September 2017

Liaisonship, the Law & Libraries: Supporting Content Mining Research

Darby Orcutt

NCSU Libraries, dcorcutt@ncsu.edu

Follow this and additional works at: <https://docs.lib.purdue.edu/atg>

Part of the [Library and Information Science Commons](#)

Recommended Citation

Orcutt, Darby (2017) "Liaisonship, the Law & Libraries: Supporting Content Mining Research," *Against the Grain*: Vol. 29: Iss. 4, Article 47.

DOI: <https://doi.org/10.7771/2380-176X.7821>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

Liaisonship, the Law, & Libraries: Supporting Content Mining Research

by **Darby Orcutt** (Assistant Head, Collections & Research Strategy, North Carolina State University Libraries) <dcorcutt@ncsu.edu>

Computational research is transforming the academic landscape, and computer-assisted mining activities are leading the charge. I have advocated strongly within the research library community for many years for attending now to what are both current and near-future needs of our research communities for a basic level of access to high-quality content for mining purposes.¹ I crafted the principles of BAM (Basic Access for Mining)² in order to create a shared understanding and pragmatic middle ground for libraries and information vendors alike to enable user access to library-provided content as data. I regularly speak to the importance of thinking in terms of “content mining” rather than simple text or data mining, to include present and future needs for image, audio, video, and other forms of information. I inked the first major agreements with commercial providers of digital historical resources to allow easy access for mining researchers to content within a field where I saw such access as a particular problem.³

Here, I’ll focus upon a facet of librarian support for content mining that deserves fuller attention: the relationship with the researcher. We librarians, particularly as liaisons to disciplinary communities, generally wear many hats. Often many, many hats. But even in a time when traditional liaison roles and activities are being reconsidered and realigned, this is happening with an eye towards making libraries more relevant to our users and increasingly central to the research lifecycle. We librarians are connectors, we are intermediaries, we are vital links between researchers and information. Connecting our users with content in computer-readable and -manipulable forms is simply an extension of our traditional responsibility, and an extension that is crucial to our continued relevance as a profession with the changing research and information landscape.

If research libraries don’t get on board in a big way with “content as data,” then we will be consigning ourselves to niche status within our user communities. While not necessarily news to those who have been paying attention, a recent headline in *The Chronicle of Higher Education* succinctly states how adaptation to the new reality is being accomplished by Elsevier — which is always smart (or sly, the preferred synonym for some librarians): “Elsevier Is Becoming a Data Company.”⁴ We need to similarly emphasize data in libraries far more and more smartly than we do currently.

Legal issues infect every aspect of content mining research support — but again, in ways that extend to the new frontiers and to the very support activities that the best libraries and librarians have already been providing. These include both proactive and reactive instruction, advising, and advocacy on issues of Fair Use, contractual law, and preferred practices within

areas often as yet unsettled with regard to case law and cultural consensus. Like many librarians, I could paraphrase the familiar catchphrase from countless television commercials of the 1980s: “I’m not a lawyer, but I play one as a liaison.”

Our users hold certain ideas about the legal contexts of mining activities. Some of these notions they need to disavowed of, some need to be refined and contextualized, and some need to be closely paid attention to, as they reflect needs, urgencies, and constructive paths forward for research. Especially as non-lawyers, we have the advantage of seeing the legal issues of content mining as just aspects of the context and constraints upon scholarly institutions and activities. While we of course wisely act within the law, we do not need to accept that present laws and practices are necessarily correct, “natural,” or firmly established, particularly with regard to new modes of research.

So, how do researchers perceive issues of accessing content for mining purposes? The specific answers certainly vary much from discipline to discipline, but except for researchers who are only working with the most clearly established, delineated, and discrete data sets, there are questions and perceptions that appear quite common. All of these user perceptions illustrate for librarians why we want to be part of the mining research workflow.

The question of “permission” to mine always arises, and researchers seem to cluster at two extremes of approaching this issue. On the one hand, some researchers blissfully assume that access of any kind equates to mining access. Until they encounter resistance, their de facto presumption is that anything is available and fair game, whether on the open Web, accessed via library-provided databases, or however. While we don’t want to break them of the ideal that ought to be, i.e., that “the right to read is the right to mine,” we do need to educate them regarding the nuances of technical barriers, terms of use, and contractual agreements that may constrain them (and may hinder access to all campus users of a particular resource when they trigger an IP block for excessive downloading).

On the other hand, many researchers presume that they must ask permission to mine any resource, even those that are open and not copyrighted. These are also users that we prefer consult with the library. The situation is often akin to that of an instructor informing a film vendor that they intend to show a film within the context of a course, and they may be incorrectly informed that they need to purchase “educational rights” or Public Performance Rights (PPR). Even in cases where no special rights or payments are needed, many vendors (through ignorance and/or greed) will insist that they are.

Most importantly, we should be creating a culture of practice around content mining where asking for permission is not a first step, but a step only taken when necessary. As the information brokers for our institutions, we librarians can take charge of this link in the chain of research — and our researchers will appreciate our doing so.

Issues of citation and data sharing often perplex new mining researchers as well, although most frequently they do not really consider these until the final stages of a project. Theoretically, the end results of most (arguably all) mining research are quantitative in nature, and therefore do not require sharing of the studied content at all (beyond perhaps for parenthetical or illustrative purposes that should generally fall well under Fair Use). Yet, I have seen vendors ask for mining agreements that limit citation using bright lines, and ones that are well below typical standards under Fair Use (in one case, a citation limit of 100 characters of text!). Again, researchers should be advised not to agree to artificial and unnecessary constraints, if at all possible.

We should be encouraged that many mining researchers want to share their data openly, even if pragmatically it is not always easy or even possible for them to do so. Certainly, it would be ideal if every mining project could share its data sets freely such that another researcher could replicate the study at hand. Yet, we must remind our researchers that this is an ideal. In reality, just as researchers frequently cite articles that are not freely available online to all readers, so too must it be acceptable to use data sets that are proprietary in nature. This is all the more reason, however, for both libraries and vendors to wherever possible adopt the principles of BAM, whereby proprietary data sets are made available for mining as broadly as possible at the institutional level rather than licensed to individual researchers, labs, or projects. Published research can describe the precise processes performed upon a particular set of proprietary content, including how data was selected, cleaned, and modified, and thus fulfill basic expectations of reproducibility. Yet again, the librarian’s role of intermediary, initiated at the outset of a mining project, would yield greater consistency and broader access for the research community.

Perhaps most importantly, we need to impress upon our faculty and other mining researchers that library mediation in obtaining access to content for mining assures the freedom of scholarly inquiry. At present, nearly all researcher requests for mining access are met with questions about the nature of the research project, often asking about funding sources, the precise searches that will be run against the content (as if mining research were

continued on page 27

not an iterative process!), where the results might be published, etc. While most of these questions are hopefully benign and likely stemming from interest in improving products and services, it is inappropriate to require they be answered ahead of granting access for mining purposes. They beg the question of what might happen if a company did not like a scholar's answers? Could they be denied access to content because of their research interests? By stepping into the middle ground, librarians help insure academic freedom. We are obtaining information access for our user community, and not interrogating them as to what they intend to do with it —aligning perfectly with our traditional roles as content brokers for our communities, paralleling the way that we traditionally purchased information in print format and circulated to any of our users without control or question as to the nature or scope of their research.

While I have focused almost exclusively above on proprietary data sets (and therefore the extension of the traditional library role as provider of published content to users), I do not want to ignore the extension of a newer but now well-established role of libraries as enablers and even publishers of content. We librarians consult on matters of copyright,

Fair Use, publication agreements, Open Access, and a host of other aspects of scholarly communication. We need to make sure that these conversations and our capacities extend as well into these areas as they relate to mining and data sets. As court rulings around Google Books have affirmed, there are certainly ways that transformative and openly shareable data sets can be produced under Fair Use from copyrighted, proprietary data sources. We should be engaging with our communities to facilitate the sharing of research data sets. We should be engaging with OA communities to ensure publication and hosting options for sets of data in all formats (not simply text and numbers, but images, audio, video, and more). We should be promoting and advocating the work and value of researcher-created data sets by encouraging consideration of their creation and sharing as a form of publication that should be appropriately valued as scholarly activity within our institutions and the disciplines.

In short, we need to strategically and fully extend the service of our profession into the research processes of content mining. This will require closer consideration of quantitative research, deeper understanding of its legal contexts, and stronger relationships with content miners, as well as a renewed sense of our mission and ability to add value across the research lifecycle. 🐼

Endnotes

1. **Darby Orcutt**, "Library Support for Text and Data Mining," *Online Searcher* 39: 3 (May/June 2015), pp. 27-30.
2. Originally "Basic Access Model," revised to "Basic Access for Mining." **Darby Orcutt**, "BAM: The Basic Access Model for Content Mining Agreements," *Proceedings of the Charleston Conference 2015*, pp. 155-157. <http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1718&context=charleston>
3. "NCSU Libraries opens pioneering new possibilities for data mining historical content," <http://www.infodocket.com/wp-content/uploads/2014/11/final-Gale-data-mining-press-release-1103142.pdf>; "Unlimited Priorities and NCSU Libraries Partner to Create Model Data Mining Agreement," <http://www.unlimitedpriorities.com/2015/03/unlimited-priorities-and-ncsu-libraries-partner-to-create-model-data-mining-agreement/>; "NCSU Libraries & Adam Matthew Digital Strike Ground-breaking Content Mining Agreement," *Southeastern Librarian* 63: 3 (Fall 2015), p. 12. <http://digitalcommons.kennesaw.edu/cgi/viewcontent.cgi?article=1581&context=seln>
4. **Paul Basken**, "Elsevier Is Becoming a Data Company. Should Universities Be Wary?" *Chronicle of Higher Education*, August 7, 2017. <http://www.chronicle.com/article/Elsevier-Is-Becoming-a-Data/240876>

Booklover — Theater

Column Editor: **Donna Jacobs** (Retired, Medical University of South Carolina, Charleston, SC 29425)
<donna.jacobs55@gmail.com>

It is late May/early June in Charleston and the Spoleto Festival is in full swing. The Festival hosts two weeks of opera, dance, a garden tour, theater, puppetry, physical theater, music, and artist talks capped with a Finale at the historic Middleton Place. This year the Galway theater company Druid has taken over the Dock Street Theater for the presentation of *Waiting for Godot* by **Samuel Beckett**. **Beckett** won the **1969 Nobel Prize in Literature** and was the subject of a previous **Booklover** column. But it is of note, in my quest to read one piece of work by every author to have won the **Nobel Literature Prize**, that this was a unique opportunity to experience an author's work presented in a format other than words on a page. A simple stage with a tree and a rock, five characters and a play about nothing — it was funny and riveting. Now it is time to read a work by another author.

Thomas Mann was presented the **1929 Nobel Prize in Literature** "principally

for his great novel, *Buddenbrooks*, which has won steadily increased recognition as one of the classic works of contemporary literature." It is unusual for the committee to reference a single work instead of the author's body of work. I have chosen instead to read a short story entitled *Little Lizzy*. It begins: "There are marriages which the imagination, even the most practiced literary one, cannot conceive. You must just accept them, as you do in the theater when you see the ancient and doddering married to the beautiful and gay, as the given premises on which the farce is mechanically built up."

This story unfolds about a couple in such a marriage. The charming, lovely and young wife, known as Amra has decided to organize a large party with entertainment. She has convinced her husband, "a perfect colossus of a man," to be the climax of this entertainment event. "Christian (the husband's name), suppose you come on at the end as a chanteuse,

in a red satin baby frock, and do a dance." Amra continues with her declaration that in addition to the dance he will perform a song. A song that Herr Alfred Lütner, her lover, will compose and provide the piano accompaniment.

"In a choked and gasping voice he sang, to the accompaniment of the piano. The lamentable figure exhaled more than ever a cold breath of anguish. It killed every light-hearted enjoyment and lay like an oppressive weight upon the assembled audience. Horror was in the depths of all these spellbound eyes, gazing at this pair at the piano and at that husband there. The monstrous, unspeakable scandal lasted five long minutes."

Thomas Mann was born in Germany in 1875. Writing was in his genes. His older brother was the author **Heinrich Mann** and three of his children became prominent German writers. He was initially designated to run his father's grain company. His father died when he was a young man and the company was liquidated releasing **Mann** from the business legacy.

continued on page 28