# Negotiating a Text Mining License for Faculty Researchers

**Leslie A. Williams, Lynne M. Fox, Christophe Roeder, and Lawrence Hunter**

**ABSTRACT**

*This case study examines strategies used to leverage the library's existing journal licenses to obtain a large collection of full-text journal articles in XML format, the right to text mine the collection, and the right to use the collection and the data mined from it for grant-funded research to develop biomedical natural language processing (BNLP) tools. Researchers attempted to obtain content directly from PubMed Central (PMC). This attempt failed because of limits on use of content in PMC. Next, researchers and their library liaison attempted to obtain content from contacts in the technical divisions of the publishing industry. This resulted in an incomplete research data set. Researchers, the library liaison, and the acquisitions librarian then collaborated with the sales and technical staff of a major science, technology, engineering, and medical (STEM) publisher to successfully create a method for obtaining XML content as an extension of the library's typical acquisition process for electronic resources. Our experience led us to realize that text-mining rights of full-text articles in XML format should routinely be included in the negotiation of the library's licenses.*

## INTRODUCTION

The University of Colorado Anschutz Medical Campus (CU Anschutz) is the only academic health sciences center in Colorado and the largest in the region. Annually, CU Anschutz educates 3,480 full-time students, provides care during 1.5 million patient visits, and receives more than $400 million in research awards.[1] CU Anschutz is home to a major research group in biomedical natural language processing (BNLP), directed by Professor Lawrence Hunter. Natural language processing (also known as NLP or, more colloquially, "text mining") is the development and application of computer programs that accept human language, usually in the form of documents, as input. BNLP takes as input scientific documents, such as journal articles or abstracts, and provides useful

**Leslie A. Williams** (leslie.williams@ucdenver.edu) is Head of Acquisitions, Auraria Library, University of Colorado, Denver. **Lynne M. Fox** (lynne.fox@ucdenver.edu) is Education Librarian, Health Sciences Library, University of Colorado Anschutz Medical Campus, Aurora.
**Chistophe Roeder** is a researcher at the School of Medicine, University of Colorado, Aurora.
**Lawrence Hunter** (larry.hunter@ucdenver.edu) is Professor, School of Medicine, University of Colorado, Aurora.

functionality, such as information retrieval or information extraction. CU Anschutz's Health Sciences Library (HSL) supports Hunter's research group by providing a reference and instruction librarian, Lynne Fox, to participate on the research team. Hunter's group is working on computational methods for knowledge-based analysis of genome-scale data.[2] As part of that work, his group is devising and implementing text-mining methods that extract relevant information from biomedical journal articles, which is then integrated with information from gene-centric databases and used to produce a visual representation of all of the published knowledge relevant to a particular data set, with the goal of identifying new explanatory hypotheses.

Hunter's research group demonstrated the potential of integrating data and research information in a visualization to further new discoveries with the "Hanalyzer" (http://hanalyzer.sourceforge.net). Their test case used expression data from mice related to craniofacial development and connected that data to PubMed abstracts using gene or protein names. "Copying of content that is subject to copyright requires the clearing of rights and permissions to do this. For these reasons the body of text that is most often used by researchers for text mining is PubMed."[3] The resulting visualization allowed researchers to identify four genes involved in mouse craniofacial development that had not previously been connected to tongue development, with the resulting hypotheses validated by subsequent laboratory experiment.[4] The knowledge-based analysis tool is open access.

To continue the development of the BNLP tools for the knowledge-based analysis system, three things were required: a large collection of full-text journal articles in XML format, the right to text mine the collection, and the right to store and use the collection and the data mined from it for grant-funded research. The larger the dataset, the more robust the visual representations of the knowledge-based analysis system, so Hunter's research group sought to compile a large corpus of relevant literature, beginning with journal articles. The text that is mined can start in many formats; however, XML provides a computer-ready format for text mining because it is structured to indicate parts of the document. XML is "called a 'markup language' because it uses tags to mark and delineate pieces of data. The 'extensible' part means that the tags are not pre-defined; users can define them based on the type of content they are working with."[5,6]

XML has been adopted as a standard for content creation by journal publishers because it provides a flexible format for electronic media.[7] XML allows the parts of a journal article to be encoded with tags that identify the title, author, abstract, and other sections, allowing the article to be transmitted electronically between editor and publisher and to be easily formatted and reproduced into different versions (e.g., print, online). XML can also indicate significant content in the text, such as biological terms or concepts. XML allowed Hunter's research group to write computer programs that can make sense of each article by using the XML tags as indicators of content and placement within the article. Products have been developed, such as LA-PDFText, to extract text from PDF documents.[8] However, direct access to XML provides more useful corpora

because the document markup saves time and improves the accuracy of results extracted from XML.

Once the sections and content of an article are identified, text-mining techniques are applied to the article. "Text mining extracts meaning from text in the form of concepts, the relationships between the concepts or the actions performed on them and presents them as facts or assertions."[9] Text-mining techniques can be applied to any type of information available in machine-readable format (e.g., journal article, e-books). A dataset is created when the text-mined data is aggregated. Using BNLP tools, Hunter's research group's knowledge-based analysis system analyzed the dataset and produced visual representations of the knowledge that have the potential to lead to new hypotheses. Text mining and BNLP techniques have the potential to build relationships between the knowledge contained in the scholarly literature that lead to new hypothesis resulting in more rapid advances in science.

## LITERATURE REVIEW

Hunter and Cohen explored "literature overload" and its profoundly negative impact on discovery and innovation.[10] With an estimated growth rate of 3.1 percent annually for PubMed Central, the US National Library of Medicine's repository, researchers struggle to master the new literature of their field using traditional methods. Yet much of the advancement of biological knowledge relies on the interplay of data created by protein, sequence, and expression studies and the communication of information and discoveries through nontextual and textual databases and published reports.[11] How do biomedical researchers capitalize on and integrate the wealth of information available in the scholarly literature? "The common ground in the area of content mining is in the shared conviction that the ever increasing overload of information poses an absolute need for better and faster analysis of large volumes of content corpora, preferably by machines."[12]

BNLP "encompasses the many computational tools and methods that take human-generated texts as input, generally applied to tasks such as information retrieval, document classification, information extraction, plagiarism detection, or literature-based discovery."[13] BNLP techniques accomplish many tasks usually performed manually by researchers, including enhancing access through expanded indexing of content or linkage to additional information, automating reviews of the literature, discovering new insights, and extracting meaning from text.[14] Text mining is just one tool in a larger BNLP toolbox of resources used to read, reason, and report findings in a way that connects data to information sources to speed discovery of new knowledge.[15] According to pioneering text-mining researcher Marti Hearst, "Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of

experimentation."[16] Biomedical text mining uses "automated methods for exploiting the enormous amount of knowledge available in the biomedical literature."[17]

Recent reports, commissioned by private and governmental interest groups, discuss the economic and societal value of text mining.[18,19] The McKinsey Global Institute estimates the worth of harnessing big data insights in US health care at $300 billion. The report concludes that greater sharing of data for text mining enables "experimentation to discover needs, expose variability, and improve performance" and enhances "replacing/supporting human decision making with automated algorithms," among other benefits. Furthermore, the McKinsey report points out that North America and Europe have the greatest potential to take advantage of innovation because of a well-developed infrastructure and large stores of text and data to be mined.[20] However, these new and evolving technologies are challenging the current intellectual-property framework as noted in an independent report by Ian Hargreaves, "Digital Opportunity: A Review of Intellectual Property and Growth," resulting in lost opportunity for innovation and economic growth.[21] In "The Value and Benefits of Text Mining," JISC finds copyright restrictions limit access to content for text mining in the biomedical sciences and chemistry and that costs for access and infrastructure prevent entry into text-mining research for many noncommercial organizations.[22] Despite copyright barriers, organizations surveyed pointed out the risks associated with failing to use text-mining techniques to further research include financial loss, loss of prestige, opportunity lost, and the brain drain of having talented staff seek more fulfilling work. JISC explores a research project's workflow and finds a lack of access to text mining delayed the publication of an important medical research study by many months, or the time the research team spent analyzing and summarizing relevant research.[23] Both reports advocate an exception to intellectual property rights for noncommercial text-mining research to balance the protection of intellectual property with the access needs of researchers. A centrally maintained repository for text mining has been proposed, although its creation would face significant challenges.[24]

Scholarly journal content is the raw "ore" for text mining and BNLP. The lack of access to this ore creates a bottleneck for researchers. "New business models for supporting text mining within the scholarly publishing community are being explored; however, evidence suggests that in some cases lack of understanding of the potential is hampering innovation."[25] BNLP and machine-learning research products are more accurate and complete when more content is available for text mining. "Knowledge discovery is the search for hidden information. . . . Hence the need is to start looking as widely as possible in the largest set of content sources possible."[26] However, as noted in a *Nature* article, "The question is how to make progress today when much research lies behind subscription firewalls and even 'open' content does not always come with a text-mining license."[27] Large scientific publishers are facing economic challenges, and potentially diminished economic returns, as the tension over the right to use licensed content heats up. *Nature*, the flagship of a major scientific publisher, predicted "trouble at the text mine" if researchers lack access to the contents of research publications.[28] And a 2012 investment report predicted slower

earnings growth for Elsevier, the largest STEM publisher, if it blocked access to licensed content by text-mining researchers. The review predicted, "If the academic community were to conclude that the commercial terms imposed by Elsevier are also hindering the progress of science or their ability to efficiently perform research, the risk of a further escalation of the acrimony [between Elsevier and the academic community] rises substantially."[29] With open access alternatives proliferating, including making federally funded research freely accessible, STEM publishers are under increased pressure to respond to market forces. "The greatest challenge for publishers is to create an infrastructure that makes their content more machine-accessible and that also supports all that text-miners or computational linguists might want to do with the content."[30] On the other end of the spectrum, researchers are struggling to gain legal access to as much content as possible.

Academic libraries have long excelled at serving as the bridge between researchers and publishers and can expand their roles to include navigating the uncharted territory of obtaining text-mining rights for content. Increasing the library's role in text mining and other associated BNLP and machine-learning methods offers tremendous potential for greater institutional relevance and service to researchers.[31] At CU Anschutz's HSL, Fox and Williams, an acquisitions librarian, found natural opportunities for collaboration including negotiating rights to content more efficiently through expanded licensing arrangements and facilitating the secure transfer and storage of data to protect researchers and publishers.

## METHOD

Hunter and Fox began working in 2011 to obtain a large corpus of biomedical journal articles in XML format to create a body of text as comprehensive as possible for BNLP experimentation that would further advance Hunter's research group's knowledge-based analysis system. The desired result was an aggregated collection obtained from multiple publishers, stored locally, and available on demand for the knowledge-based analysis system to process. Hunter and Fox soon realized that "the process of obtaining or granting permissions for text mining is daunting for researchers and publishers alike. Researchers must identify the publishers and discover the method of obtaining permission for each publisher. Most publishers currently consider mining requests on a case by case basis."[32] They pursued a multifaceted strategy to build a robust collection and to determine which strategy proved most fruitful because, during a grant review, National Library of Medicine staff wanted evidence of access to an XML collection before awarding a grant.

Fox first approached two open-access publishers, BioMed Central (BMC) and Public Library of Science (PLoS), to request access to XML text from journals in the subjects of life and biomedical science. Fox had existing contacts within both organizations and an agreement was reached to obtain XML journal articles. Letters of understanding were quickly obtained as both publishers were excited about exploring new ways for their research publications to be accessed and the potential to increase the use of their journals. Possible journal titles were identified and

arrangements were made to transfer and store files locally from BMC and PLoS to Hunter's research group.

Hunter approached staff at PubMedCentral (PMC) to request access to articles and discovered they could only be made available with permission from publishers. A Wiley research and product development executive granted Hunter permission to access Wiley articles in PMC. The Wiley executive was interested in learning what impact text mining might have on Wiley products. Hunter's research group planned to transfer Document Type Definition (DTD) format files from PMC. Unfortunately, when Hunter's research group staff requested file-transfer assistance from PMC, no PMC staff were available to provide the technical help needed because to budget reductions. PMC staff could accurately evaluate their time commitment because they had a clear understanding of the XML access and transfer process, and knew they could not allocate resources to the effort.

Hunter then began to leverage his professional network connections to obtain content from a major STEM vendor. Research and development division directors within the company were familiar with the work of Hunter's research group and were willing to provide assistance in acquiring content. However, when the research group began to perform research using this data, further investigation determined that the contents were not adequate for the research. Follow-up between Fox, the research group, and the vendor revealed that the group's needs were not communicated in the vendor's vernacular, resulting in the group not clearly understanding what content the vendor was providing. This disconnect occurred in the communication flow from the research group to the vendor's research and development staff to the vendor's sales staff (who identified the content to be shared). It was a like a game of telephone tag.

After the initial strategies produced mixed results, Hunter's research group hypothesized that they could harvest materials through HSL's journal subscriptions. Hunter's research group attempted to crawl and download journal content being provided by HSL's subscription to a major chemistry publisher. Since publishers monitor for web crawling of their content, the chemistry publisher became aware of the unusual download activity, turned off campus access, and notified the library that there may have been an unauthorized attempt to access the publisher's content. Researchers are often unaware of complex copyright and license compliance requirements. In fact, librarians sometimes become aware of text-mining projects only after automated downloads of licensed content prompt vendors to shut off campus access.[33] Libraries can prevent interruption of campus-wide access to important resources by suggesting more effective content-access methods.

Williams, an HSL acquisitions librarian, investigated the interruption in access and discovered Hunter's research group's efforts to obtain journal articles to text mine for their research. She offered to use her expertise in acquiring content to help Hunter's research group obtain the dataset needed for their research. Initially, Hunter and Fox had not included an acquisitions

librarian because that position was vacant. After Williams became involved, the effort focused on licensing content via negotiation and licensing with individual publishers.

## RESULTS

"There are a large number of resources to help the researcher who is interested in doing text mining" but "no similar guide to obtaining the necessary rights and permissions for the content that is needed."[34] At CU Anschutz, this vacuum was filled by Williams, who is knowledgeable about the acquisition of content, and Fox, who is knowledgeable about Hunter's research, serving as the bridge between the research group and the STEM publisher. By working together and capitalizing on each other's expertise, Williams and Fox were able to facilitate the collaboration that developed a framework for purchasing a large collection of full-text journal articles in XML format. As the collaboration progressed, three major elements to the framework surfaced, including a pricing model, a license agreement, and the dataset and delivery mechanism.

Researchers interested in legally text mining journal content often find themselves having to execute a license agreement and pay a fee.[35] What should the fee be based on to create a fair and equitable pricing model? Publishers establish pricing for library clients on the basis of not only the content, but many valued-added services such as the breath of titles aggregated and made available for purchase in a single product, the creation of a platform to access the journal titles, the indexing and searching functionality within the platform, and the production of easily readable PDF versions of articles. These value-added services are not required for text-mining endeavors. Rather, the product is the raw journal content that has been peer-reviewed, edited, and formatted in XML that precedes the addition of value-added services. Therefore the pricing should not be equivalent to the cost of a library's subscription to a journal or package of journals. In the end, after lengthy negotiations, the pricing model for the Hunter's research group collection of full-text journal articles in XML format consisted of

- a cost per article;
- a minimum purchase of 400,000 articles for one sum on the basis of the cost per article;
- an annual subscription for the minimum purchase of 400,000;
- the ability to subscribe to additional articles in excess of 400,000 in quantities determined by Hunter's research group;
- a volume discount off the per article price for every article purchased in excess of 400,000;
- inclusion of the core journal titles purchased via the library's subscription at no charge;
- inclusion of the core journal titles purchased by the University of Colorado Boulder at no charge because of Hunter's joint appointment at both CU Boulder and CUAnschutz campuses; and
- a requirement for HSL to maintain its subscription to the vendor's product at its current level.

"Where institutions already have existing contracts to access particular academic publications, it is often unclear whether text mining is a permissible use."[36] From the beginning, common ground was easily found on the subject of core titles purchased by the two campuses' libraries. Core titles are typically those journals that libraries pay a premium for to obtain perpetual rights to the content. Most of the negotiation focused on access titles, which are journals that libraries pay a nominal fee to have access to without any perpetual rights included.

The final challenge related to cost was determining how to process and pay for the product. Hunter's research group operates on major grant funding from federal government agencies. The University of Colorado requires additional levels of internal controls and approvals to expend grant funds as well as to track expenditures to meet reporting requirements of the funding agencies. Also, grant funding of this type often spans multiple fiscal years whereas the library's budget operates on a single fiscal year at a time. Therefore it was decided that Hunter would handle payment directly rather than transferring funds to HSL to make payment on their behalf.

"Libraries as the licensee of publishers' content are from that perspective interested in the legal framework around content mining."[37] During price negotiations, Williams recommended negotiating a license agreement similar to those libraries and publishers execute for the purchases of journal packages. A license agreement would offer a level of protection for all parties involved while clearly outlining the parameters of the transaction. Hunter and the STEM publisher readily agreed.

The final license agreement contained ten sections including definitions; subscription; obligations; use of names; financial arrangement; term; proprietary rights; warranty, indemnity, disclaimer, and limitation of liability; and miscellaneous. While the license agreement was similar to traditional license agreements between libraries and publishers for journal subscriptions, there were some notable differences. First, in the definitions section, users were defined and limited to Hunter and his research team. This limited the users to a specific group of individuals unlike typical library–publisher license agreements that license content for the entire campus.

Second, the subscription section covered how the data can be used in detail and allowed the dataset to be installed locally. This was important to make the dataset available on demand to researchers; to allow researchers to manipulate, segment, and store the data in multiple ways instead of as one large dataset; and to allow the researchers the ability to access and use the large dataset efficiently and quickly. Because the dataset would be manipulated so extensively, the license gave permission to create a backup copy and store it separately. The subscription section also required the dissemination of the research results to occur in such a way that the dataset could not be extracted and used by others. This was significant because Prof. Hunter releases the BNLP software applications they develop as open source software so that the applications can be open to peer review and attempts at reproduction. Ideally, someone could download the open source software, obtain the same corpus as input, and see the same output mentioned in the paper.

Third, the obligations section was radically different from traditional library–publisher license agreements because even though "publishers are still working out how to take advantage of text mining . . . none wants to miss out on the potential commercial value."[38] This interest prompted the crafting of an atypical obligations section in the license agreement that included an option for Hunter to collaborate with the STEM publisher to develop and showcase an application on the vendor's website and included a commitment for Hunter to meet quarterly with the vendor's representatives to discuss advances in research. Furthermore, the obligations section specified a request for Hunter and the University of Colorado to recognize the vendor where appropriate and a right for the STEM publisher to use any research software application released as open source. Up to this point, Williams had been collaborating with the University of Colorado in-house counsel to review and revise the license agreement. When the STEM publisher requested the right to use the software application, Williams was required to submit the license agreement to the University of Colorado's Technology Transfer Office for review and approval. Approval was prompt in coming, primarily because Prof. Hunter releases his software applications as open source.

Fourth, the license agreement included a "use of names" section, which is not found in typical library–publisher agreements. This section authorized the vendor to use factual information drawn from a case study in market-facing materials and a requirement that the vendor request written consent, as required from the University of Colorado System, for information in the case study to be released for market facing materials. The vendor also agreed not to use the University of Colorado's trademark, service mark, trade name, copyright, or symbol without prior written consent and to use these items in accordance with the University of Colorado System's usage guidelines.

Fifth, the vendor agreed not to represent in any way that the University of Colorado or its employees endorse the vendor's products or services. This is extremely important because the University of Colorado's controller does not allow product endorsements because of the federal unrelated business income tax. Exempt organizations are required to pay this tax if engaged in activities that are regularly occurring business activities that do not further the purpose of the exempt organization.[39]

Finally, the license agreement stated all items would be provided in XML format with a unique Digital Object Identifier (DOI) number, essential for linking XML content to real-world documents that researchers using Hunter's research group's knowledge-based analysis system would want to access.

After a pricing model and license agreement were finalized, the focus turned to the last major element of the framework: the dataset and delivery mechanism. Elements such as quality of the corpora contents, file transfer time, and storage capacity are all important. In other words, "the need is to start looking as widely as possible in the largest set of content sources possible. This need is balanced by the practicalities of dealing with large amounts of information, so a choice

needs to be made of which body of content will most likely prove fruitful for discovery. Text mines are dug where there is the best chance of finding something valuable."[40]

When building an XML corpora for research, Hunter's research group wanted to maximize their return on investment, so a pilot download was conducted to assure that the most beneficial content could be transferred smoothly to a local server. "Permissions and licensing is only a part of what is needed to support text mining. The content that is to be mined must be made available in a way that is convenient for the researcher and the publisher alike."[41] This pilot phase allowed Hunter's researchers and the vendor's technical personnel to clarify the requirements of the dataset and to efficiently deliver and accurately invoice for content. One of the initial obstacles was that a filter for the delivery mechanism didn't exit. Letters to the editor, errata, and more were all counted as an article. Hunter's researchers quickly determined that research articles were most important at this point in the development of the knowledge-based analysis system. How should a *useful* or *minable* article be defined—by its length, by XML tags indicating content type, or by some other criteria? Roeder, a software engineer, used article attributes and characteristics embedded in XML tags to define an article as including all of the following:

- an abstract
- a body
- at least 40 lines of text
- none of the following tags: corrigendum, erratum, book review, editorial, introduction, preface, correspondence, or letter to the editor

In the end, Hunter's research group and the vendor agreed to transmit everything and allow the group a fifteen business days to evaluate the content. The research group would then notify the vendor of how many "articles" were received. This process would continue until 400,000 "articles" were received.

After spending more than a year working to develop a structure to purchase a large corpus of journal articles to text mine. Just as Hunter's research group was ready to execute the license, remit payment, and receive the articles, their federal grant expired, stalling the purchase. In retrospect, this unfortunate development was the catalyst for a shift in philosophy and strategy for the researchers and librarians at CU Anschutz.

## DISCUSSION

XML text-mining efforts will continue to expand, leading to increased demand on libraries and librarians to play a role in securing content. Publishers, researchers, and libraries see the potential commercial and research value for text mining journal content and are driving the rapid evolution of this arena, in part, because "there is increasing demand from public and charitable funders that maximum value is leveraged from their substantial investment and this includes making outputs

accessible and usable. . . . Text mining offers the potential for fuller use of the existing publicly-funded research base."[42]

However, publishers identified two main barriers to text mining from their perspective—lack of standardization in content formats and in access terms—and concede that "publishers should develop shared access terms for research-driven mining requests."[43] From the researcher and librarian perspective, there are many barriers and costs involved including "access rights to text-minable materials, transaction costs (participation in text mining), entry (setting up text mining), staff and underlying infrastructure. Currently, the most significant costs are transaction costs and entry costs."[44] The significant transaction costs stem from the time it takes to navigate the complexity of negotiating and complying with license agreements for journal content. The various types of "costs are currently borne by researchers and institutions, and are a strong hindrance to text mining uptake. These could be reduced if uncertainty is reduced, more common and straightforward procedures are adopted across the board by license holders, and appropriate solutions for orphaned works are adopted. However, the transaction costs will still be significant if individual rights holders each adopt different licensing solutions and barriers inhibiting uptake will remain."[45]

In a survey of libraries, findings indicated that librarians anticipate a new role as facilitators between researchers and publishers to enable text mining.[46] Librarians are a natural fit for this role because they already have expertise in navigating copyright, requesting copyright permissions, and negotiating license agreements for journal content. "Advice and guidance should be developed to help researchers get started with text mining. This should include: when permission is needed; what to request; how best to explain intended work and how to describe the benefits of research and copyright owners."[47]

After their experience with developing a framework to license and purchase a large corpora of journal articles in XML format to be text mined, Fox and Williams came to believe that, in addition to providing copyright expertise, librarians should assist in reducing transaction costs by developing model license clauses for text mining and routinely negotiating for these rights when the library purchases journals and other types of content. Adopting this philosophy and strategy led Williams and Fox to successfully advocate for the inclusion of a text-mining clause in the license agreement for the STEM publisher in this case study at the time of the library's subscription renewal. This occurred at a regional academic consortium level, making text mining easier at fourteen academic institutions. Furthermore, the University of Colorado Libraries, which includes five libraries on four campuses, is now working on drafting a model clause to use when purchasing journal content as the University of Colorado System and to put forth for consideration by the consortiums that facilitate the purchase of our major journal packages. Given that incorporating text mining clauses into library–publisher license agreements for scholarly journals is in its infancy, there are few resources available to assist librarians adopting this new role. Model clauses include the following:

- British Columbia Electronic Library Network's Model License Agreement[48]
    - Clause 3.11. "Data and Text Mining. Members and Authorized Users may conduct research employing data or text mining of the Licensed Materials and disseminate results publicly for non-commercial purposes."
- California Digital Library's Standard License Agreement[49]
    - Section IV. Authorized Use of Licensed Materials. "Text Mining. Authorized Users may use the licensed material to perform and engage in text mining/data mining activities for legitimate academic research and other educational purposes."
- JISC's Model License for Journals[50]
    - Clause 3.1.6.8. "Use the Licensed Material to perform and engage in text mining/data mining activities for academic research and other Educational Purposes and allow Authorised Users to mount, load and integrate the results on a Secure Network and use the results in accordance with this License."
    - Clause 9.3. "For the avoidance of doubt, the Publisher hereby acknowledges that any database rights created by Authorised Users as a result of textmining/datamining of the Licensed Material as referred to in Clause 3.1.6.8 shall be the property of the Institution."

Publishers are also beginning to break down barriers perhaps, in part, because of the sentiment that "privately erected barriers by copyright holders that restrict text mining of the research base could be increasingly regarded as inequitable or unreasonable since the copyright holders have borne only a small proportion of the costs involved in the overall process; furthermore, they do not have rights or ownership of the inherent facts or ideas within the research base."[51] BioMed Central and PLoS both offer services that allow researchers to access XML text collections. BioMed Central makes content readily accessible by providing a website for bulk download of XML text.[52] PLoS requires contact with a staff member for download of XML text.[53] In December 2013, Elsevier also announced that it would create a "big data" center at the University College London to allow researchers to work in partnership with Mendeley, a knowledge management and citation application now owned by Elsevier. While this is a positive step, the partnership does not appear to make the data available to research groups beyond the University College London.[54]

However, there is still a long way to go before publishers and librarians are routinely collaborating on opening up the scholarly literature to be mined. For example, a 2012 *Nature* editorial states "Nature Publishing Group, which also includes this journal, says that it does not charge subscribers to mine content, subject to contract."[55] Repeated attempts by Williams to obtain more information from Nature Publishing Group and a copy of the contract have proved fruitless.

In January 2014, Elsevier announced that "researchers at academic institutions can use Elsevier's online interface (API) to batch-download documents in computer-readable XML format" after

signing a legal agreement. Elsevier will limit researchers to accessing 10,000 articles per week.[56,57] For small-scale projects with a narrow scope, this limit will suffice. For example, mining the literature for a specific gene that plays a known role in a disease could require a text set under 30,000 articles. At Elsevier's current rate of article transfer, a 30,000 article text set could be created in roughly three weeks. However, for large-scale projects such as Hunter's research group's knowledge-based analysis system that require a text set of 400,000 articles (or much more, if not limited by budget constraints), nearly a year of time would be required to build the corpora. Time is one of the most valuable commodities in computational biology. The elapsed time required to transfer articles at the rate of 10,000 articles per week represents a bottleneck that most grant-funded research cannot afford. Speed of transfer will also be a factor. Researchers require flexibility to maximize available central processing unit (CPU) hours because documents can take from a few seconds to a full minute each to transfer to the storage destination. Monopolizing peak hours in high performance computing (HPC) settings may mean that computing power is not available for other tasks, although many HPC centers have learned to allocate CPU use more efficiently to high volumes. Furthermore, the terms and conditions set by Elsevier for output limits excerpting from the original text to 200 characters.[58] This is roughly equivalent to two lines of text or approximately forty words. This may be insufficient to capture important biological relationships necessary to evaluate the relevance of the article to the research being represented by the Hanalyzer knowledge-based analysis system.

## CONCLUSION

Forging a partnership between a library, a research lab, and a major STEM vendor requires flexibility, patience, and persistence. Our experience strengthened the existing relationship between the library and the research lab and demonstrated the library's willingness and ability to support faculty research in a nontraditional method. Librarians are encouraged to advocate for the inclusion of text-mining rights in their library's license agreements for electronic resources.

What the future holds for publishers, researchers, and libraries involved in text mining remains to be seen. However, what is certain is that without cooperation between publishers, researchers, and libraries, breaking down the existing barriers and achieving standards for content formats and access terms will remain elusive.

## REFERENCES

1. University of Colorado Anschutz Medical Campus, University of Colorado Anschutz Medical Campus Quick Facts, 2013, http://www.ucdenver.edu/about/WhoWeAre/Documents/CUAnschutz_facts_041613.pdf.

2. Sonia M. Leach et al., “Biomedical Discovery Acceleration, with Applications to Craniofacial Development,” *PLoS Computational Biology* 5, no. 3 (2009): 1–19, http://dx.doi.org/10.1371/journal.pcbi.1000215.

3. Jonathan Clark, *Text Mining and Scholarly Publishing* (Publishing Research Consortium, 2013).

4. Corie Lok, “Literature Mining: Speed Reading,” *Nature* 463 (2010): 416–18, http://dx.doi.org/10.1038/463416a.

5. Hong-Jie Dai, Yen-Ching Chang, Richard Tzong-Han Tsai, Wen-Lian Hsu, "New Challenges for Biological Text-Mining in the Next Decade," *Journal of Computer Science and Technology* 25, no.1 (2010): 169-179, doi: 10.1007/s11390-010-9313-5.

6. Anne Hoekman, “Journal Publishing Technologies: XML,” http://www.msu.edu/~hoekmana/WRA%20420/ISMTE%20article.pdf.

7. Alex Brown, "XML in Serial Publishing: Past, Present and Future," *OCLC Systems & Services* 19, no. 4, (2003):149-154, doi: 10.1108/10650750310698775.

8. Cartic Ramakrishnan et al., “Layout-Aware Text Extraction from Full-Text PDF of Scientific Articles,” *Source Code for Biology and Medicine* 7, no. 7 (2012), http://dx.doi.org/10.1186/1751-0473-7-7.

9. Ibid.

10. Lawrence Hunter and K. Bretonnel Cohen, “Biomedical Language Processing: Perspective What’s Beyond PubMed?” *Molecular Cell* 21, no. 5, (2006): 589–94.

11. Martin Krallinger, Alfonso Valencia, and Lynette Hirschman, “Linking Genes to Literature: Text Mining, Information Extraction, and Retrieval Applications for Biology,” *Genome Biology* 9, supplement 2 (2008): S8.1–S8.14, http://dx.doi.org/10.1186/gb-2008-9-S2-S8.

12. Eefke Smit and Maurits van der Graaf, “Journal Article Mining: The Scholarly Publishers’ Perspective,” *Learned Publishing* 25, no. 1 (2012): 35–46, http://dx.doi.org/10.1087/20120106.

13. Hunter and Cohen, “Biomedical Language Processing,” 589.

14. Clark, *Text Mining and Scholarly Publishing.*

15. Leach et al., “Biomedical Discovery Acceleration.”

16. Marti Hearst, “What is Text Mining?” October 17, 2003, http://people.ischool.berkeley.edu/~hearst/text-mining.html.

17. K. Bretonnel Cohen and Lawrence Hunter, "Getting Started in Text Mining," *PLoS Computational Biology* 4, no. 1 (2008): 1–3, http://dx.doi.org/10.1371/journal.pcbi/0040020.

18. JISC, "The Model NESLi2 Licence for Journals," 2013, http://www.jisc-collections.ac.uk/Help-and-information/How-Model-Licences-work/NESLi2-Model-Licence-/.

19. Ian Hargreaves, "Digital Opportunity: A Review of Intellectual Property and Growth," May 2011, http://www.ipo.gov.uk/ipreview-finalreport.pdf.

20. James Manyika et al., "Big Data: The Next Frontier for Innovation, Competition, and Productivity," McKinsey & Company, May 2011, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

21. Hargreaves, "Digital Opportunity."

22. Diane McDonald and Ursula Kelly, "The Value and Benefits of Text Mining to UK Further and Higher Education," JISC, 2012, http://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining.

23. JISC, "The Model NESLi2 Licence for Journals."

24. Smit and van der Graaf, "Journal Article Mining."

25. McDonald and Kelly, "The Value and Benefits of Text Mining."

26. Clark, *Text Mining and Scholarly Publishing.*

27. "Gold in the Text?" *Nature* 483 (March 8, 2012): 124, http://dx.doi.org/10.1038/483124a.

28. Richard Van Noorden, "Trouble at the Text Mine," *Nature* 483 (March 8, 2012): 134–35.

29. Claudio Aspesi, A. Rosso, and R. Wielechowski. *Reed Elsevier: Is Elsevier Heading for a Political Train-Wreck?* 2012.

30. Clark, *Text Mining and Scholarly Publishing.*

31. Jill Emery, "Working In A Text Mine: Is Access about to Go Down?" *Journal of Electronic Resources Librarianship* 20, no. 3 (2009):135–38, http://dx.doi.org/10.1080/19411260802412745.

32. Clark, *Text Mining and Scholarly Publishing*: 14.

33. Van Noorden, "Trouble at the Text Mine."

34. Ibid.

35. Ibid.

36. JISC, “The Model NESLi2 Licence for Journals.”
37. Smit and van der Graaf, “Journal Article Mining.”
38. Van Noorden, “Trouble at the Text Mine.”
39. Internal Revenue Service, “Unrelated Business Income Defined,” http://www.irs.gov/Charities-&-Non-Profits/Unrelated-Business-Income-Defined.
40. Clark, *Text Mining and Scholarly Publishing*: 10.
41. Ibid: 14.
42. McDonald and Kelly, “The Value and Benefits of Text Mining.”
43. Smit and van der Graaf, “Journal Article Mining.”
44. McDonald and Kelly, “The Value and Benefits of Text Mining.”
45. Ibid.
46. Smit and van der Graaf, “Journal Article Mining.”
47. McDonald and Kelly, “The Value and Benefits of Text Mining.”
48. British Columbia Electronic Library Network, BC ELN Database Licensing Framework, http://www.cdlib.org/services/collections/toolkit/.
49. “Licensing Toolkit,” California Digital Library, http://www.cdlib.org/services/collections/toolkit/.
50. JISC, “The Model NESLi2 Licence for Journals.”
51. McDonald and Kelly, “The Value and Benefits of Text Mining.”
52. “Using BioMed Central’s Open Access Full-Text Corpus for Text Mining Research,” http://www.biomedcentral.com/about/datamining.
53. “Help Using This Site,” PLOS, http://www.plosone.org/static/help.
54. Iris Kisjes, “University College London and Elsevier Launch UCL Big Data Institute,” *Elsevier Connect*, press release, December 18, 2013, http://www.elsevier.com/connect/university-college-london-and-elsevier-launch-ucl-big-data-institute.
55. “Gold in the Text?”
56. Richard Van Noorden, “Elsevier Opens Its Papers to Text-Mining,” *Nature* 506 (February 2, 2014): 17.
57. Sciverse, Content APIs, http://www.developers.elsevier.com/cms/content-apis.

58. “Text and Data Mining,” Elsevier, , http://www.elsevier.com/about/universal-access/content-mining-policies.