Library Support for Text and Data Mining

A Report for the University Libraries at Virginia Tech

by the Text and Data Mining Task Force
Philip Young (Chair)
Collin Brittle
Inga Haugen
Edward Lener
Virginia Pannabecker

June 22, 2017

This report is issued under a Creative Commons Attribution 4.0 International (CC BY) license

Suggested citation:
Young, P., Brittle, C., Haugen, I., Lener, E., Pannabecker, V. (2017). Library support for text and data mining: A report for the University Libraries at Virginia Tech. Retrieved from http://hdl.handle.net/10919/78466

Introduction

Text and data mining (TDM) uses methods of automated extraction, combination, and analysis of data to create new information by revealing trends, patterns, and relationships. While considered together in this report, the mining of text and of data usually require different considerations. Text mining, sometimes called text analytics, can be viewed as a subset of data mining.

Text and data mining is used extensively in fields such as bioinformatics, business, computer science, and the humanities, though it can be applied to any field. As the number of peer-reviewed articles and related data increase, TDM is often viewed as a way to make sense of this information. TDM can be used to identify research trends, reveal data relationships that were previously undetected, and to review the literature in a field. The use of TDM as a research technique is expected to increase and to expand into new disciplines. Academic libraries have traditionally been a major source for peer-reviewed and other scholarly literature on campuses, and therefore should play a role in facilitating access for TDM.

TDM is already used by several researchers at Virginia Tech, though usually without support from the Libraries. The purpose of this report is to explore how the University Libraries has supported TDM in the past and how it might do so in the future, with a focus on access to peer-reviewed research and other scholarly sources.

University Libraries Support for TDM

Past support for TDM by the University Libraries is not well documented. Faculty requests seem to have been relatively few in number and unique in their needs. Support from the Libraries has primarily come in the form of purchasing data and/or the rights to mine content. For example, the Libraries has purchased access to databases such as Wharton Research Data Services and Compustat, sometimes with cost-sharing arrangements. However, it is not clear whether purchased data is used for TDM or for other research purposes.

Selectors have used funds from their collections budget to purchase data by faculty request. This data sometimes comes in physical form (such as a CD, USB drive, or hard drive), due to vendor concerns about the security of proprietary data as well as problems involving the online transfer of very large datasets. These physical formats do not seem to be “checked out” to a library user. Lack of a catalog record or other record of provenance could be the result of a license or purchase agreement that names a researcher rather than the Libraries. Additionally, data may be time-sensitive and lose value over time. In any case, it appears that at least some data purchased by the Libraries is not retained.

On the other hand, some data resources retained by the Libraries have little if any information associated with them. For example, the Libraries currently possesses a hard drive containing archives of *The New York Times*, and has transferred 3 terabytes of 17th and 18th century images from hard drives to a non-accessible server. The task force could not determine the origin or users of these resources.

One of the best documented TDM projects involving the Libraries is described in *An Epidemiology of Information: Data Mining the 1918 Influenza Epidemic Project Report*. This project involved a library faculty member, Bruce Pencek, as a co-principal investigator. Problems of access to newspaper content are described on p. 44 of the report:

> Unfortunately, it was not possible to utilize the proprietary newspaper databases for text mining, although they were used in conventional, manual analyses that enriched the grant researchers' knowledge bases. Three related reasons prevented our utilization of the library's historical newspapers: cost, novelty, and contracts. Additional fees for actually acquiring the content for computational analysis were required, even though these were not well spelled out in the library's "perpetual access" licenses. For one vendor, for example, it was not technically possible to extract only the years required for this project from its multiple papers. The additional costs were unanticipated in the initial project budget. In addition, questions about the size of the data transfer and storage concerns complicated acquisition. Researchers, library informaticists, and vendor representatives felt their way across unfamiliar terrain for four months, until there simply was not enough time left to use the proprietary content. Ultimately, it became clear in conversations with the database vendors that text mining of historical periodicals was a use of their content not anticipated in their contracts or their practices.

Inclusion of TDM language in the licenses agreed to by the Libraries and by the Virtual Library of Virginia (VIVA) could address the permissions and cost problems of similar projects. Also, the contract language would serve as notice to vendors of this potential usage, though additional communication with vendors would assist in their readiness for TDM projects.

In 2008, Tamara Kennelly worked with Steven D. Sheetz, a professor in the Pamplin College of Business, and his students on a data mining project that involved exploring words used in about 200 poems received by Virginia Tech after April 16, 2007 (now part of MS2008-020). Poems were digitized and the analysis was enabled through optical character recognition (OCR). The Libraries worked with University Counsel to address copyright and confidentiality issues. No report of the findings was published.

In 2015, three librarians (Inga Haugen, Andi Ogier, and Ginny Pannabecker) worked with a research team led by Kathryn Ziemer (Bioinformatics Institute) investigating how much attitude research crosses disciplines such as sociology, anthropology, psychology, and computer science. They were interested in how different disciplines label and define 'attitude research'

(the term used in psychology), which refers to research on attitudes, sentiments, opinions, and beliefs, and their effects on human behavior. As an example, computer science researchers also explore topics that overlap with this area, such as “sentiment analysis” - the computational analysis and categorization of opinions, beliefs, attitudes in text based on natural language processing; and “human-computer interaction” studies that include looking at how interactions with machines, software applications, etc. affect or are affected by human behavior, and by human attitudes, sentiments, opinions, beliefs, past experiences, etc. To address this research goal, they conducted a citation network analysis by identifying the most influential articles on attitude research (or similar) from a given discipline (via expert recommendations, citation counts, and other methods), and then used snowballing methods to document all the citing and cited articles connected to those foundational articles from each discipline. Included in this process was a goal to pull the citation results using the API of Web of Science to make the initial data collection easier. Following this, the team investigated tools and methods for identifying and visualizing overlapping and connecting citations in the network within the same discipline and between disciplines. Inga and Ginny worked with the team on developing the overall method, especially for identifying experts in the field, and Andi Ogier, in collaboration with Annette Bailey, worked on negotiating API access to Web of Science for the team. Network exploration tools (including visualization of the data) considered during the project included Gephi (permalink) and Sci2 Tool (permalink).

Support for TDM in Academic Libraries

Academic libraries provide support for TDM in several ways, including creating educational materials (such as library guides and websites describing TDM), highlighting key tools (software, APIs, techniques) for accomplishing TDM, and recommending major content sources for TDM. Academic libraries also provide institutionally specific information about TDM provisions (or prohibitions) included in library resource licenses, as well as offering contact points for researchers to inquire about obtaining access beyond what is already allowed. Intellectual property laws and issues related to TDM are also prominent in academic library guides and website information. Within this realm, libraries often address how copyright law (and the fair use exemption) affect current access to and allowable use of content for TDM applications.

Academic libraries and consortia are including TDM provisions in licenses. While some libraries are also actively purchasing subscriptions to content or full datasets that can be used for TDM, many are taking a ‘by request’ approach that requires creating user awareness and providing a method to suggest dataset purchases or access subscriptions for their projects.

Links to selected academic library guides supporting TDM can be found in Appendix B.

Licensing for TDM

Content providers vary widely in their approaches to allowing TDM. Some vendors offer this capability at no additional cost while others charge extra, with fees commonly in the range of $500 to $1,000. While TDM is becoming more commonplace, many vendors still do not offer it as part of their service package. New licenses or license addendums are often required to initiate TDM, so incorporating discussions about TDM provisions when initiating new products or when renegotiating existing licenses can minimize the additional time required. VIVA, the statewide consortium to which Virginia Tech belongs, has recently been successful in incorporating TDM language into some of its licenses for new products, although it is not yet considered an essential element for approval or renewal.

There are several examples of license stipulations to permit TDM, such as those from the California Digital Library (Word document, section IV, p. 3; permalink), the Canadian Research Knowledge Network Model License (permalink), and the Jisc Model Licenses (permalink). Perhaps the best known is the 2014 Liblicense Model License agreement (section 3.2j; permalink), which allows for use of licensed materials by authorized users "to perform and engage in text and/or data mining activities for academic research, scholarship, and other educational purposes, utilize and share the results of text and/or data mining in their scholarly work, and make the results available for use by others, so long as the purpose is not to create a product for use by third parties that would substitute for the Licensed Materials." It is important to note the emphasis here on research and scholarly use of data. Commercial providers have legitimate concerns about broader sharing of data extracted from their products in such a way that might undermine their own offerings. Analysis, synthesis, and transformation of raw data by academic researchers can help avoid these concerns.

There is no specific exemption for TDM in U.S. copyright law, although TDM can be accommodated under fair use (see below). Most commercial products require a license which governs exactly who may use a resource (typically known as "authorized users") and what they may do with it. Therefore, licensing is a critical element in efforts to expand access to TDM rights for researchers at Virginia Tech. Elsevier's text and data mining policy (permalink) is part of new or renewing institutional licenses and allows affiliated researchers to perform TDM after they also agree to the terms. However, the policy has numerous limitations: it requires use of the API, provides access to text only (a separate agreement is required for images), limits re-publishing to text snippets not exceeding 200 characters, and requires outputs to prohibit commercial use through the CC BY-NC license. Wiley's text and data mining agreement (permalink) allows TDM by any researcher for non-commercial use, though library licensing terms supersede it for affiliated researchers.

One problem with TDM for scholarly resources is that researchers may need licenses from multiple vendors in order to gather items of a particular type or from a particular discipline. This

is an extremely time consuming process, which also involves varying licenses and technical accommodations. Some services centralize vendor permissions, such as Crossref (permalink) and the Copyright Clearance Center (permalink). However, it is not clear whether the Libraries could have a role in facilitating cross-vendor TDM, other than alerting researchers to those options. Additionally, many vendors prefer to provide TDM licensing to specific researchers rather than an institution, thereby limiting permissions to a specific project. In this case, the absence of continued and/or simultaneous research availability reduces the value of TDM resources purchased by the Libraries.

Legal Aspects of TDM

In addition to the problem of access to content, legal aspects of TDM for copyrighted content must be addressed (please note that the authors are not attorneys, and this report does not constitute legal advice). Since copyright law is territorial, this analysis focuses on TDM that is performed in the United States. Under U.S. copyright law, TDM can be accommodated under fair use (U.S. Copyright Law 17 U.S. Code § 107; permalink). Two countries, the United Kingdom and Japan, have specific copyright exceptions for TDM.

The Association of Research Libraries (ARL) issue brief Text and Data Mining and Fair Use in the United States (permalink) states that TDM is almost always a fair use, as long as a researcher is not bound by a contract that forfeits fair use rights, and does not "make the full text, or substantial portions, of the underlying articles publicly available." The issue brief cites eight cases in which courts ruled favorably for TDM as fair use, and analyzes the four fair use factors from a TDM perspective. Courts generally find TDM a highly transformative fair use that does not serve as a market substitute. However, this does not mean that fair use provides a comprehensive exception, and the four factor test under fair use should be applied to any TDM project utilizing copyrighted material.

Libraries should ensure that their researchers' fair use rights are not limited, as outlined in the Code of Best Practices in Fair Use for Academic and Research Libraries (permalink):

> Under some circumstances, fair use rights can be overridden by contractual restrictions. Thus, these principles may not apply if a library has agreed, in a license agreement, donor agreement, or other contract, to forgo the exercise of fair use with respect to some set of collection materials**.** If fair use rights are to be preserved, library personnel in charge of acquisitions and procurement should be vigilant as they negotiate and enter into contracts related to collections materials. (p. 4)

Libraries should also be aware of disciplinary norms regarding fair use, which are increasingly becoming documented in codes of best practice (permalink). Contract language that limits use

of corpora within interdisciplinary frameworks should be avoided, particularly if one of the disciplines involved has a much broader range of fair use practices.

The type of content being mined should be considered in respect to its copyright status. Text mining is generally covered by fair use. Numerical data, to the extent that it represents facts about the world, is not copyrightable, though database rights may be in effect in the U.S. for creative or original arrangements of data (in some other countries, database rights are stronger). Images, like text, should be presumed to be under copyright unless indicated otherwise. Reproduction of individual images or sections of text in research papers may require permission unless a separate fair use case can be made.

The role of copyright as a barrier or limitation to TDM research should lead academic libraries to continue advocating for the role of open licensing and the public domain. For example, a best practice for licensing data for reuse is to place it in the public domain with the Creative Commons Zero (CC0) dedication (permalink) or assign the Open Database License (permalink). Even when data might be considered facts, and therefore not copyrightable, these designations create greater clarity for reuse. A variety of open licenses are already available to those submitting material to VTechWorks and VTechData.

The task force has suggested questions to support a research interview with someone interested in TDM (see Appendix A), some of which address the legal aspects of working with the content. Though not comprehensive, these questions provide a starting point. For example, researchers may need to know whether TDM results can be accessible according to funder and/or journal data policies, and how much of the original resource can be displayed in a journal article. A library professional can guide researchers to openly available corpora for TDM during the research interview, which can also be found in the TDM online guide.

The Hague Declaration (permalink), endorsed by the ARL and many other organizations, is an international call for clear rights and improved infrastructure for content mining that

> ... aims to foster agreement about how to best enable access to facts, data and ideas for knowledge discovery in the Digital Age. By removing barriers to accessing and analysing the wealth of data produced by society, we can find answers to great challenges such as climate change, depleting natural resources and globalisation.

Technical Aspects of TDM

TDM impacts multiple steps in the research data lifecycle, and may be implemented in many ways. Downloading data is not necessary for TDM, but is common. The raw data may be created during research, or may be transferred from a colleague without reaching a network. Legal aspects aside, downloading raises other considerations. For instance, when scraping a

site to retrieve data, one should pause between fetches to prevent the site from experiencing denial of service-level loads. Where available, a site's official API, or bulk download service should be used to retrieve the data. However, use of an API sometimes requires agreeing to a license which limits the intended research (such as the Elsevier license mentioned previously).

Processing may be necessary to begin analysis. For instance, the corpus may exist as a PDF, which will likely require text extraction for use. Text may be stored in an encoding the TDM method cannot understand. Tabular data might require splitting, joining, or creating columns. Common office software, like Microsoft Excel and Adobe Acrobat, are usually sufficient for these tasks, but custom or niche software may be required. The desired corpus and TDM analysis software will determine what processing is required, and should be a major consideration while planning research.

By definition, TDM refers to the analysis step. In text mining, common techniques include sentiment analysis (determining a text's attitude towards a topic), topic modeling (identifying a text's topic), and term frequency-inverse document frequency (scoring a word's importance based on frequency across documents in a corpus). Depending on the data, clustering or numeric analysis methods may be appropriate as well. Commercial tools may already exist to perform the desired technique, but custom code is common. The programming languages Python and R both provide numerous libraries for users to perform their own analysis.

Due to continuing advances, this section provides only an introduction to material that will be kept updated in the Tools and Technology section of the TDM online guide.

Discussion

There are numerous challenges for researchers employing text and data mining. The Libraries could mitigate some but not all of these challenges. Many of the challenges are related to proprietary text or data for which TDM services are viewed as an additional revenue stream. Due to this, it is very important that researchers plan their projects well in advance so that costs or other challenges don't become barriers. In addition, outreach to researchers should warn that unauthorized crawling or scraping of Libraries-licensed resources could result in loss of access to them by the provider, and potentially to other university researchers if accessed by proxy (wireless or off-campus use). Cost-sharing could be considered for some projects when the Libraries has an existing vendor relationship that does not include a TDM agreement. For grant-funded projects, the Libraries might expect researchers to include all costs, particularly since many licenses fall between $500-$1000. Additionally, the Libraries might encourage the development of data management plans even when they are not required, in order to ensure planning for data storage, access, and reproducibility.

In some cases, vendors are not aware of the potential of TDM research, or have insufficient technology or expertise to support it. The inclusion of TDM language in licenses agreed to by the Libraries could spur better vendor preparation for TDM. Researchers should be aware of the varying quality of OCR for digitized resources. In some cases, the quality of vendor-supplied OCR may be superior to resources freely available, but will come at a cost. Researchers should also be aware that while a larger corpus generally produces more robust results for TDM, it also presents problems for data transfer and storage. File formats can also present difficulties for TDM. While many scholarly journals now provide articles in XML, some journals only provide articles in PDF, which can be a challenge for TDM tools. This is also a consideration for text and data mining in VTechWorks, where the PDF format is prevalent.

The Libraries has several opportunities to facilitate TDM for Virginia Tech researchers. Foremost among these is improving communication with vendors and reducing the transaction costs of researchers. The Libraries could promote TDM as a research tool, particularly in disciplines that are less aware of its potential, and the possible sources that could be utilized. In some cases there is a lack of familiarity with TDM methods, which might be more frequently used if they were better known. Libraries TDM resources may serve as an introduction to the field for many researchers, particularly students. By defining and connecting TDM-related services, the Libraries can help researchers become more aware of the considerations needed for their research. By including TDM in vendor licenses and by continuing to fund open access articles and resources, the Libraries can ensure wider and ongoing access to TDM sources, as opposed to project-specific use by a single researcher or research group. Resources such as Twitter and Wikipedia are text mined often because they are well known and available. The Libraries could create awareness of lesser known TDM sources as well as increase its contributions to corpora available for permission-free TDM.

The Libraries should weigh the limited requests for TDM support against the resources needed to build and expand that support. Indications are that TDM research methods will increase, and their frequently interdisciplinary nature favors a central resource like the Libraries. Increased TDM support would deepen the contribution of the Libraries to the research process and the entire research lifecycle.

Recommendations

The Text and Data Mining Task Force recommends that the University Libraries:

- Develop **TDM outreach** to inform researchers at Virginia Tech about support from the Libraries. One purpose of the outreach should be to create awareness of support and potential issues and/or costs before grant proposals are submitted. Outreach should identify available TDM sources and their terms of

use. The Libraries could also describe what kinds of research are possible for TDM sources.
    - Online outreach: develop an online guide
    - Training: develop short introductions/workshops (such as NLI sessions)
    - Liaisons: communicate how the Libraries can assist with TDM, and use the suggested TDM reference interview questions (Appendix A) to help interested researchers

- Develop **TDM expertise** within the Libraries in areas such as licensing, tools, methods, storage, preservation, and visualization. Since expertise will probably be distributed, a decision tree or contact list could be developed. Referrals and relationships could be made with other forms of campus support such as Advanced Research Computing's visualization services (permalink), as well as between experienced and inexperienced TDM researchers. Demonstrated expertise in the Libraries could result in co-investigator status on TDM research projects. Examples such as systematic reviews would combine a librarian's knowledge of scholarly literature with TDM techniques in a process that would benefit any discipline.

- Include **TDM in content licenses** to the greatest extent possible, and ensure that licenses do not limit fair use. Both the University Libraries and VIVA should insert the following language from the LibLicense Model License (section 3.2j; permalink) into new and renewing vendor contracts:
    - "Text and Data Mining. Authorized Users may use the Licensed Materials to perform and engage in text and/or data mining activities for academic research, scholarship, and other educational purposes, utilize and share the results of text and/or data mining in their scholarly work, and make the results available for use by others, so long as the purpose is not to create a product for use by third parties that would substitute for the Licensed Materials. Licensor will cooperate with Licensee and Authorized Users as reasonably necessary in making the Licensed Materials available in a manner and form most useful to the Authorized User. If Licensee or Authorized Users request the Licensor to deliver or otherwise prepare copies of the Licensed Materials for text and data mining purposes, any fees charged by Licensor shall be solely for preparing and delivering such copies on a time and materials basis."

  Within VIVA, the Libraries may need to advocate for the benefits of TDM, particularly when additional costs are involved.

    - If the use of TDM continues to increase, the Libraries might emphasize institutional licensing rather than licensing to a specific researcher. This would allow for continued and/or simultaneous availability beyond a specific project.

    - Where access to TDM resources has a cost, the Libraries could develop guidelines for cost sharing with grantees, colleges, and/or departments.
    - Identify which resources we already provide that allow for a granular level of search (such as searching the full text of all documents in the resource, e.g., ProQuest interface). This could help us prioritize which licenses to look at for access that facilitates TDM practices. These could then be added to the online guide as identified.

- Develop one or more **TDM resource centers** within the Libraries that have tools installed on computers, and that are capable of hosting events and consultations. Potential locations include Port (which already has some TDM tools like RapidMiner) and/or the new D-Hub.

- Continue to **support open access** through funding and infrastructure, which are often overlooked contributions to the research corpus available for TDM. Through the Open Access Subvention Fund, the Libraries help disseminate openly licensed research that can be reused in many ways, including for TDM. Subsidizing open resources such as Open Library of Humanities, Knowledge Unlatched, and Reveal Digital's Independent Voices project also enable TDM.
    - Develop explicit policies and technical accommodations for TDM in VTechWorks, VTechData, and Odyssey. For VTechWorks, consider facilitating TDM by providing for a bulk download, API, and/or a robots.txt crawl directive, as well as providing article access in TDM-friendly formats such as XML. While most items in VTechWorks are not openly licensed, almost all are available to the public, and the transformative nature of TDM provides a strong fair use case under U.S. copyright law.
    - Continue or expand digitization efforts, which add new material to be text mined through machine-readable OCR or transcription. Campus researchers might be alerted to these new research opportunities, and TDM potential can be a factor in prioritizing digitization projects.

- **Record metadata for TDM resources** licensed or owned by the Libraries. Provenance would allow the Libraries to track the location and history of resources, both physical and digital, and respond to audits. Metadata would allow for discovery of datasets, tools, and other resources for TDM, along with their licensing information. For data held in a physical format, determine a centralized location for storage and whether items can be checked out to users. For data held in a digital format, develop guidelines for data transfer and storage, and the potential need to limit access to specific users.

- Explore **library use of TDM**. Examples from the literature indicate that TDM tools could be used to gain more value from internal library data, to improve discovery systems, and/or connect library resources to the larger web.

Acknowledgments

The task force wishes to thank Annette Bailey, Ladd Brown, Shane Coleman, Tamara Kennelly, Mike Linkous, and Bruce Pencek for providing information for this report. The task force takes primary responsibility for the content of the report.

Appendices

Appendix A: TDM Interview Questions

- What research is planned? What is your research question or interest?
- Have you identified a source or sources of content that may address your interest or answer your question?
- What specific information would you like to extract from this source and what format would you like it to be in (to facilitate your analysis, etc.) if you have a choice?
- Does the source provide a method for pulling / exporting and saving the content you desire (API or download), and / or does it allow you to pull out the content you want and interact with it / manipulate it online?
    - Where available, a site's official API, or bulk download service should be used to retrieve the data
    - If not, the Libraries may be able to provide support in identifying a method or negotiating access; and failing both of those, identifying alternative content sources.
- Is permission needed and/or a fee charged to pull content from this source?
    - Note that permission for some types of content access is often requested even for openly available/free resources
- Has the content provider supported TDM projects before, or made preparations for doing so?
- How will the file type affect TDM?
- How large is the data? How will it be transferred? Where will it be stored?
- If the data is digitized text, what is the quality of the OCR? Is it possible to get a sample of the OCR to check for quality?
- How might the research benefit the content owner/provider?
- For copyrighted/proprietary content, how much can be displayed in the resulting research?
- Results of your TDM project:
    - To what extent will replication of the research be an issue?
    - Will researchers be able to deposit the data with their research, or otherwise provide reviewers/readers access?

- What are the data policies of the funder(s), if applicable, or the journal(s) in which the researchers hope to publish?
- What metadata will be needed to describe the data, such as provenance?

Appendix B: Selected Academic Library Guides to TDM

Boston College
Carnegie Mellon University
Duke University
East Carolina University
Florida State University
Indiana University Bloomington (archived version, 2013)
Macquarie University
Massachusetts Institute of Technology (list of APIs)
NIH Library
University of California, San Diego
University of Chicago
University of Massachusetts, Amherst
University of Michigan
University of Minnesota, Duluth
University of Southern California
University of Tennessee
Western Michigan University
Yale University, Medical Library

Selected Resources

Ananiadou, S., Rea, B., Okazaki, N., Procter, R., & Thomas, J. (2009). Supporting systematic reviews using text mining. *Social Science Computer Review* 27(4), 509-523. https://doi.org/10.1177/0894439309332293

Association of Research Libraries. (2015). Issue brief: Text and data mining and fair use in the United States. Retrieved from http://www.arl.org/storage/documents/TDM-5JUNE2015.pdf, archived at https://perma.cc/M9YJ-4CH5.

Clark. J. (2013). *Text mining and scholarly publishing*. Publishing Research Consortium. Retrieved from http://publishingresearchconsortium.com/index.php/prc-documents/prc-guides-1/4-prc-text-mining-and-scholarly-publishin-feb-2013/file, archived at https://perma.cc/47VD-EPNB.

Emery, J. (2008). Working in a text mine; Is access about to go down? *Journal of Electronic Resources Librarianship 20*(3). https://doi.org/10.1080/19411260802412745

Ewing, E.T., Hausman, B.L., Pencek, B., Ramakrishnan, N., Gad, S., Seref, M., Kerr, K., & Eysenbach, G. (2014). *An Epidemiology of Information: Data Mining the 1918 Influenza Epidemic Project Report*. http://hdl.handle.net/10919/46991

Green, H., & Dickson, E. (2017). Expanding the librarian's tech toolbox: The "Digging Deeper, Reaching Further: Librarians Empowering Users to Mine the HathiTrust Digital Library" project. *D-Lib Magazine*, *23*(5/6). https://doi.org/10.1045/may2017-green

Grewal, P., & Huhn, K. (2016). Text & data mining clauses in academic library licenses: A case study. Retrieved from http://library.concordia.ca/about/staff/forum/files/Presentation_Grewal_Huhn.pdf, archived at https://perma.cc/9HNM-BH2V.

Harris, L.E. (2016). Text and data mining (TDM): Copyright and licensing issues. Special Libraries Association Conference, Philadelphia, June 12. Retrieved from http://www.copyrightlaws.com/wp-content/uploads/2016/07/TDM-Slides-v3-12-June-2016-version-for-posting-PDF.pdf, archived at https://perma.cc/E3SX-PRB2.

International Federation of Library Associations and Institutions. (2013). *IFLA statement on text and data mining*. Retrieved from https://www.ifla.org/publications/node/8225, archived at https://perma.cc/VM3Y-GCVS.

Jeffery, K., Houk, K., Nielsen, J., & Wong-Welch, J. (2017). Digging in the mines: Mining course syllabi in search of the library. *Evidence Based Library and Information Practice, 12*(1), 72-84. https://doi.org/10.18438/B8GP81

Jisc. (2012). *The Value and Benefit of Text Mining to UK Further and Higher Education*. Digital Infrastructure. Retrieved from https://www.jisc.ac.uk/sites/default/files/value-text-mining.pdf, archived at https://perma.cc/LSV6-JZSC.

Joorabchi, A., & Mahdi, A.E. (2014). Towards linking libraries and Wikipedia: automatic subject indexing of library records with Wikipedia concepts. *Journal of Information Science 40*(2), 211-221. https://doi.org/10.1177/0165551513514932

Kaufman, R., & Iarrobino, M. (2016, August 11). Structured XML versus the blob of text. *Research Information*. Retrieved from https://www.researchinformation.info/news/analysis-opinion/structured-xml-versus-blob-text, archived at https://perma.cc/7DPC-UBS4.

Kilicoglu, H. (2017). Biomedical text mining for research rigor and integrity: Tasks, challenges, directions. *bioRxiv* [preprint]. https://doi.org/10.1101/108480

King, L. (2015). Data mining on vendor-digitized collections. *Library Connect*. Retrieved from https://libraryconnect.elsevier.com/articles/data-mining-vendor-digitized-collections, archived at https://perma.cc/DS55-LUW4.

Knoth, P. & Pontika, N. (2016). How can repositories support the text-mining of their content and why? OpenMinTeD workshop, Open Repositories conference. Retrieved from https://www.slideshare.net/openminted_eu/how-can-repositories-support-the-text-mining-of-their-content-and-why

Lammey, R. (2015). CrossRef text and data mining services. *Insights 28*(2), 62–68. https://doi.org/10.1629/uksg.233

Lease Morgan, E. (2012). Use and understand: The inclusion of services against texts in library catalogs and "discovery systems." *Library Hi Tech*, *30*(1), 35–59. https://doi.org/10.1108/07378831211213201

Lee, T. (2012). Text mining from three perspectives: An e-resource librarian's view. Charleston Conference on Acquisitions and Collection Development, November 2012. Retrieved from https://www.dropbox.com/s/k0r8qnvt95ifiyi/Text_Mining_TLee.pdf

Nicholson, S. (2003). The bibliomining process: Data warehousing and data mining for library decision making. *Information Technology and Libraries, 22*(4), 146-151. http://hdl.handle.net/10150/106392

Nicholson, S. & Stanton, J. (2003). Gaining strategic advantage through bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries. In Nemati, H. & Barko, C. (Eds.), *Organizational data mining: Leveraging enterprise data resources for optimal performance* (pp. 247-262). Hershey, PA: Idea Group Publishing. Retrieved from http://hdl.handle.net/10150/106383

NISO/ICSTI. (2016, June 30). Text and data mining: The way forward as seen by the library, publisher and researcher communities. [Webinar]. http://www.niso.org/news/events/2016/webinars/june30_joint_webinars/, archived at https://perma.cc/RH3P-JER9.

Okerson, A. (2013). Text & data mining: A librarian overview. IFLA WLIC 2013, Singapore. http://library.ifla.org/252/1/165-okerson-en.pdf, archived at https://perma.cc/ZW6C-UGP8.

Przybyła, P., Shardlow, M., Aubin, S., Bossy, R., de Castilho, R. E., Piperidis, S., ... & Ananiadou, S. (2016). Text mining resources for the life sciences. *Database*, *2016*, baw145. https://doi.org/10.1093/database/baw145

Publishing Research Consortium. (2016). *Text mining of journal literature 2016: Insights from researchers worldwide*. Retrieved from http://publishingresearchconsortium.com/index.php/prc-documents/prc-research-projects/54-prc-text-mining-of-journal-literature-2016/file, archived at https://perma.cc/43U6-LERA.

Rathemacher, A. J. (2013). Developing issues in licensing: Text mining, MOOCs, and more. *Serials Review*, *39*(3), 205-210. https://doi.org/10.1016/j.serrev.2013.07.016

Rebholz-Schuhmann, Dietrich, Anike Oellrich, & Robert Hoehndorf (2012). Text-mining solutions for biomedical research: enabling integrative biology. *Nature Reviews Genetics 13*, 829-839. https://doi.org/10.1038/nrg3337

SPARC. (2014). Developments in publishers' text and data mining (TDM) policy. Retrieved from http://sparcopen.org/our-work/developments-in-tdm-policy/, archived at https://perma.cc/3TAT-3ZBA.

Tsueng, G., Nanis, S. M., Fouquier, J., Good, B. M., & Su, A. I. (2016). Citizen science for mining the biomedical literature. *Citizen Science: Theory and Practice 1*(2): 14, pp. 1–11. https://doi.org/10.5334/cstp.56

Van Noorden, R. (2012, March 8). Trouble at the text mine. *Nature News & Comment.* https://doi.org/10.1038/483134a

VandeCreek, D. E. (2016). Text mining at an institution with limited financial resources. *D-Lib Magazine*, *22*(7). https://doi.org/10.1045/july2016-vandecreek

Williams, L.A., Fox, L.M., Roeder, C., & Hunter, L. (2014). Negotiating a text mining license for faculty researchers. *Information Technology and Libraries, 33*(3). https://doi.org/10.6017/ital.v33i3.5485

Websites

ContentMine http://contentmine.org/, archived at https://perma.cc/62L6-UXBZ

FutureTDM http://www.futuretdm.eu/, archived at https://perma.cc/T26P-7WXK

The Hague Declaration http://thehaguedeclaration.com/, archived at https://perma.cc/8Y3C-Z5UL

OpenMinTeD http://openminted.eu/, archived at https://perma.cc/2LUF-L7Z8