Charleston Library Conference

Text and Data Mining Contracts: The Issues and Needs

Nancy Herther
University of Minnesota

Daniel Dollar
Yale University Library

Darby Orcutt
North Carolina State University Libraries

Alicia Wise
Elsevier

Meg White
Rittenhouse Book Distributors

Nancy Herther, Daniel Dollar, Darby Orcutt, Alicia Wise, and Meg White, "Text and Data Mining Contracts: The Issues and Needs" (2015). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284316233

Text and Data Mining Contracts: The Issues and Needs

Nancy Herther, Sociology Librarian, University of Minnesota

Daniel Dollar, Director of Collection Development, Yale University Library

Darby Orcutt, Assistant Head, Collection Management, North Carolina State University Libraries

Alicia Wise, Director, Elsevier

Moderator: Meg White, Executive Director of Technology Services, Rittenhouse Book Distributors

Meg White: Good afternoon, folks, and welcome. This is our second Neapolitan session of the day in ballroom number two, so if that is not what you were looking for now is your time to make your escape. We thank the audience for your stamina at the end of a very long day, but we're glad you're here for what promises to be a very interesting session. So I'm going to turn this over to Nancy and her panel for "Text and Data Mining: Licensing Issues."

Nancy Herther: And thank you, Meg, and welcome to all of you here today. I'm glad you've been able to find us and I certainly again agree. It's kind of been a long day. It's a great conference; I'm sure you agree as well, but we're getting so much information. We're hoping that we can organize the session a little bit differently. We're going to be having some presentations and then really we're going to turn it over to you. I had a chance to work on a couple of articles for Katina for the website on text and data mining, and one of the things that I found in doing that was that, at least here in the United States and Canada and North America, it's still so fresh, it's still so new, there are so many questions and a lot of the answers that one organization doesn't necessarily work at all for another. So it was very interesting, and quite an education for me, and because of that the lesson I took from that was we need to start talking a lot more about text and data mining, what's happening around us. Maybe if you don't have a lot going on, then understand and listen and share with the people around you who maybe have a little bit more experience. There is no one formula. We don't have any one way of doing it. Someday we might but we don't today. I want to introduce our three speakers who are going to speak for 5, 7, 10 minutes on their experiences with text and data mining. We have three wonderful perspectives here. They're going to present to you certainly not the only perspectives that we might have, but I think all of these are valuable. They have all agreed that they would be happy to take your questions, and answer whatever questions you might have and give you whatever sort of information or insight that they have to give you. I'm going to go ahead now and introduce our speakers; excuse me for the paper here. So, our first speaker is Daniel Dollar. He's the director of collection development at Yale University Libraries, and he has been involved and he's currently involved with their budget management, an important aspect when it comes to collections, the collection steering committee, and also the executive committee for the library as a whole. He has written on a variety of different issues, certainly e-books and a lot of transformations that we have seen here, and he's going to give us a perspective being from a large, very multidisciplinary institution and the issues that he sees that are important, and how it impacts his practice as well as the current structures that exist at Yale. Our next speaker is someone who has been an academic, understands academic research. She has a PhD in anthropology and research, and that is Alicia Wise. She has been at Elsevier since June of 2010. She is leading the universal access team. In this role she is responsible for access and related policies, building relationships with other stakeholders in scholarly communication, and she has also worked for the UK Joint Information Systems Committee first to manage national

http://dx.doi.org/10.5703/1288284316233

negotiations for access to a broad range of intellectual property issues; she has experience on different aspects of these sorts of issues that we're looking at today. In Great Britain text and data mining is developing very differently than it is here in North America, so I really appreciate her perspective on this as well.

Our last speaker here, our last specialist, is Darby Orcutt. He is the assistant head of collection Management and chair of the humanities and social sciences subject team at the North Carolina state University Libraries. He has negotiated content mining access with numerous vendors, including the first ever blanket mining agreement between an institution and a major commercial vendor of historical abstracts, and that was from Gale, and similar first institutional agreements with accessible archives, Unlimited Priorities, Adam Matthew, and others that are going to be announcing in the near future. So I want to turn this over to Daniel Dollar now, our first speaker, and again we will have lots of time for questions, insights, concerns, etc., from you as we go. Thank you.

Daniel Dollar: Thank you, Nancy. All right. So, let me start my timer here. Okay. "Text and Data Mining: Licensing Issues." You saw, some of you may have seen, the video that Elsevier created on text and data mining; it's a nice overview. I'm taking this from LIBER, the Association of European Research Libraries Text and Mining website, where it talks about text and data mining as basically its machine-read material; its copying of large quantities of material which can sort of concern people when you talk about copying, but it's in order to extract data and then recombine it to find patterns. I thought it was very interesting that this morning we had Jim O'Neil talking and saying how many of you have done text and data mining. And it was like okay, every time you do a Google search, so we're all text and data miners whether we knew it or not. But, it is certainly something that is critical if we're going to try to understand a large corpus of material. We are moving beyond I think, like the horse and buggy era where we call things "e-journals" and "e-books." I mean, as the content has moved into a digital age and we have to think in new ways in order to make sense of this huge corpus of material and text, and data mining is a tool to do that and really see this as becoming an everyday tool or part of a toolkit for the discovery of new knowledge, and even in the humanities it is going to become a mainstream part of scholarly inquiry. That is what we see coming, or envision.

What are the challenges? I see four challenges: legal, pricing, access, and library support. On the legal licensing side, our position is, wait for it, we really don't see a reason to have to have separate TDM licenses. The output is subject to the same terms and conditions undertaken with any research using licensed resources, and making lawful use of the content when employing TDM is still governed and subject to the license agreements that we've signed. We can't do text and data mining and just throw content out on the web or re-create the products that we've licensed for the whole world, but we can make research use of that content again within the context of our licenses. We do accept the premise that the right to read is the right to mine whether it's close, human reading or distant, machine reading of the material and if it—Susan Riley from LIBER who was supposed to be here on this panel, you got me instead, yay! But, if she was here from LIBER I think she would talk a little bit about how in Europe they feel that they are behind. The researchers there are behind us in terms of our ability to make use of text and data mining because of fair use and the way it is interpreted in the United States. And I would direct you to the LIBER website for more information about their rationale for that.

Dropping into pricing: as a research library, we are paying a premium for content and so if we are paying a premium for content, for digitized content, then it should come with text and data mining as a matter of course. You know, I sort of think of the analogy of when back in the 1990s we started getting content on discs but you couldn't print or you couldn't save it because you were worried about what would happen to the content. We sort of got past that—we got past the early days of e-journals when we were worried about electronic ILL, I sort of feel we're in the same sort of early stages with text and data mining. There's a lot of concern—what's going to happen? The content is going to get out on the web and bad things are going to happen, but from the library's standpoint and we're paying this premium for content, if you're going to have these kinds of restrictions that in a few years I'm going to have to go back and buy text and data mining rights, then didn't I pay full royalties the first time around? Actually I already paid it more than once because I bought it on microfilm but let's—you know, oh well. We'll get it right one day, but I'm quite concerned as a chief collection development officer at my institution, going in and making six-figure purchases, or high five-figure purchases, and I'm going to have my colleagues or the person who comes after me saying, "What in the heck was he doing?" By not including this and not making sure that this was clearly understood when we made this acquisition.

Access: so in terms of access, you know, getting the raw data either on a drive or putting it in a secure location in the cloud where we can actually do text and data mining. We've had some success with vendors where we have been able to do that, we were able to get research output from that text and data mining and then point back to the publisher's website for the human readable content, and that's a real win because we're making new discoveries, our researchers can make new discoveries; it points back to the platform where we have the human readable licensed content, and what do you know? It drives up usage so it's a good thing.

So, APIs come out and are a major issue when we talk about text and data mining, and there's good and bad. We look at what JSTOR is doing with the Data for Research Program as good. It's not mediated. We are also optimistic about what the HathiTrust Research Center is working on. They're trying to work on an API to mine all of the content within HathiTrust, and that includes the post-1923 content, which is quite problematic. So, we see those as advantages. We take a Dem view with API, where it's going to be mediated or potentially mediated by others sort of looking over the shoulder of the researcher, potentially. And that accessibility: this is something that I can only touch on briefly. I understand that UC Berkeley is making some use of text and data mining for accessibility issues, and so I don't know all the mechanics of how they're doing that but it is an interesting idea that maybe they can take that content and put it into forms that can be used with someone who has disability needs in very controlled ways.

And then the final challenge is library support. We think of the sciences, the labs, linguists, those folks generally have quantitative data science skills or expertise in their labs or their departments. The humanities not so much, and so we see digital humanities centers as helping bridge the gap and helping the humanists sort of tap into these rich resources and make use of them in new and exciting ways, but that has a financial cost. Even, wait for it, at Yale it has a cost. I mean every position we hire for has to go through our Executive Council, and we have long debates. I mean we have caught up to the rest of the world and so—but we have to make these investments. We've received a grant. We're hiring and building out a digital humanities lab, and again it is a significant financial impact for us. It's not free, and even if we can make use of the content using TDM we have to make an investment on our side.

So, thank you. These are three of my colleagues who are instrumental in helping me with this presentation but don't hold them responsible for what I said. Thank you.

Alicia Wise: Hey, everybody, it's terrific to be here today and I really enjoyed already learning from my fellow panel members. We've had some lively discussions in preparation for this, which got me thinking afresh and I hope the discussion today will continue that. So, let me offer a publishing perspective on the same space. TDM is interesting and it's challenging for all of us, and it's really important I think that libraries and publishers work together because we are all supporting researchers so that the more we are aligned the easier it is for them. We have a wide array of TDM programs. We're actually working in close partnership with specific universities. We have a policy but we also have service channels, so we're supporting TDM miners through a development portal and we provide quite a lot of technical support and services beyond just access to the content. Some of our corporate customers, for example, work with us quite closely to develop TDM software and tools. That isn't a call that's been made on us for the academic community, or at least not yet, but it's perhaps something that will come in the future.

We have been engaged in text and data mining support for almost 10 years now. It started off in a very small way with one group of researchers in California who wanted access to our publications for developing text and data mining tools back in 2006, and so that was interesting—those really early adopters. We stepped up I think and made a scale change in our support for TDM, and in about 2013 when we engaged in around 30 pilot projects to better understand the challenges of mining for researchers and libraries and publishers and we kind of thought it would explode at that point, but interestingly I think we along with other publishers are seeing a slow, steady growth in interest in text and data mining and it hasn't been that explosion yet. It still feels like we're very much in the early phases. We're still actually engaging with early adopters, which is interesting.

Okay, so some of the challenges from a researcher's perspective: they need access to computers and code and things that they may not easily have. A lot of the early adapters are actually writing their own code, and that's a barrier potentially to the wider use and uptake of mining. They have to deal with access to content in a variety of different formats across different platforms, and there's a lot of refinement and testing and learning and gurning through data and the outputs to kind of get meaningful results, and I think gaining access to those different platforms and permissions to mine across all the different platforms is of course a challenge for them and something that would be helpful for all of us to help resolve.

I hesitate to say to librarians what some of your challenges may be. I'm sure you know that better than I do, but we do hear from our library partners that actually it is pretty demanding for you all because you're supporting a wide array of research projects on your campuses and they may have really different mining requirements, so developing the expertise, hiring people with the right expertise to really be able to provide tailored support, is a new challenge. It's a hassle for you as well if there are different access requirements and permissions and things across different platforms, and something that Daniel and Darby really brought home to me is that your user privacy is really important to you. So what your users are doing across different platforms and how that is being monitored or tracked is a concern. Any additional cost at all is going to be a challenge, and we are also hearing from librarians that they're sometimes concerned about how TDM figures are being factored into the counter usage statistics, so they want to see publishers reporting separately on machine users and human users.

From a publisher's perspective, there are other publishers in the audience, for example, from Sage, so this is one perspective, but again we're having to support TDM projects across a wide range of subject areas—arts, sciences, and humanities; those have very different support requirements. We also are supporting users with a wide array of technical experience, so when they're mining they're actually often writing programs that are then being run on our platforms, and many miners are very sophisticated and they have terrific software—that's not always the case though. We do have examples where somebody lets their software loose and they haven't really tested it well and it kind of goes crazy, so this is one reason we do ask miners to tell us what their e-mail address is so that if their program kind of goes wild unintentionally we can get into quick contact with them and ask them to please stop running that code.

Legal challenges: we are a global publisher. Every law in every country and every state is different, and yet we need to have a simple easy-to-use service that is understandable that works across all of these different boundaries, national boundaries and institutional boundaries. So, in our case we have come up with one clause that we insert into all of our site license agreements that covers text and data mining from an institution's perspective. We agree that separate agreements aren't necessary, but we do need some kind of license in place because not every country has the same approach. And privacy for users is also important to us to maintain their respect and our credibility. We need to very much respect user privacy and we do.

Ok, and then, finally, we have to find ways of supporting miners, the machine users in a way that won't undermine the quality of service that we are providing to human users, and this is actually the reason we use an API service and let the miners go to town on that. It's a completely different platform actually than our human users are experiencing and interacting with. So, one doesn't affect the performance and speed of the other. We have a TDM policy. We don't make any additional charge for text and data mining. I know other publishers have other approaches, and I've got some sympathy for that. It is not inexpensive to develop the infrastructure and so forth for text and data mining, so if you have concerns about that from other publishers, you know, engage in dialogue with them. But I would urge you to be open-minded and sympathetic about the fact that they may be incurring new costs as well. And we also have made a firm commitment that if any researchers in academic institutions want to mine content that their institution doesn't actually subscribe to or have access to—that we'll respond and be really flexible there. So we have universalacess@elsevier.com, where you can make those sorts of inquiries if you need that support.

There is a registration portal and I've personally registered for an API key, and you guys are right—it is a pain in the ass! So, we have some work to do, I think, to get this streamlined a little bit more, and the idea is that we do want that user's e-mail address in case we need to get in contact with them about their code. We used to ask them questions about their subject area and their projects so that we could learn ourselves what they were requiring in different subject areas, but we want to step away from asking intrusive questions. They don't want to fill in forms for us, and we now have pilots so we've got that understanding. There's a simple click through license for our TDM service,s and I've shared it with Daniel so he can tell me afterward if it's simple enough or if we need to continue refining it, but one of the things I've discovered in preparing for this session is that the way it's been installed, there's actually another click-through license before you get to the TDM license, which is not intended. That's just an error in how it's been implemented, so again, on behalf of Elsevier and other publishers, if you or your users are interacting with our TDM services and you find stupid, awkward things like that, let us know. It's probably a mistake rather than a conspiracy or deep evil, honestly.

Okay, so we've rolled out the TDM access clauses into all of our site licenses, again at no additional cost. If anybody has been missed and wants them, just get in touch with your account manager or contact that universal access e-mail address. The big thing though is that we recognize that miners, that librarians don't only need access to Elsevier content. We honestly do get that, so we're partnering with other publishers through CrossRef on an industry-wide text and data mining service, and this essentially gives researchers a way to get API keys and permissions and to go to town across our platforms so they can access a wide variety of content. So, here are the publishers currently participating. We have about 13 1/2 million articles at this point in the service, and that can grow quite drastically. It is estimated that we are on a course for about 36 million by the middle of next year, and that service from CrossRef and the cross-platform mining is a completely free service as well. So, there's more to be done; there is more refinement; we want to work closely with librarians and we want to work closely with other publishers, and I'm looking forward to learning more about how we can do that more effectively. Thanks, everybody.

Darby Orcutt: Well, my thanks to Daniel and Alicia for really spelling out in a clear fashion the issues and the complexities of all this. I think what I'm going to do is simplify things. I'm a simple guy; I'm a "get it done" kind of guy, and my experience has been in the trenches of working with many, many vendors to try to cut to the heart and to make a deal to make this happen, and I think I have some solutions to some of these and at least one solution that we need to go ahead and move on with regard to this complex area.

A few years ago, or a couple of years ago, actually, this really came to my attention when I attended a campus colloquium at NC State on mining. This was sponsored by one of our social sciences units, and I was shocked to find hundreds, yes hundreds, of faculty and graduate students in attendance, and when one of the speakers asked, “How many of you are engaged in this activity?” almost every hand went up. I knew that there were a lot of folks working in these areas, but what I didn’t know was just how many were already doing this and how many were not using necessarily library resources to do this. Now, certainly we have some library resources that particularly social scientists know they can get datasets from, and that’s great but I also found in the course of this session that a lot of these folks were using whatever dataset they could find on the web—whatever was convenient, whatever seemed to be well structured data but not necessarily the best data to answer their particular research questions, and we in libraries particularly are worried about, well, if you’re like me you’re worried about the role of the library and research heading into the future. Here we are providing these wonderful resources, and are our users really actually taking advantage of them? We need to make it easier. We need to open up our collections for computer-assisted research. We need to open it up for these sorts of research informatics like mining.

I also, as I looked around the library community at the time, found a lot of folks who said, “Oh yeah, we’ve had that on our radar for a long time. Oh yeah, we have that as part of our license agreements, yeah, part of our checklists, we ask for mining rights or access rights.” And I had one colleague from another institution who said, “Oh yeah, we’ve negotiated those rights with a couple of dozen vendors.” I said, “Send me a list.” And they sent me a list, and probably the most interesting was, I’m making this up, but the *Journal of Esoteric Studies* and the three titles they published, they had mining rights to that; there were no major publishers on the list. None of the content that our users would really be wanting to do was opened up for this. I think really as I found in rolling up my sleeves and working with vendors on this, I found a “push me/pull me” on both sides of this equation. On the vendor side there’s this fear of letting the data out, and I think that it is sometimes a very well founded fear. There have been instances I’ve learned of, in confidence, where content has gotten out there and been posted on the web—usually it’s in China but, hey, and so there’s this fear of letting things out, and I think on the other hand a lot of vendors have a real fear of being last to the gate. They recognize that this is an area that they really need to respond to, but they’re really not sure how to do that. They’re afraid that their content will become devalued if they don’t offer it, but they’re really not sure how to do that. On the other hand, I think we have this “push me/pull me” of librarian confusion. For one, I think that I find on the library side a lot of misunderstanding of vendor capacities, and that’s a big problem. This is something again I learned from talking with vendors. One of these issues, for example, is the issue of siloed content. A lot of vendors are a little reticent to release their data because they’re a little embarrassed. What on the front end looks like a seamless product or a seamless set of products is actually all kinds of mishmash of different datasets, because these are the things that have been developed over time and they shouldn’t feel embarrassed about that. Certainly we in libraries understand that issue, but I think that is one of the things that is going on, and there’s this fear that maybe people won’t know what to do with that. They won’t how to grapple with this data that is not in a single structured format. But that’s okay. That’s what researchers do. I think also that librarians oftentimes expect an awful lot—and Alicia alluded to this—expect an awful lot for no additional cost, and that’s something that I have found—that’s actually confusion both on the library inside and oftentimes on the vendor side. Librarians think, “Hey, you can create an API or you can customize the dataset and that’s no big deal.” Well sometimes it is, and I know also that in talking with the vendors sometimes the vendors think that providing data is a very expensive proposition because their only experience with it has been with providing customized datasets that have required a lot of staff hours from very experienced, expert programmers and such.

I think also librarians grapple, and a colleague was mentioning this to me last night, addressing the present needs and addressing the future. One of the things that we should be doing, and libraries is thinking toward the future about, is thinking about what our users will need versus the sort of—now it's just the early adopters, as Alicia said. And so I think that really thinking about what are we doing when we put together agreements for mining, we're not building a bridge to nowhere. We're building the bridge that is across that great chasm where all of our researchers are headed toward, or a great many of them. And I think the biggest confusion all around on the part of both libraries and vendors is what do we do now? Where do we go? How do we take all this morass of confusing things and turn it into meaningful action? And I appreciate that when we talk about these things, talking about mining implies new services, it implies new support—it implies new roles; this is for libraries and for vendors, and I can't answer all of that and what that will look like, but I can tell you the first step.

I've been advocating, and this is what has allowed me to ink so many of these first of kind deals with a great many vendors, so I've been advocating for a very basic access model, and I realized just this morning I finally have my snappy name for it: the Basic Access Model, BAM. Let's move toward BAM, you know—hey, in the real world there's oftentimes a BAM that comes right before mining activities, right? Am I taking it too far? The Basic Access Model, well our idea of basic access will change over time but what does that mean right now? How do we get access to this particularly, and at this point it's our early adopters, it's our high-end researchers; first of all, we need to get all the data. We need to get all the files for those researchers. I've learned that speaking of raw data, that's not the terminology oftentimes to use in the vendor community. For some, actually creating raw data is an extra step for them so I say, "Well, what format do you have the data in? Send me those files." I want to include not just the text files but also the image files, and there are a couple of important reasons for that. First, depending on the researcher and what they want to do, those things can be extremely, extremely important. Secondly, as we look to the future, we talk now about text and data mining. I prefer to use the term "content mining" because there already are researchers who are or are thinking about mining images, mining video. Let's think about—let's have a generic term for talking about all of these formats that will become an important part of mining activities in the future.

Next, we need to have clear and appropriate cost recovery agreements. It's not fair to ask the vendors to do all this for free, but at the same time let's have a reasonable cost model for this. Now, when I worked out the agreement with Gale, the solution that we came to, a very inelegant one, but maybe the best one, was that the content we needed—they delivered all of the files to us on physical hard drives and they charged us the cost of those hard drives. That's fair. That was what we negotiated. Why? Because even with all the wonderful bandwidth that we have at NC State it would take a long, long time to download all of that data and, in fact, one of my colleagues from the library took the hard drives and walked them over to our laboratory for analytic sciences where one of our faculty, Paul Fife, got together a team and started to mine 19th century British newspapers—an inelegant solution. That's what is basic at this time. It will change. Bandwidth will increase; hey, we can make changes, but if he is right now, he and a team of researchers are working on this data. If we waited, and again, people ask me, "Why didn't Gale provide an API? Why didn't Gale do something better?" Gale should instead be applauded for going ahead and taking this step and becoming the first major commercial vendor of historical archives to offer this type of blanket access to a campus, and they've now turned around and opened that up to any of their customers. If we waited for them to get it all together it would've taken some time, and again that's not just Gale—that's any vendor right now. They're all working on these issues, but we need to make sure that we nail down that basic access for our researchers now.

I mentioned blanket access. That is another important key. We want to mirror the access that we have for close reading of this content, so we don't want to have an additional login. We don't want to have—as it's been a point of contention between me and one vendor, we came to agreement on everything except they want to know, they want every single person who's going to mine that content to register their name and their project with the vendor. Absolutely not! Would we sign a license for content for a database that had those terms? Absolutely not. There are a lot of good reasons why our researchers would need to keep their research private at least for a time. I had one researcher who was going up for an NSF grant, a very competitive grant, and he wasn't quite ready to reveal his great idea for what he was going to do. On the flip side, do probably 99% of our researchers want to report that to the vendor, want to be able to have these conversations about the data and what they're doing? Absolutely. But that needs to be their decision, not a requirement.

Lastly, we really need to have in this basic access model. We need to make sure that there are no special restrictions on mining activities or the outputs. And Daniel alluded to this; these things are already covered by our terms, they're covered by copyright, they're covered by the provisions of fair use. The irony is I've had a lot of vendors who say, "Oh, we want to limit the number of words that can be put out there, we want to limit the number of characters." No. We don't need to do that. We don't have to have a special agreement, and in fact the irony is that of folks who are mining, and this is particularly in text mining environments, their outputs generally cite less of the text than those people who were doing close readings because their outputs are quantitative, not to mention that many of these researchers are doing close reading and computer-assisted distant reading and putting their interpretations together. Well, where do you draw the line? Do they have to follow these restrictions because they had a mining element to their research, or can they follow these because they were also doing regular just using the database? Again, we don't need to complicate things. That's why I'm advocating for this: the Basic Access Model. BAM! Thank you.