BAM: The Basic Access Model for Content Mining Agreements

Darby Orcutt
North Carolina State University Libraries, dcorcutt@ncsu.edu

BAM: The Basic Access Model for Content Mining Agreements

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

Abstract

The Basic Access Model (BAM) provides a reasonable and practical framework of business terms for libraries and vendors to agree on how to facilitate user access to digital content for content mining purposes, as well as a principled and agreed upon industry foundation for future cooperation. BAM has already opened up significant content for mining access. The sooner we can open up our collections—both as libraries and as vendors—to the new and emerging tools and methods of content mining researchers, the more relevant we and our collections will be.

Background

The need for much deeper library support for text and data mining (TDM) research came especially clear to me a couple of years ago at a campus colloquium on mining. Hundreds of faculty and graduate student researchers showed up—a far larger crowd than I had expected, and a crowd that was not simply interested in mining activities, but already engaged in them. Yet, I was simultaneously dismayed at the quality of the datasets that many of these researchers were using. Certainly many were using datasets from the ICPSR (Inter-university Consortium for Political and Social Research), federal government, and other standard sources, yet it seemed at least as many were simply downloading whatever datasets they could conveniently find via Google.

As a librarian, I knew we had better content available through the Libraries that could much better answer many of their research questions. But I also knew that we had thus far done little to expose to our users what data we had, and that most of it was not available in ways that would easily allow for computational research. Furthermore, this was not an institution-specific situation. Surely, I thought, research libraries could do better than we were. If we are serious about being relevant sources of information for researchers now and in the future—indeed, if we value the continued existence of the institution of the research library—then I realized we needed to be opening up our collections for computational research immediately.

Push Me/Pull Me

Librarians have been talking about establishing text and data mining agreements with vendors for quite a few years now, but little progress has been made. The current scene seems dominated by vendor fear and librarian confusion, but each of these conditions operating within a push me/pull me relationship to itself. Vendors, especially commercial vendors, simultaneously fear letting their data out into the world and not adapting well to the changing context of scholarly activity; in other words, they want to hold and protect the value of their content on the one hand, while also not missing the opportunity to increase the value of their content (or not allowing its value to diminish) by withholding it from the new and emerging ways in which academics now and increasingly will need to use it.

Librarians themselves largely seem to feel torn in different directions too, confused mainly by pragmatic versus idealistic desires. At the institutional level, they generally privilege present needs of access over what they perceive as future needs of supporting mining activities (resulting, for one, in not pushing too hard for TDM agreements). At the larger professional level, however, librarians as a whole seem rather more idealistically than pragmatically focused. They misunderstand the current or near capacities of vendors, including how silo-ized is much content that may appear homogenous via a front-end interface, how historicized and uneven is much metadata, and especially how costly it might be to

http://dx.doi.org/10.5703/1288284316247

provide the "perfect" interface or even APIs, as well as quibbling amongst themselves as to exactly what researchers will need. The result is near-deadlock, with continued non-access for computational researchers.

Breaking Through the Logjam

The big confusion all around centers on what to do and where to go from here. For both vendors and libraries, the move toward computational research implies new services, new means of support, and new roles. I recently began exploring the possibilities for these on the library side in an article in *Online Searcher* (Orcutt, 2015). While I can't answer all of these questions—and moreover, I believe that there are multiple acceptable and contextually appropriate decisions around this support that will need to be made at the local—I am certain that our first step needs to be nailing down agreement on a basic level of access for research.

This is what I've been doing for the North Carolina State University (NCSU) Libraries, but also with a hand toward getting commercial vendors to offer like terms to their other library customers. Beginning with Gale in June 2014, we have inked several first-ever blanket mining agreements with commercial providers of especially primary source historical content. And, as Gale did just a few months later in November 2014, many of these vendors have (or will soon) be offering these same terms to other institutions. Several press releases (with more to come) detail some of these successes:

Gale:

http://www.infodocket.com/wp-content/uploads/2014/11/final-Gale-data-mining-press-release-1103142.pdf

Unlimited Priorities/Accessible Archives:

http://www.unlimitedpriorities.com/2015/03/unlimited-priorities-and-ncsu-libraries-partner-to-create-model-data-mining-agreement/

Adam Matthew:

http://news.lib.ncsu.edu/blog/2015/08/14/ncsu-libraries-adam-matthew-digital-strike-groundbreaking-content-mining-agreement/

The Basic Access Model (BAM)

The key to opening up these rich datasets to computational research is to focus on a core level of access for mining purposes, with both library and vendor meeting at a reasonable, practical, and sustainable agreement regarding how to make this content available to researchers now, without the need to define and articulate the complete future landscape of what vendor and library roles, services, and support for mining activities may become. My Basic Access Model (BAM) is not a model license, but rather a model framework of business terms to ensure researchers can access these often superior (although never perfect) datasets. And while our shared understanding of what constitutes "basic" may change a bit over time, this is what it includes now:

1. Library customers can access *all* data—This includes all associated metadata and image files. In fact, I much prefer to speak of "content mining" than TDM, as the former is more inclusive of the sorts of formats that only the most advanced researchers are currently mining but will eventually become much more commonly mined.
2. Clear and appropriate cost recovery agreements—These should spell out in advance actual and reasonable costs of such recovery. Too many institutions have found that they and vendors differ over what constitute the bases of "cost recovery," so agreement in principle alone is not sufficient. Vendors would do well to publish clear and reasonable cost schedules for particular datasets. Providing basic access for content mining should not be a profit center for vendors,

although I do see opportunity for vendors to develop at scale and monetize add-on services (particularly teaching resources for mining activities) (Orcutt, 2015). Although it may sound archaic, the easiest and least expensive delivery option for very large datasets at present is actually physical delivery of hard drives (something that will certainly change over time!).

3. Access for researchers must be blanket access—In other words, access to data for mining (or distant reading) purposes should mirror usual individual (or close reading) access, allowing for anonymous access so long as institutional affiliation is confirmed, and without need to indicate what, how, or why research is being conducted (as some vendors desire). Libraries have vested interests in protecting their users' privacy, as well as ensuring against potential overt or *de facto* censorship of a line of research. And while the overwhelming majority of content mining researchers welcome conversation and transparency with vendors about their research questions, there are legitimate reasons why a researcher may wish to delay or forgo divulging their ideas (for example, a pending grant application).
4. Lastly, there should be no special restrictions on mining activities or their products—These uses by scholars are already covered by existing contract terms, copyright, and fair use. Access to data for content mining does not give a researcher any special rights to share or publish, but neither should it remove any usual abilities to do so. Motivated by fear, some vendors attempt to more strictly limit, for example, textual citations by mining researchers (to a certain number of words or characters); ironically, the products of text mining research are often more quantitative or algorithmic in nature, and are likely to cite less (and less of the essence) of texts than more traditional scholarship.

Next Steps

Libraries and vendors alike should look to the BAM model as both a present means for content mining access to data, and as a principled and agreed upon foundation for future cooperation between content providers and libraries. Certainly tools, support, instruction, services, staffing, infrastructure, and needs related to content mining will give us plenty to discuss, develop, and figure out over the conceivable future. Yet, the sooner we can open up our collections—both as libraries and as vendors—to the new and emerging tools and methods of computational researchers, the more relevant we and our collections will be to academic research, scholarship, and the discovery and production of new knowledge.

References

Orcutt, D. (2015). Library support for text and data mining. *Online Searcher, 39*(3), 27–30.