, as most agreements for accessing limited-access data seem to include some sort of “non-commercial” stipulation.

Research libraries need to embrace now our growing constituency of computer-assisted and even non-human users. As a research library community, we should immediately stop privileging only the traditional individual human reading model of research. While I certainly do not expect that practice to go away, computational research will have a profound impact upon and establish itself as a vital part of every academic field, even in those disciplines in which it remains a minority activity. With the advent of the internet, libraries could have drawn a line and identified ourselves only with print or physical information objects. I’m glad we didn’t. At this juncture, I hope to see us again naturally extend our understanding of our traditional mission, and thus work to support the full range and future of computational research, with the next and most urgent step being to secure access for mining purposes for researchers to significant limited-access datasets.
### Strengths

- Libraries are the traditional brokers of information for the research community, the traditional hub for high-quality content; providing access to data for mining is a very logical extension of our mission.
- Libraries already have business relationships with many/most of the entities controlling desired limited-access data.
- Libraries have expertise in metadata and interoperability, which are key to providing data access.
- Libraries are the equal-opportunity, democratic, and most cost-effective units of our institutions to serve as the research infrastructure for information resources increasingly valuable across disciplinary lines.

### Weaknesses

- The term “text and data mining (TDM)” should NOT be used. I advocate the much more inclusive term “content mining.” We need to pay attention to ALL formats, including text, data, text as images, images, audio, video, and more. These are being mined now, and will be increasingly in the future.
- There is a lack of reasonable and standard business models for ensuring researcher access to limited-access content for mining, particularly for content that is generally leased rather than owned/licensed in perpetuity.
- Much/most of the academic library community seems to fundamentally misunderstand the big picture with regard to content mining research, perhaps due to an in-group bias based on the researchers at their institution or who approach their library for support. In negotiating models and articulating strategy as a community, we need to prioritize the current and near-horizon needs of high-end researchers rather than those of the majority of researchers, instructors, and students. The cutting-edge research will drive the understandings, tools, and future of mining activities by the rest.
- Our libraries and larger institutions have antiquated and limiting policies regarding privacy and IRB matters.
# Opportunities

- Libraries can lead rather than follow when it comes to articulating models and enabling computational research.
- We can help create structures for interoperability and AI access to research collections/information.
- We can establish ourselves as essential to the cutting-edge of research support.

# Threats

- If academic libraries don’t pivot quickly on many of these issues, our researchers will conduct their research apart from the library or even apart from the university. (This is already happening).
- Requirements by vendors (including ways of implementing RA21) could threaten academic freedom, especially for researchers in this area.
- Commercial entities could outpace us in establishing standards, and particularly with “lock in” practices.
- Commercial entities will seek to monetize access for mining purposes (effectively discriminating against computationally assisted research versus traditional human reading).
- The newspaper industry will continue to preclude or seek to monetize access for mining research, which is especially problematic given their self-described role as the “Fourth Estate” and essential to our democracy.
- The library community will not fully embrace the new role of serving non-human users.
FORUM STATEMENT

Working in libraries affords many opportunities to engage challenges associated with text and data mining (TDM) limited access datasets. Challenges afford themselves across interactions with diverse disciplinary communities at institutions with varying resources. Increasingly, data requested by research communities fall outside the traditional purview of library collection development and research support. Solutions to TDM challenges are few given in understanding and misaligned values. Continued activity in light of these factors fosters an environment where weaknesses and threats are many and seemingly tractable opportunities are few. In the space of this brief statement I will introduce four challenges: (1) underdeveloped and inconsistent content provider effort to meet text and data mining need, (2) misaligned values reinforced by ambiguous and/or overly restrictive content provider terms, (3) debt incurred by technical abstraction, and (4) purposeful technical opacity.

To a large extent content provider efforts to meet TDM needs are underdeveloped. Wide variation in mechanisms for provisioning limited access datasets evidence that underdevelopment. From one provider to the next there is little indication of concerted development activity which leads to a wild and wooly data access environment. This situation places undue burden on users and library service providers. Variation in development of application programming interfaces, where they are available at all, creates user learning curves and library service debts that are difficult to anticipate let alone bear. Variable positions on data documentation inclusion – that is to say which accompanying information a content provider deems in scope for data acquisition can be difficult to anticipate. For example, there have been instances of content providers refusing to share document type definition files along with a bulk data purchase given a claim of “intellectual property”. In a similar vein, content providers have been hesitant to provide documentation that details the component parts of a dataset where they are constituted from multiple publications. Where that data is provided upon request, post time delay, it can appear in forms that needlessly obfuscate data evaluation – e.g. a PDF of tabular data describing the dataset. Further, data documentation arising from a user paradigm wherein bibliographic description is fit for purpose is less well matched to a data driven paradigm.

Underdeveloped and/or misaligned content provider values relative to TDM are readily apparent at the level of terms associated with limited access datasets. It is not uncommon to be referred by a content provider to a provision for sharing “snippets of text” when inquiring on the ability to share data utilized and/or generated in the course TDM research. The ability to share data has a direct correspondence to the ability to conduct transparent research oriented toward reproducibility or replicability. Reference to a “snippets of text” provision serves to inhibit library guidance in the use of resources for TDM given the ambiguous nature of the provision. Some institutions will move forward with that ambiguity and others will not. Guidance of this kind further compounds the long-term ability of university, domain, and scholarly repositories to acquire and preserve TDM research. In a slightly different vein, engagements with non-
traditional content providers that vend social media data entail onerous and cost prohibitive terms that impact the viability of library support and broader research potential. For example, a National Science Foundation proposal focused on studying linguistic change in the Midwest was pursued by 3 institutions. In order to support this research, the principal investigator sought to acquire data through a social media data vendor. The data vendor required each participating institution to acquire the data and when asked about sharing data for the purposes of peer review suggested that an additional acquisition would need to occur. According to these terms the possibility of moving forward was essentially dead on arrival.

Finally, efforts to enable TDM with limited access datasets have been challenged by degrees of technical abstraction or intentional technical opacity that can be difficult to weather. For example, HTRC’s data capsule project is a wonderful sign of progress but abstraction via the non-consumptive paradigm holds the potential to create a technical and methodological debt that compounds researcher effort in the course of peer review and library effort to promote the viability of the service. On the content provider side, a related tension with very different motivations, can be observed in proprietary machine learning approaches to generating open data and scholarly metrics. Without better understanding of the method generating the data and the metrics serious questions arise as to the viability of that data. Furthermore, this data constitutes a potential threat insofar as administrative pressure is set to push building upon those data in the pursuit of demonstrating scholarly impact. This essentially creates a lock in with seemingly open data vended on an opaque, proprietary foundation. It is akin to a “DPLA as platform” approach without the corresponding commitment to openness.

These are just a few, high level glosses of issues I have encountered in the library community. I look forward very much to learning from my peers at the IMLS supported, Data Mining Research Using In-copyright and Limited-access Text Datasets national forum.
## STRENGTHS

- Broad thematic and temporal coverage
- Centralized library community attempt to enable non-consumptive use
- Decentralized library community attempt to support TDM work

## WEAKNESSES

- Service lock-in > data acquisition
- Access facilitated by technical abstraction
- Technical opacity
- No data acquisition clearinghouse
- Access enabled by corporate privilege
- Lack of ethical provisions for data
- Unsustainable costs – e.g. double dipping for data that haven’t been enhanced
- Ambiguous terms
- Overly restrictive terms

## OPPORTUNITIES

- Align terms with normative or near-normative conventions in research communities that support reproducibility and/or replicability efforts
- Centralized acquisition clearinghouse for data – ensure broad access on terms favorable to research; too many examples of multiple purchases for datasets across a university

## THREATS

- Service lock-in > data acquisition
- Technical opacity
- Costs inhibit multinational research collaboration
- Costs inhibit multi-institutional research collaboration
- Research rich vs. resource poor – resort to consortial solutions only extend the life of an unsustainable model
FORUM STATEMENT

I assume the role of a Librarian, having served in numerous public services and technical positions within the Cornell University Library for over 30 years. Currently as Digital Curation Services Lead, I coordinate Cornell’s involvements with HathiTrust, which includes oversight of deposits of digitized content, correction of bibliographic records, coordination of permissions projects aiming to expand access, and coordination of our involvement in HathiTrust initiatives. I also monitor developments at the HathiTrust Research Center (HTRC) for affordances for our scholarly community’s projects and classes. I fill the role of text analysis expert in Cornell’s Digital Humanities Collaborative (DHCollab), fielding referrals for consultations with scholars about the text analysis projects they envision. I help acquaint them with the methods and strategies that might be fruitful for their goals. Sometimes I find resources that assist them in bootstrapping to the appropriate skill level to conduct their research; other times I help make introduction to other researchers or assistants with complementary skills in hope of catalyzing effective collaborations. I even might directly assist a project technically if the effort is scoped to fit within my skill and available time. Often I act as a translator between technical staff and the researcher to demystify jargon and to facilitate productive conversation. I regularly lead workshops and classroom experiences in for-credit courses to introduce methods of text analysis to humanities scholars. I am not a librarian by training; I most often draw upon my experience as a systems and network administrator (for example, technical instruction, data management, markup languages, coding and translation of technical aspects of work to non-technical audiences). Through my management of Cornell’s Google digitization project, I became familiar with the structure of digitized books and their potential for algorithmic reading, and was captivated by the possibility they held for transforming humanistic scholarship, not just at the scale of books that could be read (albeit distantly), but also the very nature of the scholarly questions we can now hope to productively ask and answer.

Even as traditional methods of close reading continue to provide value, they will be supplemented by complimentary data-driven methods. It is completely necessary to anticipate and provide for this expansion in methods within the humanities, because the evolution brings with it some basic tensions. Some of these tensions come from the cultural shift underway as the humanities incorporate methods and strategies long embraced by the sciences. Scientific inquiry often moves forward in teams of experts with diverse skills that coordinate to break new ground, which is alien to the traditional approach of the humanities where individual scholars read, reflect and debate. Similarly, scientific methods seeks to gather and quantify data, drawing conclusions through statistical significance, which is a far cry from discussions about tone, voice and plot. The methods and strategies of computational approach to text require that humanists borrow from the sciences, and the resulting cultural awkwardness can be palpable as scholars try to navigate the different modes of working and thinking, the diligence required around documenting, sharing, and interpreting results, and the skill gap in tools required to make effective progress.
The nature of limited content adds an additional layer of challenge, for it is rare that a researcher is satisfied with non-restricted corpora alone. Even if primary texts of interest are open, it is likely that important critiques for these texts are not, and a scholar restricted to open corpora may find himself working on the less-interesting portion of a project (for instance, interpreting a primary text while ignoring the manner in which that text has been interpreted through time). Scholars who focus on the recent age are doubly hampered since copyright law adversely affects access. Not least of all, even material that is not affected by copyright can get locked behind paywalls through the vendor platform by which it is made available. If we don’t find ways to release restricted content, major portions of human intellectual thought remain locked away, obtained only by the few clever enough or rich enough to obtain it.

Fortunately, the academic research library is a natural partner to support scholars in the humanities through these evolutionary shifts and is well-positioned to assist in freeing limited corpora for wider use. Libraries have long been a gathering place for interdisciplinary work because they support a wide range of disciplines through resource collection and procurement, and staff expertise. Collaboration and support of departmental goals are key traditional offering with new meanings as the library performs the role of incubator and collaboration broker. The “lab” model of the sciences is springing up in libraries to help bootstrap teams that benefit from both disciplinary expertise and technical strengths. The library is a natural place for skills workshops in subjects like data analysis, bibliographic metadata, markup languages, mapping skills, taught as often by specialists in the formerly hidden ranks of IT and Technical Services as often as they are by subject area experts. Libraries already procure digital resources, make them discoverable, and assist scholars in their use. But now on a fairly routine basis we procure limited corpora through special arrangement with vendors, who extract the text from their vended products to a form useful for computational analysis. The resulting text corpora is described and made available (discovery and access) to scholars as appropriate to the licensing agreement. But our descriptions are not limited to the data set as a whole - we also have used our forensics unit to describe the internal structure and filenaming within the set, so as to provide guidance documentation that allows the researcher to more quickly understand organization, codes and internal relationships within the data, and thus more easily and efficiently use the acquired corpora. In an era where some question the continued relevance of the library, I see an open opportunity to revitalize the library as a hub of scholarly collaboration and cross-pollination, curiosity and discovery, instruction and skills acquisition, assuring the traditional mission of the library is never more vital than in this digital age.
**STRENGTHS**

- Researchers who are curious and willing to engage the edge can use new methods to discover or confirm patterns in text.
- Libraries can exercise their core mission to describe and make available all acquired text corpora as library assets. We paid money for them, so we should retain them and make them available as appropriate.
- Librarians and library staff can contribute coordination of projects and serve as an information and skills brokerage. The library can be a major coordination hub in text analysis research.
- Library IT and metadata staff can assist researchers to develop skills necessary for text analysis.

**WEAKNESSES**

- Algorithmic analysis of text requires skillsets that are often beyond the average humanities researcher. Technical skill development is essential.
- Access to full-text of in-copyright data often requires extraordinary and non-routine steps on the part of library staff and researchers, as well as additional money (beyond subscription fees). These have chilling effect for this use.
- Once ordered from vendors, corpora can take many months to be available, and can come on any type of media.
- Text corpora are not available in a standard format or markup, and arrangement of the data is idiosyncratic to the vended product.

**OPPORTUNITIES**

- Introduces systematic, algorithmic logic into humanities research. This expansion of type of logic is arguably the most important underlying development in our culture in the last two decades, and I predict, a basic understanding of algorithmic approaches are already essential to a well-rounded education.
- Serves as a platform that brings staff and researchers with a diversity of skills together. Effective coordination of these skills brings immense value to the research community.
- A researchers desire to digitize in order to create a text corpus can be leveraged to create access copies for close reading to appropriate audiences, and preservation copies.

**THREATS**

- Access to text is often licensed, and possibly subject to disappearance if licensing is not kept current. This can be true even of text that is not in copyright if access if proprietary.
- Once acquired, in-copyright text corpora often lack a plan as a library asset (discovery and appropriate access by researchers). Acquired datasets can essentially disappear due to poor organization.
- University may lag behind library and/or research needs in terms of program support.
FORUM STATEMENT

I am a researcher who regularly uses data to study things like the discourse of prestige in the novel, institutional bias in academic publishing, and gender bias and social networks in the contemporary novel. Most of my research focuses on text-data within documents so my primary interest is in making sure that digitized resources are available as full text. I also engage more recently with bibliographic data and abstract data in order to study trends within my discipline of literary studies.

In all of these cases there are legal limitations to what I can do with the data. In all cases I believe that I am not allowed to share the data and in some cases I am not sure whether I am even allowed to be analyzing the data. This legal ambiguity causes a great deal of uncertainty and disincentive in my work and makes it harder to collaborate. It also violates basic norms of research which are predicated on the transparency and reproducibility of scientific research.

My vision is that rather than solve all in-copyright problems ever for researchers I would like to see a concerted effort to make the research library a place that actually facilitates rather inhibits digital research. Right now it feels like the library’s job is to sign agreements that limit researchers’ access to data. I would like to see a (inter)national strategy put in place that empowers researchers to study the data that is *in* their libraries. We have reached a historic tipping point where the library’s mission of making information accessible and circulatable is no longer being achieved. It’s time to reverse that trend. Instead of placing limits on the number of materials accessible and the ways in which that data is accessible (via proprietary and often very poor interfaces) I’d like to see infrastructure developed for data mining library data. Why have we privileged only one kind of access (reading) at the expense of another (data analysis)?

I come at this as someone who finds it very challenging to understand how our current state of affairs was arrived at and why change is so difficult. I hope to have a better understanding of the issues involved by the end of the conference as well as some hope that change may indeed be possible.
## Strengths

- There is an amazingly large and growing trove of digital material to work with
  - This can include historical documents through services like Gale and HathiTrust
  - Or it can be the output of scholarly articles and monographs
  - Or contemporary novels.
- There are amazing new tools being developed everyday through open-source computing like R that allow for quick learning and applied uses to study language and texts at large scale

## Weaknesses

- All of that great digital material is subject to some kind of copyright limitations with the exception of a tiny, tiny slice
- HathiTrust limits based on Google agreements
- Academic monographs and journals are not freely accessible for mass download through the library
- Contemporary novels aren't either

## Opportunities

- HathiTrust Research Centre Data Capsules that allow for non-consumptive analysis of in-copyright texts
- New research tools to improve analytical methods

## Threats

- Lawyers, librarians, publishers, journal editors -- anyone invested in sending take-down notices about data for research purposes
FORUM STATEMENT

My research on copyright law and its application to copy-reliant technology played a non-trivial role in securing the favorable fair use rulings in the Authors Guild v. HathiTrust and Authors Guild v. Google cases. I continue to write about copyright and non-expressive/non-consumptive use and champion the cause of the digital humanities.

I also use text mining tools in my own research, for example, in The New Oral Argument: Justices as Advocates, Notre Dame Law Review, Vol. 94, (forthcoming 2019) Tonja Jacobi and I conduct a comprehensive empirical inquiry of 55 years of Supreme Court oral argument using text mining tools. Using various metrics we show that oral arguments changed dramatically in the mid-1990s in response political polarization.

Lawyers are notoriously verbose, so in addition to my one-page SWOT analysis, I have composed a 15 page draft paper on The Legal Infrastructure for Research in the Digital Humanities. The paper is not ready for general release, but it can be made available to forum participants. It will also be revised in the coming weeks.
### Strengths

The legality of non-expressive use/non-consumptive use of copyrighted works as exemplified by text mining and other digital humanities practices is now well established under U.S. law.

### Weaknesses

- The precedents confirming the legality of non-expressive use/non-consumptive use are purely creatures of U.S. law.
- Digital works hosted on computer systems may be subject to terms of use restrictions and technological restrictions designed to prevent data mining.
- Researchers should expect that terms prohibiting text-mining and scraping the contents of those repositories are valid and enforceable.
- Gaining unauthorized access to a digital repository might trigger both civil and criminal liability under the Computer Fraud And Abuse Act.
- Some actions taken to make that reproduction possible may well violate the Digital Millennium Copyright Act prohibitions on the circumvention of technological protection measures.

### Opportunities

- As a discipline, the digital humanities is still very much in the pioneering stage where substantial resources are being devoted to assembling sources of data, and developing tools and toolsets.
- Once these tools and practices have been taught to a new generation of undergraduates and PhD’s, i.e., once they have become a normal part of the curriculum in the humanities, we will see them being used in ways beyond our present imagination.

### Threats

- Researchers may exceed the scope of fair use under the Authors Guild precedents.
- The size of snippets or text previews and security issues are the main concern.
FORUM STATEMENT

Legal experts in libraries can provide value to computational text analysis (CTA) researchers through: (A) creation and deployment of pragmatic workflows conveying essential CTA literacies, and (B) advocacy on behalf of researchers for expanded CTA capabilities.

A. PRAGMATIC WORKFLOWS CONVEYING ESSENTIAL CTA LITERACIES

In accessing, building, and working with collections of copyright- or contractually-limited texts, CTA scholars make elections that carry legal impact—yet may not be aware of the impact of choices they make. The requisite legal literacies that CTA scholars must competently navigate are robust—encompassing not just an understanding of copyright and fair use, but also a sense of how a university database license, publisher-provided API (application programming interface), archives agreement, or website “terms of use” fit into protocol for access, collection, and analysis. Guiding scholars through these literacies before they use or build research corpora is key, yet currently few outreach programs assist scholars proactively before point-of-need.

Research libraries can be instrumental in offering this education, particularly as they often employ librarians with legal expertise who can also train others in foundational legal principles. Training materials or services that libraries create ought first to distinguish between the use of corpora “behind a black box” (i.e. where the content is not viewable to the researcher) versus corpora that researchers create themselves, as well as the attendant scenario in which researchers subsequently republish portions or all of the corpus they create. These three categories—precompiled corpora, creation of corpora, and republication of corpora—form a useful framework around which to build a legal literacy workflow such as the following:

1. USE OF PRECOMPILED CORPORA
   a. Address scope of the corpus:
      A researcher should be able to articulate, and understand the reasoning for, the boundaries of her corpus—not simply in terms of the total number of items represented, but also taking into consideration what materials may not be included in the corpus, and whether or not those items were omitted for reasons related to copyright.

   b. Account for mode of access
      While CTA researchers may initially seek to bulk download materials, familiarity with emerging modes of access enabled by fair use (e.g. derived downloadable datasets, secure computing environments, and web-based tools for interacting with a corpus) will open new opportunities for access and analysis in a manner compliant with copyright law. This necessitates understanding why Google Books allows snippet views, or why HathiTrust Digital Library enables search features but restricts full-text viewing. Likewise, it is essential to understand how copyright and contract law’s effects on access may impact a scholars’ ability to answer the research questions at hand.
2. **Creation of Corpora**

A CTA researcher seeking to develop a corpus must rely on additional literacies that integrate a more nuanced understanding of copyright and contract law:

**a. Copyright and fair use rights**

Building on what researchers know from what makes Google Snippets or HathiTrust Digital Library’s text search features possible, researchers should be equipped to consider whether the content of the corpus they build is protected by copyright and, if so whether it would be fair use to create a searchable and annotated database of these materials—tapping into such cases as *Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014); *Authors Guild v. Google Books*, 804 F.3d 202 (2d Cir. 2015); *White v. West*, 29 F. Supp. 3d 396 (S.D.N.Y. 2014); and *A.V. ex rel Vanderhye v. iParadigms*, 562 F.3d 630 (4th Cir. 2009).

**b. Assess means of content access**

*i. Via institutional license agreement*

Researchers should have an understanding of when they are utilizing institution-licensed resources and databases. This requires a subsequent awareness of whether that license permits the creation of a collection, and/or restricts CTA uses that would otherwise constitute fair use. Scholars should also understand that web scraping in violation of a database license agreement might impede database access for other campus users.

*ii. Via website*

Before compiling a corpus via a website, a researcher should consider whether the same content is available through an institutionally-licensed database, because license agreements may expressly allow CTA even if generic “Terms of Use” from a website bearing that same content do not. If a researcher is using materials on the Web and not through an institutionally-licensed resource, the researcher should: (1) understand the scope of permissible uses defined by a website browsewrap or clickwrap agreement that the user may be deemed to have entered into; (2) understand the formal web services for legal access to web content (e.g. the availability of an API or other programmatic or bulk access point); (3) consider best practices concerning programmatic access, including limitations and prohibitions documented in a site’s robots.txt and “robots” <meta> tag.

*iii. Via archives, museum, or library special collection*

All researchers, CTA and consumptive readers alike, should be aware of any use or republishing restrictions they may be asked to accept when acquiring copies of materials from library special collections, archives, or museums. These conditions may be passed on to researchers if the archives or museum has signed donor agreements that restrict the reuse of records—even if doing so would constitute fair use.

3. **Corpus Publishing**
Publishing the content or the annotated database that a researcher has created can help other scholars test their own algorithms, and provide raw material upon which to conduct research and testing. Yet, it is often in the republishing of corpus content that the limits of fair use are reached, or the bounds of license agreements exceeded. CTA researchers should thus separately undertake a fair use and contract analysis (especially of database license agreements) if they intend to publish excerpts or content from the corpora they assemble.

B. ADVOCACY ON BEHALF OF RESEARCHERS
Legal experts and libraries (and legal experts in libraries) are well positioned to advocate for improving researcher workflows. For instance, they can participate in amicus briefing or commentary to help guide courts or legislatures in shaping critical issues of fair use and license agreement interpretation surrounding CTA uses. They can also advocate for scholars when negotiating with content providers, so that institutional license agreements include contract terms that promote CTA. Finally, legal experts and librarians can help internationally-collaborating scholars streamline their understanding of jurisdictional variations of CTA legal principles.

*This Forum Statement is based on collaborative work by Rachael G. Samberg (Scholarly Communication Officer) and Cody Hennesy (E-Learning Librarian) at the UC Berkeley Library. It briefly encapsulates a workflow that we first proposed in “Law & Literacy in Non-Consumptive Text Mining,” a chapter currently under review for a forthcoming book, Copyright Conversations: Rights Literacy in a Digital World (ACRL).
### STRENGTHS

- Understanding how complex legal landscape affect computational text analysis (CTA) —most importantly with copyright and contract law, but also inclusive of privacy and publicity rights which can factor into the republishing of CTA corpora for downstream use.
- Familiarity with how and what researchers want to publish, and combining that with legal expertise to develop a pragmatic workflow for CTA scholars.
- Ability to communicate, train, and develop resources for researchers regarding core literacies affecting their ability to conduct CTA.
- Ability to provide tailored guidance for individual researchers’ context- and content-specific questions.

### WEAKNESSES

- Limited proliferation of literacies trainings for researchers working with a pre-existing corpus or building their own corpus (e.g. Researchers scraping content for which terms of use constrict what would otherwise be fair use).
- Not acting as counsel to the university or as researchers – limiting ability to provide “answers” to certain fair use or contract term inquiries.
- CTA support constituting only one portion of overall support for scholarly publishing that is necessary to create and dispense (limiting time and resources we can devote to supporting scholars).
- Shifting boundaries of fair use, and ensuring that guidance to researchers about scope of what constitutes fair use remains current.
- Getting access to suitable content—e.g. DRM-protected content, or content that is expensive to subscribe to—without having research compromised by the limitations.
- International publishing and incorporating international CTA rules.
## Opportunities

- Guiding researchers in fair use (and how to transform content), public domain, and copyright licensing (e.g. Creative Commons) landscapes
- Literacies training for researchers working with a pre-existing corpus or building their own corpus. (e.g. Guiding researchers through landscape of understanding why they are unable to view text; or if they are creating their own, what they are allowed to download or republish).
- Training service providers to support researchers—lessening the load of providing guidance to thousands of users on case-by-case basis
- Creation of an implementable, practicable workflow for CTA researchers
- Advocacy to help guide courts or legislatures in shaping CTA space, and advocacy when negotiating with content providers to include contract terms that promote CTA
- Libraries can use this work to become further integrated into research of users, enabling libraries to assist throughout the publishing lifecycle
- Testing boundaries of “transformativeness”
- Finding ways to facilitate scholarship even when IP or contracts impose limitations; the answer is never “no, you can’t do this”

## Threats

- Database license agreements or website terms of use that attempt to curtail uses of content that would otherwise constitute fair use but for the contract terms superseding copyright law.
- Confusion for researchers as to the enforceability of website Terms of Use and what to tell researchers about terms of use. (This is because some jurisdictions recognize that website “terms of service” can constitute a valid “browsewrap” agreement—as opposed to a “clickwrap” which solicits assent; however enforceability of “browseswraps” may need to be determined on a case-by-case basis given the differences in ways that “Terms of Use” are displayed.)
- Economics of scholarly publishing system incentivizing database or content providers to restrict usage rights
- Potential legislative changes or case law limiting CTA on the basis of copyright or contract law
- Researchers not wanting to involve library if they believe library is © or license agreement police
Forum Statement

Elsevier’s Text and Data-Mining Policies

As a content provider, I understand the complexities of opening access to databases to permit text and data-mining. Data content integrity is one of the core values offered by content providers. To ensure the content retains its integrity, access needs to be carefully optimized. Elsevier is committed to working with individual researchers wishing to access content to perform various analytical studies so researchers’ needs are met. Elsevier develops its resources, including content databases, to support researchers as it wishes to aid the improvement of research performance. It uses a license-based approach for text and data-mining for non-commercial research purposes. Its policy for non-open access content may be found at: https://www.elsevier.com/about/our-business/policies/text-and-data-mining. Registration for API keys to download content for text and data-mining is easy and fast. Once registered, API keys are available from the developers’ portal at http://dev.elsevier.com/index.html. Registration not only provides downloading capabilities, but access to Elsevier’s technical support.

Text and data to be mined is downloaded from Elsevier’s systems to local ones for further manipulation. Open-access content does not require an API to be text-mined; however, it is highly recommended as support is provided to registered users. Images can be text-mined as well with a different API. Instructions on sharing abilities with downloaded text and data-mined content is available at https://www.elsevier.com/about/our-business/policies/text-and-data-mining. Access to content for text-mining is available to everyone including researchers at non-subscribing institutions – just email: universalaccess@elsevier.com. I am not sure how many researchers are aware of this ability to consult with content providers to assess the potential for a partnership.

Content Providers’ Value

Content providers can add extra value to research through the expertise and knowledge they bring to the table regarding their content. They are aware of the differences between data that is recorded as part of journal publications (articles with accompanying data) and “raw data” or unpublished data. They also create the database platforms that structure the data; they know the strengths and weaknesses of the platforms. They are aware of idiosyncrasies and can navigate through the databases with finesse and familiarity. They can apply artificial intelligence and natural language processing to the content as well as other techniques. They also can offer guidance about approaches and methodologies for text and data-mining.

For researchers, gathering text from multiple content providers may or may not be desirable. Working in partnership with local company experts eliminates the need for researchers to become familiar with each content source and database infrastructure. Elsevier text and data-mined content can be combined with other publishers’ content; we recommend you use CrossRef’s API to do so.
Often data is the value proposition of a company, as it has invested time and resources into producing quality content, and therefore broad sharing of it may reduce the sustainability of the company’s mission. Content providers provide quality data to mine. Companies seeking revenue from their information analytic capabilities may not wish to share their data freely. This does not mean they are not willing to partner to apply their data to research. As more academia and industry are working together, the analysis of content seems to be a natural collaboration opportunity for both researchers and commercial content providers. The best way to determine companies’ interests in such partnerships is to inquire. Never assume content behind a paywall is not available for text-mining related to research.

Collective Solutions

It is exciting to see so much energy and creativity being applied by data stakeholders to derive standards and agreements on citing data. These efforts will enable interoperability between data repositories, creating tools for discovering data sets, and giving appropriate attributions for the intellectual property. It affords opportunities for the various data stakeholders to work together and to gain understandings of variant viewpoints and perspectives. It also gives some structure to an amazing web of content that is yet ineffectively managed and recorded for time.

As data is much more variable than journal articles or book formats, there is a lot of work left to do to figure out how to appropriately select and collect data, describe it effectively and authoritatively, store it for ease of access, retrieve it from distributed sources, analyze it with appropriate tools and resources, and report it for others to reuse, review and comment. There is also the discussion of how to attribute academic scholarly reward for its production and for its ability to be shared for future research. The FAIR principles of Findable, Accessible, Interoperable, and Re-usable, derived by Force11, are guiding efforts. More principles and standards for describing data are needed to optimize discoverability. Offering the complete picture of research outcomes through the linkage of journal articles with source data enables the re-use and cross-discipline review of results. Being able to search the data repositories to identify relevant data sets is another great tool being developed by several. We really are in the early stages of a “tower of Babel” however, the great news is we are aware of this and are together shaping the future of big data so it can be of benefit to health, to our societies, and to ourselves. Only through effective and frequent collaborations among the various data stakeholders can we derive solutions to meet our personal data management needs and those of our respective institutions.

Forum Participation Benefits

With deep appreciation, I am honored to be a participant of the IMLS National Forum on conducting data research mining using in-copyright and limited-access text datasets. I applaud the creators and planners of the Forum, and I look forward to learning more about the different perspectives participants will bring to the continuum of conversations. I have always believed we need to work across professional silos to shape the collective future of scholarly communications - including big data - as is evidenced from my work and leadership of the Chicago Collaborative (https://sites.google.com/a/umn.edu/chicago-collaborative/). Thank you for including me in this groundbreaking and beneficial learning experience.
<table>
<thead>
<tr>
<th>STRENGTHS</th>
<th>WEAKNESSES</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Elsevier has an existing policy supportive of researchers’ access to</td>
<td>• Content must be downloaded and stored locally to be text-mined, limiting central access and reuse</td>
</tr>
<tr>
<td>content for mining purposes.</td>
<td>by multiple researchers.</td>
</tr>
<tr>
<td>• Elsevier has committed to research data being made freely available</td>
<td>• Raw data is not centrally or easily available for review.</td>
</tr>
<tr>
<td>• Many resource platforms are created with the researcher in mind.</td>
<td>• Journals are just starting to require data accompany published articles.</td>
</tr>
<tr>
<td>• Elsevier has been very active in standard-setting initiatives, such as</td>
<td>• Standards for data are premature in development and not widely adopted.</td>
</tr>
<tr>
<td>Force11/data citation efforts.</td>
<td>• Knowledge about how to efficiently and effectively manage data is still being generated</td>
</tr>
<tr>
<td>• Elsevier has extensive experience working with institutions and</td>
<td>• Skills are just starting to be strongly developed among information professionals.</td>
</tr>
<tr>
<td>individual researchers on mining text and datasets.</td>
<td>• Library schools need to partner with researchers to understand the training needed for librarians</td>
</tr>
</tbody>
</table>
### Opportunities

- Collective initiatives on data standards need to continue to be supported and matured.
- Many stakeholders are willing to cooperate to formulate standards and offer data management guidance.
- The awareness for the need to corral data is growing among multiple types of professions and industries.
- Industry and university partnerships are more common, enabling innovative applications of data.
- Machine-learning and AI is on the cusp of enabling extensive analysis of vast amounts of data.
- Journal editors are starting to require data be available, if not connected, to journal publications.

### Threats

- Huge corporations, such as Google, will have an impact on how future data will be shared.
- IP and other commercial considerations.
- U.S. Net neutrality changes may impact the sharing of data easily among institutions of higher education and with industry.
- Selection of all kinds of data may create extensive noise that interferes with the collection of appropriate data.
- The silos that result from academic researchers’ rewards may inhibit the potential of data sharing and thus collective knowledge.
- Researchers’ and university administrators’ lack of awareness of the contributions librarians can add to the management of raw and published data.
A crucial milestone in the history of the scientific revolution was the 17th century recommendation of the Royal Academy that scientists publish their results in printed articles (https://royalsociety.org/about-us/history/). This public communication of science results enabled scientists to stand on the shoulders of their predecessors and greatly accelerated the progress of science. Of course, the technological artifact that enabled this process was the printing press, which had been invented (in Europe) in the 15th century.

A technological artifact of the 21st century may be as (or more) revolutionary for science than the printing press: the global network of digital computers. Whereas the printing press was the enabler of scientific communication, the global network of computers is both the enabler of new modes of scientific communication and of new modes of scientific performance. Some observers have said that digital computers have introduced two new modes of science beyond the traditional modes of theory and observation/experiment: computational science and data intensive science (https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/). Moreover, digital computers support the combining of science communication and performance in ways undreamed of in earlier times. It is as if the printing press and (say) the telescope were the same technological artifact.

This article will focus on some of the new modes of communicating science with only brief mention of the equally important new modes of scientific performance. The first part of the article will describe an “ideal” state of unencumbered use of the new technology and the second part will describe a practical application that can be utilized today and can serve as a bridge to less encumbered uses of the future.

Open science in the 17th century versus “ideal” open science today

Open science in the 17th century meant the printing of science articles. Today’s technology of networked computers enables scientists to publish all their results, not just summary articles. That is, in addition to articles, data, software, workflows, and other outputs can be “easily” published. That is, easily published when today’s technical, legal, and cultural barriers are overcome. This ideal world has recently been described by the policy goal of FAIR data (https://www.nature.com/articles/sdata201618). Data in this sense refers to all science products, including articles, which are the focus of this workshop. The acronym FAIR is decoded to Findable, Accessible, Interoperable, and Reusable. The next paragraph expands on these attributes.

Today articles are Findable primarily by keyword searches initiated by people, but the new technology will enable semantic searches initiated by both people and machines. Articles today are either freely Accessible or hidden behind paywalls based on copyright, government mandate, and other
considerations. In an ideal future, all articles will be freely available and the costs of publication will be supported by article processing charges or other mechanisms. Today, articles are Interoperable only in the sense that a scientist can read multiple articles and integrate the knowledge gained from each. In the future, the technology will enable computers to “read” the articles for humans in the sense of linking the knowledge from multiple articles. This capability is crucial due to the huge number of articles published—impossible for any scientist to read more than a small fraction. And this capability will enable a new mode of scientific discovery: “connecting the dots of science factoids” found in multiple articles. (An example of this is described below.) Finally, articles are usually Reusable today only in the sense of referencing them in other articles. In the future, Reusability will include the construction of derived works (with proper attribution). All of the futuristic uses depend on technological, legal, and cultural advances, which will (imho) occur in due course once their compelling advantages are widely understood. But “due course” is seldom quick, because the forces of the status quo always resist the revolutionary potential of new technologies (https://quod.lib.umich.edu/j/jahc/3310410.0002.217/-brian-winstons-media-technology-and-society-a-history?rgn=main;view=fulltext).

In the case of science articles, cultural and legal resistance is, of course, centered around the publishers’ holding the copyright to most science articles and their current business models of selling journal subscriptions to the scientific institutions where the science was conducted (http://www.arl.org/news/arl-news/3643-fair-use-in-text-and-data-mining-arl-publishes-issue-brief#.WrurBFopChA, https://sparcopen.org/our-work/developments-in-tdm-policy/). In the days of the printing press, this was a reasonable process, but today it is inefficient, resistant to the revolutionary capabilities, and prone to monopoly rents. As stated above, I believe that all this will change, but not tomorrow. Hence, the subject of this workshop about what to do in these times of In-copyright and limited access texts. The second part of this article will describe one approach to obtaining some of the revolutionary capabilities in today’s encumbered environment.

**Semantic processing of article abstracts**

Semantic processing of biomedical science articles has been accomplished at NIH/NLM over the last several decades. This experimental system, called *Semantic Medline* (1), is an existence proof of what could be accomplished for science articles in general, if the appropriate “text mining” research across the disciplines were conducted. An enabling feature of Semantic Medline in today’s copyright environment is that article abstracts are often made freely available even when the article is copyright protected. That is because reading the abstract may motivate a scientist to purchase the article. This section will describe Semantic Medline and its potential generalization to all science.

Semantic Medline consists of four subsystems: the Medline database of the titles and abstracts of (virtually) all biomedical research articles, a controlled vocabulary (to solve the synonym problem), a natural language processing system (nlp) to discover the “key sentences” in the abstract and extract the core predication of the sentence (of the form subject-verb-object), and a knowledge processing software system such as the Semantic Web. The following paragraphs will describe these systems.
Medline is one of the databases that NLM has been assembling for some time (https://www.nlm.nih.gov/bsd/pmresources.html). It currently consists of between 25 and 30 million article citations and abstracts. Other disciplines wishing to produce a similar system might need to wrangle citations and abstracts from multiple locations.

The controlled vocabulary system utilized by Semantic Medline was developed at NLM for other purposes, but served Semantic Medline well at “no additional cost.” It is a compendium of biomedical vocabularies known as Unified Medical Language System (https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/). Assembling such a vocabulary for each scientific discipline would be a major research undertaking. Some disciplines have begun this effort, but others have not.

The core activity of the Semantic Medline project has been to develop the nlp techniques to parse the abstract’s sentences and identity which are the key sentences—those that describe what the article claims to have accomplished. (The article itself describes how the results were achieved. Thus for finding relevant articles, the abstract may actually be better than the article itself.) The developers estimate that the average abstract has four key sentences and the current system can identify three of them. The set of all key sentence “cores” (of the form subject-verb-object) form a knowledge graph where the nouns are the nodes of the graph and the verbs label the arcs. An important research question is how easy or hard it would be to translate Semantic Medline’s nlp code to process the scientific language of other disciplines.

Finally, semantic processing of the Semantic Medline knowledge graph is accomplished by knowledgebase processing systems such as the Semantic Web (https://en.m.wikipedia.org/wiki/Semantic_Web). The Semantic Web has been under development for almost 20 years and should be ready for use in any discipline.

An example of the use of Semantic Medline will give an indication of its value (2). Two biomedical factoids have been known for some time. First, that older men have lower testosterone levels than young men, and second that older men have more sleep problems than their younger colleagues. The biomedical community occasionally “wondered” if the two facts were related, but no proof of a relationship was forthcoming. A very simple search in Semantic Medline discovered two existing articles that provided the connection. One article demonstrated that testosterone is an inhibitor of the hormone cortisol and the other demonstrated that cortisol is a sleep disrupter. Obviously no scientist had read both articles, probably because they were published in different sub disciplines.

In conclusion

This article has claimed that 21st century open science will revolutionize the conduct and the communication of science by publishing and processing all science products—articles, data, software, etc. In the current environment to copyright-restricted access to and processing of articles, some of the revolutionary benefits of digital processing can be had by processing abstracts, which may be open even
when the articles are not. An existence proof of the value of such processing is provided by Semantic Medline, a system developed at NLM to turn biomedical abstracts into a knowledge graph for semantic processing. Additional research is called for to implement a Semantic Medline work-alike for each scientific discipline.


2. Miller, Christopher M.; Thomas C. Rindflesch; Marcelo Fiszman; Dimitar Hristovski; Dongwook Shin; Graciela Rosemblat; Han Zhang; and Kingman P. Strohl. 2012. A closed literature-based discovery technique finds a mechanistic link between hypogonadism and diminished sleep quality in aging men. SLEEP 35(2):278-85.
FORUM STATEMENT

This memorandum summarizes the legal mechanisms through which text data mining (TDM) - enabling solutions may be implemented. It focuses exclusively on the process for implementation, not the substance of any remedy. There are five distinct approaches that may be taken, two non-governmental from the bottom-up and three that governments can take, individually or collectively, from the top-down. These include, in ascending order of effectiveness: (1) common-use licensing; (2) non-governmental agreements (or Memoranda of Understanding – MoUs); (3) national governmental legislation and regulation; (4) inter-governmental executive agreements or MoUs; and (5) governmental treaties.

There is an inherent paradox in these approaches. On the one hand, there is a range of effectiveness, beginning with individual, non-governmental action all the way to collective, intergovernmental approaches, as set out above. On the other hand, the more effective the mechanism, the more difficult it is to conclude and implement it. Normally, one would choose the most effective mechanism and ignore the others, but this is neither realistic nor advisable. Moreover, although each approach is distinct and has very different parties and legal effect, they also can be pursued concurrently, depending on a strategic perspective and the levels of maturity and receptiveness in each instance. The latter considerations involve more than just the nature of the actors and legal effect, but include scientific and technical (S&T) and socioeconomic factors that need to be weighed. In order for this outline of issues to be tractable in a short document, however, this memo does not take a strategic view and limits the outline to the tactical approaches with legal effect, only.

Also, at the outset are some approaches that are not included, but that have some legal relevance. These are: (1) individual contracts; (2) normative behavior; (3) the body of writings by legal and S&T experts; and, most important (4) other (non-copyright or IP) legal considerations and governmental mechanisms that are either already in place or may become significant.

Contracts, rather than licenses, are concluded with the specific assent of all parties, frequently just two parties. Licenses are “portable” and are good against the world. They are backed up by existing legislation and can be either more restrictive than copyright (think End User Licensing Agreements – EULAs – for commercial software or databases) or less restrictive, with “some rights reserved”, such as Creative Commons common-use licenses. Both contracts and licenses are legal instruments with legal effect. There are distinctions as well as overlaps between contracts and licenses, which is the topic of another article, but the key distinction here is that contacts are typically one-of-a-kind whereas licenses are usually standard. This means that the use of individual contracts (as opposed to large, multi-party agreements, which can be a contract as well) should not be encouraged for use in reducing TDM restrictions, since they are narrowly limited in scope to the parties giving assent.

Normative behavior does not have legal effect, per se, but is an important precursor to the development of policies and laws. If there is a groundswell in the application of TDM for research purposes, it is indicative of the need for more formal codification of such practices in the law. At the same time, although all research disciplines engage in TDM, some are much more prolific than others and some are more proprietary than others (think, for example, the STM literature or databases in astronomy vs.
industrial chemistry). The nuances exhibited by the different disciplines are not only instructive but can present models to emulate in an approach with more legal effect.

The publications of experts, whether in law or in other sectors, are also indicative and can propose solutions that may be adopted in more formal legal instruments. Both the normative practices of the research communities as well as expert analyses in the literature are important to survey in advance of setting up mechanisms that have the force of law.

Finally, there are many other laws and practices with legal significance to TDM, that need to be considered in developing a legal strategy. Although intellectual property laws can be waived or modulated by individuals or organizations through license or contract, there are other restrictive laws that cannot be so modulated, such as the protection of personal privacy, national security, endangered species, and the like. Moreover, there are the restrictive licenses of STM publishers and commercial vendors of databases or “collections of information” that form existing practice that either need to be changed or potentially emulated (if sufficiently enabling) as a model for future formal adoption. Although these other laws and practices with legal effect are important to finding solutions to the restrictive TDM problem, they too lie outside the scope of consideration for this memo but need to be considered at the workshop and beyond.

The rest of this memo, then, focuses on the five broad approaches and gives some preliminary suggestions about the use of each. The purpose is to provide a framework within which solutions may be developed.

**Common-use Licenses**

As noted above, these are licenses that promote open access and user rights, and derogate from full copyright with some rights reserved or not at all with a CC0 waiver. Common-use licenses are not mechanisms for ameliorating restrictive TDM laws that are already in place, however. They provide an *ex ante* approach that authors and institutions can adopt on a voluntary basis, rather than an *ex post* remedy to restrictive TDM. Such licenses are also blunt instruments, in the sense that some of them enable users to perform TDM of the literature so licensed, but they are silent on TDM practices and therefore are not nuanced or protective of other potential author concerns.

There are now many common-use licenses being used (over one billion Creative Commons licenses and many other types), so this is becoming indicative of TDM-enabling practice in research communities. Furthermore, there are over 11,000 open access journals self-registered in the Directory of Open Access Journals ([https://doaj.org](https://doaj.org)). This practice is now becoming more visible and may be among the approaches to consider.

**Non-governmental Agreements (MoUs)**

Many research institutions, universities, and other non-governmental organizations (NGOs) enter into various kinds of consortia and cooperative agreements for research and other purposes. These agreements are often legally non-binding, unless they are written as a binding contract, but both versions set out provisions that guide all participants, including the affected employees of those organizations.
There are two potential legal effects of non-governmental agreements, both as indicators or even models for governmental actions with more binding legal effect. The first is an extension of the individual common-use license approach. Research projects and consortia are increasingly using those licenses. Such non-governmental agreements also have the possible specific inclusion of a clause that specifically addresses TDM, although this observer is not aware of any.

The other approach is related to this latter possibility; that is, one can envision universities and scientific societies concluding an MoU with one or more major subscription journal publishers with the express purpose of working out an enabling TDM agreement. If successful and broadly applicable to other disciplines, it could form the basis for national legislation or even a protocol to the Digital Copyright Treaty of 1996.

**National Legislation and Regulation**

Legislation enacted by the national legislature (in the case of the US, the Congress) and implemented by one or more agencies of the executive branch through some regulations has the advantage of a law that is binding on all people within the jurisdiction. In the case of restrictive TDM, such legislation almost certainly would be in the form of an amendment to the existing copyright law (the 1977 Copyright Act in the US).

However, it is very difficult to bring an issue up for legislative action and usually even more difficult to negotiate a compromise among all the parties interested in the issue. In the case of TDM, many entities—corporate, academic, and governmental—on both sides of the issue would be involved. This would require some national track record with potential solutions. The law’s protagonists would be well advised to develop a legislative strategy and model language prior to pushing the topic for congressional action. Even with a careful plan and a receptive legislature, the process can be overtaken by parties with the deepest pockets and most experienced lobbyists.

**Treaty**

Although inter-governmental executive agreements can be considered as an intermediate step between national legislation and a treaty, we will consider the treaty next because it is similar to the national legislation process. The main difference is that it is even more difficult to come to an agreement on a treaty, even one that is just an amendment or protocol to an existing treaty, such as copyright. Several nations should already have passed a legislative fix to the TDM issue to serve as a model for the treaty. Moreover, a treaty is even more complicated than national legislation, with more parties, especially governments, at the table. Finally, the inter-governmental treaty will need to be implemented in national legislation and regulations, potentially a long process in and of itself. Thus, a treaty is the last option to consider, after most of the other approaches have been implemented, in multiple countries, sectors, and disciplines.

**Inter-governmental Executive Agreement (MoUs)**

In many ways, an inter-governmental agreement is similar to a non-governmental agreement (both are often called MoUs). An inter-governmental MoU is concluded among ministries or government agencies, rather than by entire national governments, so the effect is limited to the activities within those
institutions’ purview. An intergovernmental MoU also can legally binding or non-binding. Sometimes, a new intergovernmental research facility or organization can be formed, which also may be explored in the TDM context. Because this type of mechanism is more limited in scope, it usually does not trigger the intense lobbying and competition for control by outside parties, such as the national legislative or treaty-making processes do. At the same time, it has a narrower or more limited, legal effect.

**Conclusions**

The reader is encouraged to see the accompanying SWOT analysis for each of the five main approaches, for a list of the pros and cons of each.

With regard to common-use licenses, open-access literature and publishing has made great strides in the last 15 or so years and will likely continue to expand. It may well be that most research literature at some future point will be open access and TDM will not be restricted. This, however, may take a long time and is uncertain, so the other approaches should be pursued as possibly more near-term and certain solutions.

There are two tracks that have been laid out above, one is the “agreement” approach and the other is the “legislative.” Agreements or MoUs are much easier to control and conclude, and form the basis and models for the more formal legislative solutions. Major representatives and umbrella entities in academic research should be encouraged to form an agreement with subscription STM journal publishers, both as a precursor to the legislative fixes and also as serious solutions in their own right. If comprehensive enough, the agreement approaches may be all that is necessary, since a solution can be carved out from the legislative context.

The national legislative changes to copyright should not be initiated until significant agreements and potential models have been implemented and tested. A treaty fix would be the last approach, after the successful implementation of all other approaches. Neither legislative option should be attempted without the formation of a coalition that is guided by a strategic plan and under what are perceived to be optimal conditions. A corruption of the legislative process and the outcomes, beyond the control of many of the parties involved, may be expected in any case.
**STRENGTHS OF A TREATY (OR EXECUTIVE AGREEMENT)**

- An international or global approach, probably as a new protocol or annex to the Digital Copyright Treaty (1996), that has the broadest and strongest possible legal and policy effect.
- Enacted in national legislation as statutory law of the land, all having minimum standards based on the treaty.
- Promotes certainty and lasts a long time.
- Legally interoperable approach, if many of the major countries adopt it.
- If successfully concluded, a treaty theoretically should be the best approach yielding the most benefits.

Note: An intergovernmental executive agreement has many of the same SWOT features of a treaty, but is narrower.

**WEAKNESSES OF A TREATY (OR EXECUTIVE AGREEMENT)**

- Treaties, even a protocol or annex to an existing treaty, can take many years to negotiate and may even have an unsuccessful outcome.
- There is a political antipathy to treaty mechanisms and resulting implementing legislation in the US Senate.
- Databases in different disciplines have varying contexts, especially institutionally, but this approach is inflexible, with a one-size-fits-all solution.
- Any mistake(s) (i.e., wrong assumptions, effects) are propagated.
- Difficult to change.
- Lasts a long time in a fast-moving technological context.

**OPPORTUNITIES OF A TREATY (OR EXECUTIVE AGREEMENT)**

- Brings all major parties to the table to work out a comprehensive and lasting solution.
- Presumably under WIPO auspices, with resident experience and expertise.

Note: An executive agreement would not be negotiated in WIPO, but by the government ministries directly responsible for its implementation.

**THREATS OF A TREATY (OR EXECUTIVE AGREEMENT)**

- The process can (is likely to? will certainly?) be captured by the deepest pockets. This may be heightened in an executive agreement, which is not nearly as transparent as the treaty process.
- Because buy-in is difficult when the process is broader, one can end up with a lowest common denominator outcome that does not satisfy the concerns of all parties, who then have to live with the results. As a result, there likely will be winners and losers.
- New or innovative approaches can be stymied for a long time, if it’s broadly adopted. An executive agreement, however, is more flexible and easier to abrogate.
<table>
<thead>
<tr>
<th><strong>STRENGTHS OF NATIONAL LEGISLATION</strong></th>
<th><strong>WEAKNESSES OF NATIONAL LEGISLATION</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Same strengths as for a treaty regime, but with more attenuated effects, because it’s effective in only one country, rather than multilaterally.</td>
<td>• Same weaknesses as for a treaty regime, but with more attenuated effects, because it’s one country rather than a multilateral regime.</td>
</tr>
<tr>
<td>• More likely to be more flexible and responsive to the nation’s specific needs than a treaty, because it’s not entangled in multilateral considerations.</td>
<td>• Small countries have a harder time influencing other, bigger countries.</td>
</tr>
<tr>
<td>• Can be quicker than a treaty and easier to change.</td>
<td></td>
</tr>
<tr>
<td>• Big countries can benefit the most and develop standards for others to use.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>OPPORTUNITIES OF NATIONAL LEGISLATION</strong></th>
<th><strong>THREATS OF NATIONAL LEGISLATION</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Same opportunities as for a treaty regime, but with more attenuated effects, because it’s one country rather than a multilateral.</td>
<td>• Same threats as for the second and third bullets under a treaty regime, but with more attenuated effects, because it’s one country rather than a multilateral regime.</td>
</tr>
<tr>
<td>• In the US, the legislation presumably would be negotiated under the auspices of the Judiciary Committees in Congress.</td>
<td>• More exposed to party politics and leading personalities, which can be bad.</td>
</tr>
<tr>
<td>• More exposed to party politics and leading personalities, which can be good.</td>
<td></td>
</tr>
</tbody>
</table>
### Strengths of NGO/Academic MoUs
- Can tailor the agreement according to community (discipline or institution or research project) norms.
- Can be quite fast, faster than official government legislation or treaties, but slower than individual action.
- Builds common understanding (trust) and practice among the participants. Can be emulated by others, including government(s).
- Can be binding, but among the parties only.
- It’s a relatively inexpensive process, both in terms of travel, time, and expertise.

### Weaknesses of NGO/Academic MoUs
- The agreements are limited in scope and effect (geographically, by project or discipline, institutionally), only to the parties that sign on.
- Probably not fully enforceable, just normatively.

### Opportunities of NGO/Academic MoUs
- The consortium can be emulated by others, including government(s).
- It can be the quickest way to start and build accepted practice for all disciplines or sectors.

### Threats of NGO/Academic MoUs
- The more players that are involved, the more compromises may need to be made, resulting in some undesirable outcomes for some (the weakest parties).
<table>
<thead>
<tr>
<th>Strengths of Common-use Licenses</th>
<th>Weaknesses of Common-use Licenses</th>
</tr>
</thead>
<tbody>
<tr>
<td>• They are very flexible, reflecting the wishes of the individual.</td>
<td>• They are only valid for the information they cover.</td>
</tr>
<tr>
<td>• They are very fast to implement.</td>
<td>• They may not cover all eventualities or preferences.</td>
</tr>
<tr>
<td>• Standard, common-use licenses are by far the least expensive option.</td>
<td>• They may not be widely adopted.</td>
</tr>
<tr>
<td>• Already have about a 15-year track record.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Opportunities of Common-use Licenses</th>
<th>Threats of Common-use Licenses</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Practice based on such licenses can build up into more formal arrangements.</td>
<td>• Can be prohibited by institutional policies.</td>
</tr>
<tr>
<td></td>
<td>• May be ruled illegal in some jurisdiction(s), although this has not happened and is not likely to occur.</td>
</tr>
</tbody>
</table>
FORUM STATEMENT

As Executive Director of the California Digital Library (CDL), I identify with the library community. CDL’s touch-points with Text & Data Mining (TDM) are:

- We are a founding member of HathiTrust, and CDL has coordinated UC’s contribution of over 4 million volumes to this effort, which are a significant portion of the 16 million volumes in HathiTrust. HathiTrust content is made accessible for TDM through the Hathi Trust Research Center. I serve on the HathiTrust Board of Governors.
- We negotiate approximately $50 million of electronic content licenses on behalf of the UC system. Approximately 60 of our licenses from 30+ providers include provisions that allow for TDM, and we actively seek to add TDM provisions on an ongoing basis. We also publish a public model license which includes the following terms for TDM:

  “Text and Data Mining. Authorized Users may use the Licensed Materials to perform and engage in text and/or data mining activities for academic research, scholarship, and other educational purposes and may utilize and share the results of text and/or data mining in their scholarly work and make the results available for use by others, so long as the purpose is not to create a product for use by third parties that would substitute for the Licensed Materials. Licensor will, upon receipt of written request, cooperate with Licensee and Authorized Users as reasonably necessary in making the Licensed Materials available in a manner and form most useful to the Authorized User. Licensor shall provide to Licensee, upon request, copies of the Licensed Materials for text and data mining purposes without any extra fees.”

(The CRL LibLicense Model license includes similar terms.)

I believe the main question facing us could be encapsulated as follows: what investment are required to turn TDM into a standard first-class research activity, and TDM outputs into first-class research data products?

In terms of TDM as an activity, it becomes “first class” predominantly when researchers have the tools and knowledge to routinely use it as a technique to answer their research questions. Adoption of this new technique will be driven by compelling exemplars of excellent and break-through research produced utilizing these techniques. Among the prerequisites are:

- A compelling, comprehensive corpus containing content that can be clearly characterized
- The ability to constrain or subset that corpus for any given research question
- The ability to leverage the tools of choice to analyze the subset
In terms of TDM outputs, these outputs are not confined just to publications of conclusions (those are already first-class research products), but also encompass access to the underlying dataset on which these conclusions are based so they can be independently verified.

Some of the most prominent attributes of a first-class research data product are:

- It can be published.
- It can be cited.
- It persists.
- It can be reused.

Publishers in particular want to curtail the scope of output that can be shared and whom it can be shared with, fearing that these outputs might substitute for access to the source content itself, while to the researcher, sharing a broad range of outputs, including underlying datasets, may be an important aspect of the value of their work.

In summary, the distributed nature of the corpus for TDM (for starters, library digitized content and library licensed content); the scale of the overall corpus; and the legal framework (IP rights as well as contractual obligations) surrounding it pose significant challenges to making TDM a routine scholarly activity, and turning TDM outputs, including the datasets underlying publications, into a first-class research data product.
IMLS National Forum  
Data Mining Research Using In-copyright and Limited-access Text Datasets  
Prepared by: Günter Waibel, California Digital Library (CDL)

**STRENGTHS**

- Library mass-digitization projects resulting in large-scale aggregations such as HathiTrust create a meaningful corpus  
- Availability of technology that begins to mitigate use of rights-restricted textual content such as HTRC data capsule  
- Existence of model license that includes TDM provisions for licensed content, widely implemented at the University of California and beyond  
- First exemplars of TDM potential as a research tool

**WEAKNESSES**

- Bringing together a larger corpus of materials by necessity has to cut across digitized and licensed content  
- Intellectual challenges of defining the content of any given corpus, its bias (exclusions / overrepresentation) and limitations.  
- Legal and technical challenges in defining and fixing a corpus subset at a moment in time as the underlying dataset for an investigation  
- Legal and technical challenges in making corpus subsets available for verification and replication of results  
- The idiosyncratic and project-specific nature of TDM makes it difficult to scale approaches or build on existing work

**OPPORTUNITIES**

In the short-term  
- Leveraging the HTRC data capsule as a technical environment which could provide TDM access to digitized and licensed corpus content  
- Engaging publishers in a conversation about practical implementation of existing TDM clauses in licensed content contracts  
- Identifying and promoting inspiring use cases and relevant methods to highlight the research potential of TDM

In the mid- long-term  
- Direct Open Access to the scholarly literature obviates legal hurdles for the currently licensed literature

**THREATS**

- Lack of researcher uptake of TDM as a first-class research activity  
- Lack of publisher incentives to make licensed content available for TDM outside of their proprietary environment  
- Lack of library investment in TDM as a first-class research activity  
- Lack of consensus around how to mitigate legal and contractual risks through a technology and best-practice framework
FORUM STATEMENT

Collections of scholarly publications are not usually thought of as big data. However, they can contain material that is of great value to researchers, including text-based scholarship, information about authors, citations, equations, and figures, as well as large-scale related data and structures such as images, slides, and data sets. Increasingly, scholars in a variety of fields are asking for access to large collections of material in order to pursue new research questions and methodologies. As a result, publishers are receiving an increasing number of requests for access to their content from research libraries and end-users for this kind of scholarly data mining. While there has been a growing interest in this sort of service, many publishers do not see this as central to their business, and may feel that responding to these requests is a distraction from the work that they need to do to advance their publishing programs. In addition, it has become clear that in most cases, scholars want access to data from a large body of publications rather than just one publisher.

For these reasons, it will be useful to explore whether it is possible to develop a text and data mining service that can serve this emerging need of the community, while addressing the concerns and requirements of publishers, libraries and end-users. There are a number of questions that will need to be answered, including whether publishers would allow access to their content for this type of use, and whether the community would be willing and able to support such a service. In addition, we will need to understand the costs for technical development, the requirements for interfaces, tools, and user support, as well as what sort of cost-recovery model would sustain such a service over the long term. Any development process will require both a clear understanding of the growing need for text mining for scholarly research, and what is needed to ensure the security and responsible use of publishers’ copyrighted content.

In 2017 ITHAKA and the HathiTrust Research Center conducted some preliminary research on TDM with librarians, publishers, and scholars. Here are some of the key takeaways from this project:

- The current level of requests to perform TDM is low, although it is trending upward
- It will be difficult for many libraries and publishers to respond to an increase in demand for assistance with TDM requests with existing skills and resources
- A centralized TDM service should focus primarily on clearing rights and providing normalized content with high quality metadata, as many libraries are able to provide the tools and support needed by scholars at this time
- A centralized service should be able and willing to provide participating publishers with robust monitoring and usage reports
- Libraries’ willingness to support a centralized TDM service will depend on the cost of participation and the value of the service to their users
- There is a need for education across the scholarly communications ecosystem regarding the use of TDM and the available services in order to increase awareness and use
<table>
<thead>
<tr>
<th><strong>STRENGTHS</strong></th>
<th><strong>WEAKNESSES</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Perceived need from the researcher and library communities for a TDM service</td>
<td>• Possibility that there is not a large enough need for this sort of service to sustain it</td>
</tr>
<tr>
<td>• Potential ability to provide a large selection of published content for TDM use</td>
<td>• Question as to who would pay to support this service</td>
</tr>
<tr>
<td>• Potential to provide published scholarly content from a variety of disciplines in one place</td>
<td>• Possibility that publishers may not be willing to provide permission for TDM use</td>
</tr>
<tr>
<td>• Ability to leverage established relationships with scholarly publishers</td>
<td>• Need for funding to develop and launch the service</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>OPPORTUNITIES</strong></th>
<th><strong>THREATS</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Opportunity to help scholars undertake new forms of research and analysis</td>
<td>• Possibility that development costs will be too high to manage within existing budgets</td>
</tr>
<tr>
<td>• Ability to leverage work done reformatting of scholarly content</td>
<td>• Possibility that there is not currently enough need to justify costs of developing and then maintaining a TDM service</td>
</tr>
<tr>
<td>• Opportunity for useful collaboration among preservation services, archives, libraries, and publishers</td>
<td>• Possibility that there will not be sufficient collaboration, and the community will end up with a group of siloed TDM services</td>
</tr>
</tbody>
</table>
**FORUM STATEMENT**

**Who I am and what I do:** I am the Digital Humanities Librarian in the Stanford University Libraries, and as such I act as principal data provider (or broker) for the humanities community on campus, identifying, acquiring, and provisioning data on their behalf. I have a modest collections budget with which to purchase data, and often advise and collaborate with other subject librarians on data purchases in their areas. My advice is largely technical, e.g., to ensure that data formats and license terms are appropriate for scholars’ needs. I am not responsible for negotiation of licenses myself; that is the role of our Digital Library Program Officer. But he often consults me, particularly on TDM license terms.

Rather than thinking of myself as an expert in licensing, copyright, or fair use, I identify as a public-facing librarian who works closely with researchers, and as an active member of the DH community: I observe what DH researchers require to do their work, what does and does not work for them, and what our Library practices and policies regarding data provisioning are. In addition to observing and advising, I also advocate for generous terms of data access and for fair uses of data. I also interact with vendors fairly frequently, mainly to request delivery of raw data, but also to negotiate terms of delivery, data formats, etc.

**What sorts of limited-access datasets I work with:** There are two broad categories: first, licensed full-text data purchased from a handful of vendors, or acquired through sharing arrangements with other libraries or consortia; second, in-copyright texts owned and digitized by the Libraries (or contracted out for digitization, e.g., to Google.

The data types I encounter range from essentially error-free, rekeyed texts, to uncorrected OCR; these are also generally marked up in XML of various schemas, granularity, and utility. Most of the researchers I currently support tend to extract metadata from XML header elements, then strip out all markup from the text proper before engaging in data enhancements (e.g., part-of-speech tagging, named-entity recognition, algorithmic OCR correction) subsequent to text mining itself.

**What sorts of licenses I prefer:** Before text- and data-mining services were widely seen as a profit center for vendors, I relied heavily on what might be called an “implied license” for text mining: in the earliest days of library electronic text collections (late 1990s-early 2000s), we commonly licensed datasets with the explicit intention to take delivery of the raw data, then locally load and serve it to our community through our own search interface. When text mining became important to our researchers, I simply interpreted the terms of that license to mean that I could likewise provide raw data to authorized users to use in their own text-mining applications – which, after all, are themselves simply another sort of search interface. I didn’t ask permission or seek to renegotiate original licenses (nor did I seek the advice of counsel) to allow for text mining, because it seemed (and seems) immanently clear to me that this practice follows both the spirit and the letter of the original licenses.
I still prefer that approach to our current one (both imposed by vendors and recommended in the library licensing community) of explicitly including text- and data-mining provisions in the license. While seeking protective terms is of course a reasonable impulse for both parties, it strikes me that the very explicit and overly fine-grained terms of use now preferred by vendors get in the way at both ends of a data transaction: they are often the subject of disputes during license negotiations, and they tend to hinder, rather than enable, the wide variety of research practices in which scholars engage (not to mention locking in whatever the current TDM practices happen to be: such licenses are not at all future-proof). Furthermore, in my view, strict terms do absolutely nothing to protect the intellectual property of a vendor: accidental (or purposeful) data leaks should be corrected (or punished) without hesitation – but these have nothing to do with access or use restrictions on authorized users. It strikes me that the only truly essential license term for non-open data is one that prohibits distribution outside the authorized user community.

If there are to be TDM terms, I much prefer broad (or even implied) ones, granting as much direct and unrestricted researcher access to raw data as possible, and a strict policy of non-redistribution without permission: as long as licensed data stays within the authorized user community, it should be utterly unrestricted. Vendors don’t always understand or agree with this – but I strongly believe that they are overstepping when they attempt to restrict scholarly access and uses.

On the other end of the research cycle, the results of text-mining research must be open for verification and replication. I strongly advocate for public sharing of derivative data as a fair use: by the same principle that one can’t claim copyright for a phone book, we should argue that data about a text corpus (word counts and vectors, feature sets, n-grams, etc.) are facts, and thus not subject to copyright.

It is poor stewardship of collections budgets to spend large sums licensing data that is restricted in ways that might hinder research. I believe that the research library community should leverage its collective consumer power by walking away from overly restrictive text- and data-mining licenses.

**What I consider fair TDM use of in-copyright data:** Finally, on the matter of non-expressive uses of in-copyright (unlicensed) text corpora, I believe that our community should rely heavily on the decision in Authors Guild v. HathiTrust: that digitization for research purposes is a fair use. The best argumentation in favor of that decision (and the one most applicable to the research community) is the amicus brief of “Digital Humanities and Law Scholars” (Matthew Sag, et al.). I believe that this should be required reading for librarians, scholars, and students in the Digital Humanities – and that we should enshrine its principles in a community-created “best practices for fair use in DH” document.
<table>
<thead>
<tr>
<th><strong>Strengths</strong></th>
<th><strong>Weaknesses</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Many of the texts that researchers want to study are still under copyright, or have been digitized and provided by vendors who license the data. Its very availability is more than an affordance: it’s a requirement for many research questions.</td>
<td>• Many vendors or licensors either aren’t inclined to allow direct access to their data, or are inclined to monetize it excessively, requiring libraries to pay essentially double for content.</td>
</tr>
<tr>
<td>• From a vendor point of view, licensed data sets are a source of income and economic viability. (This may not seem like a primary strength to the research community, but by analogy with our shared desire to see a thriving publishing industry, it is.)</td>
<td>• Many vendors want to provide data mining services only on their own servers, e.g., via an API, motivated by a misplaced desire not to lose control over it. Many of our researchers are not able to use vendor-provided tools or platforms, and require more direct access.</td>
</tr>
<tr>
<td></td>
<td>• The perceived risk, or perceived illegality, of gaining “back-door” access to data. Some researchers are under the mistaken impression that copyright disallows them from doing what their research requires.</td>
</tr>
<tr>
<td></td>
<td>• The difficulty (or inability) to provide data access to others (e.g., reviewers or other scholars), for validation and replication of research results.</td>
</tr>
<tr>
<td><strong>Opportunities</strong></td>
<td><strong>Threats</strong></td>
</tr>
<tr>
<td>-------------------</td>
<td>------------</td>
</tr>
<tr>
<td>Promoting the idea that derivative datasets (e.g., word counts, vectors, n-grams, statistical measures) should be made public as a fair use.</td>
<td>The threat of civil copyright litigation is still very real, in spite of the favorable Google Books and HathiTrust decisions.</td>
</tr>
<tr>
<td>Libraries can become (or be seen as) ever more relevant in the era of text mining as being the best poised to negotiate licenses, to pay for licensed data, and to acquire, manage both in-copyright and licensed texts.</td>
<td>Criminal liability under the Digital Millenial Copyright Act’s anti-DRM-hacking provision.</td>
</tr>
<tr>
<td>Engaging (rather than fighting) the author community in promoting the responsible use of their works, and proclaiming the significance of TDM research methods.</td>
<td>Librarians are threatened with an overwhelming number of data requests requiring high-touch interactions.</td>
</tr>
<tr>
<td>Creating a set of community best practices for fair use, and licensed use, of textual data.</td>
<td></td>
</tr>
</tbody>
</table>