IFLA WLIC 2013
SINGAPORE

Submitted on: August 7, 2013

Text & Data Mining - A Librarian Overview

Ann Okerson
Senior Advisor, Center for Research Libraries, Chicago IL, USA
aokerson@gmail.com

Copyright © 2013 by **Ann Okerson**. This work is made available under the terms of the Creative Commons Attribution 3.0 Unported License: http://creativecommons.org/licenses/by/3.0/

Abstract:

Text and data mining offers exciting research opportunities over a broad range of fields. As large corpora of data accumulate, automated and semi-automated analysis of their contents (and often of many different data sets correlated together) reveals patterns and allows establishment of fact patterns invisible to the naked eye. Libraries and librarians have an exciting opportunity to support this work.

This paper reviews some of the possibilities for such work and outlines the challenges and the way ahead for librarians. One challenge lies in the terms by which data sets are licensed and made available to academic and other users; librarians need to be proactive in ensuring that these terms are favorable for the kind of use researchers will need and that the resources themselves are available in a format that allows innovative mining-based research. Another challenge is the need to support users who wish to engage in text and data mining with limited experience, especially when they approach data sets made available through library resources. Librarians should develop the expertise to support their users by making data resources available to them on favorable terms and supporting their mining efforts .

Keywords: TDM, Data Mining, Text Mining, Center for Research Libraries, Licensing

My goal here is to provide a quick librarian overview of text and data mining, or TDM as we will now call it. The Center for Research Libraries has been instrumental in raising awareness of this emerging field of scholarly and research endeavor among its 270 library members, plus partners, in the US, Canada, UK, Hong Kong, and other areas, through a recent set of meetings and webinars. At the outset, many thanks to Alicia Wise (Director of Elsevier's Universal Access program and a member of this standing committee) and a number of publishing colleagues for their thoughtful insights and documents.

That said, just what is TDM and why are we talking about it more and more these days? Our time is short, so we will glance at that question; provide a few examples of such work; note some challenges for those who would do TDM; and raise, quickly, some issues and opportunities for librarians who work in communities where TDM is likely to become of interest or already is.

WHAT IS TDM?

Definitions abound. Bernie Reilly, CRL's President, speaks of it as "automated processing of large amounts of structured digital textual content, for purposes of information retrieval, extraction, interpretation, and analysis."[1] He distinguishes it from data mining, which extracts and analyses data, rather than text, from chosen sources. But for our overview purposes here, and since many of us support sciences or social sciences data researchers, it makes sense to include data in our remit.

In his 2013 report for the Publishing Research Consortium, independent consultant Jonathan Clark also separates the two types of activity. He notes that there is no universally accepted TDM definition, as "it is used by different communities for different purposes." He describes text mining as an sophisticated, smart type of indexing, which aims "to extracts the meaning of a passage of text and to store it as a database of facts about the content and not simply as a list of words." He defines data mining as "an analytical process that looks for trends and patterns in data sets that reveal new insights. implicit, previously unknown, and potentially useful."[2] In her benchmark 2011 report, Eefke Smit refers to TDM as "automated tools, techniques or technology to process large volumes of digital content that is often not well structured -- to identify and select relevant information; to extract information from the content, to identify relationships within/between/across documents and incidents or events for meta-analysis."[3]

WHY TDM?

It's time to chant the Ecclesiastes: "There is nothing new under the sun." Don Swanson (1924-2012), an American Information scientist, is known for his work in literature-based discovery in the biomedical domain. His method came to be called "Swanson linking," which involves connecting two pieces of information that were previously thought to be unlinked. In fact, some 25 years ago, I was privileged to be at his dinner table and listen to his story of the excitement he felt when the *Index to Scientific Information* (ISI) became available and launched the research that he continued for the rest of his life. Passionately, he would say things like, "if only AAA (biologist) could have known about the articles by BBB (chemist), biochemistry would have made CCC discovery 20 years earlier and healthcare would have made a giant leap." Swanson called these separate pieces of knowledge "disjoint data" and "undiscovered public knowledge."[4]

The next generation of miners began to extract and process data on a larger scale. The Perseus Project,[5] which started in the 80s, automatically analyses grammar and vocabulary of ancient Greek texts and makes predictions about the function of a specific word in a specific passage. If a reader is not sure about a specific word and clicks on it, a parallel window opens and shows dictionary entries with similar spellings, gives statistical information meanings, and inks to further information.

[1] Bernard F. Reilly, "When Machines do Research, Part 2: Text-Mining and Libraries," *Charleston Advisor*, October 2012: 75-76.

[2] Jonathan Clark, *Text Mining and Scholarly Publishing*, a report for the Publishing Research Consortium. Loosdrecht, The Netherlands & London: 2013: pp. 5-6.

[3] Eefke Smit and Maurits van der Graaf, "Journal Article Mining, a research study into practices, policies, plans.....and promises. Commissioned by the Publishing Research Consortium, Amsterdam, May 20, 2011. http://www.publishingresearch.net/documents/PRCSmitJAMreport20June2011VersionofRecord.pdf

[4] For Don Swanson' short bio, see: http://en.wikipedia.org/wiki/Don_R._Swanson

[5] The Perseus Project is found here: http://www.perseus.tufts.edu/hopper/about

The government sector, as well as the commercial sectors (business, finance, media, scientific industry), have been quick to embrace and deploy TDM to advance their goals. And today, students and scholars have access to huge amounts of electronic information, both data and textual - statistics about availability and growth abound. The availability of these sources has, in turn, spurred interest in academia among scientists, social scientists, and humanists. Clark notes four main reasons to engage in TDM:[6]

1. To enrich content -- Mining can improve indexing, be deployed to create relevant links, to improve the reading experience.

2. Systematic review of literature -- Mining can help researchers systematically review larger bodies of content, faster than they could do it themselves and to keep up with their field, without missing relevant information.

3. Discovery -- Mining can be used to create databases that can themselves be mined.

4. Computational Linguistics research -- Mining itself is the subject of research, for example to improve the extraction of meaning from texts.

For example, in the CRL November 2012 Text Mining webinar,[7] participants had the opportunity to hear about works such as:

> * Matthew Jockers, whose focus is on computational text analysis, specifically an approach he calls “macroanalysis,” to advance his scholarship is in Irish and Irish-American literature of the late 19th- and early 20th-century.
>
> * Robert Nelson -- "Mining the *Dispatch* seeks to explore and encourage exploration of the dramatic and often traumatic changes in the social and political life of Civil War Richmond. It uses as its evidence nearly the full run of the Richmond *Daily Dispatch* from the eve of Lincoln's election in November 1860 to the evacuation of the city in April 1865."[8]
>
> * The Corpus at the HathiTrust Research Center, which enables computational access for nonprofit and educational users to published works in the public domain and, in the future, on limited terms to works in-copyright from the HathiTrust.[9]
>
> * The IMPACT project, funded by the European Commission, to which Gale has contributed historical content. IMPACT aims to significantly improve access to historical text and to take away the barriers that stand in the way of the mass digitization of the European cultural heritage.[10]
>
> * Max Haeussler's UC Santa Cruz project to extract DNA sequences from millions of papers, to facilitate bio-curation and discovery. This required identification of "all

[6] Jonathan Clark, *op. cit.*: p. 7.

[7] The slides from this CRL webinar (Text Mining Opportunities and Challenges) are found at: http://www.crl.edu/events/8391

[8] To experiment with this site, see: http://dsl.richmond.edu/dispatch/

[9] For information, see: http://www.hathitrust.org/htrc

[10] IMPACT is described at: http://www.impact-project.eu/about-the-project/concept/

> words in full-text articles resembling DNA sequences, which are extracted and then mapped to public genome sequences."[11]

CHALLENGES FOR OUR TDM USERS

TDM requires that researchers be able to process large amounts of content in an automated way. The up front effort is in identifying the questions that should be asked (a complex process), finding the sources that should be mined, accessing those, downloading them to a local host, programming to ask the questions, and in the end analyzing and interpreting the results.

While these steps can sound straightforward, currently researchers may encounter a number of obstacles to launching, let alone completing their research projects. These can involve gaining access to materials (particularly when an institution has not paid for subscription access). If one has access, how does one get the assurance that mining will be allowed and facilitated, given that many publishers regard this as wholesale downloading? What can the researcher do with the output? Publishers do not necessarily use standard formats, so cross-publisher mining is not assured even if permissions exist -- but by its nature, TDM can get some of its most exciting results by bringing together very disparate and different bodies of data. There can be huge amounts of formatting and preparatory work for most text miners. Sending, retrieving, and storing vast amounts of content are not yet easy for most researchers. And then what happens to the mined knowledge when the research is completed? Can we standardize formats, hope for one or few mining platforms, standardized license terms?

Publishers report that demand is still limited, and this provides, I believe, some breathing room for the sectors to come together, to resolve challenges in TDM, demand for which will surely grow rapidly. Some publishers have begun to respond by working with researchers who ask them for permission, and helping them along. Others allow downloading or may provide an API to mine content. Gale, for example, makes certain raw files available to scholars for download and mining. Others are creating new tools to be built into their publishing platforms. I'm hopeful that our final panelist, David Tempest (Elsevier) will walk us through a number of publisher initiatives aimed at supporting researchers in the TDM space.

ISSUES AND OPPORTUNITIES FOR LIBRARIANS

I suggest that we have two main areas of activity with respect to meeting TDM needs in our research institutions.

I. **License language.** A growing number of librarians have been active in developing language for license/contract language and permissions, so essential for TDM. We, as well as publishers, are coming to a rapidly opening new world of research possibilities, and many of us find that our electronic resources licenses do not speak to this topic at all. Much can be said here, and time is short, so I will point you to an excellent overview of institutions active in this area, outlined in a November 2012 presentation by Teresa Lee, University of British Columbia,[12] and suggest that at this particular moment, the license language from the

[11] See PPT presentation by Judson Dunham (Senior Product Manager, Elsevier) at the CRL webinar.

[12] Teresa Lee, "Text Mining from Three Perspectives: an e-resource librarian's view," in the *Charleston Conference on Acquisitions and Collection Development*, November 2012. See: https://www.dropbox.com/s/k0r8qnvt95ifiyi/Text_Mining_TLee.pdf

California Digital Library and also of JISC (UK) are worth a second and third look -- these organizations have invested a great deal of time and thought in "getting it right." As this library work progresses, we can hope that various licensing checklists and model licenses will start to present language that all the rest of us can learn from and improve. In Fall 2013, the Center for Research Libraries hopes to launch a community process that will undertake to completely revise the original 1999 CLIR/DLF Model License that has been widely adapted by numerous libraries and consortia in the US.[13] TDM sits at the top of our list of new contract elements.

Publishers have also been developing TDM license language, with two notable results: the "PDR pharma" license, which aims at supporting the needs of the commercial, pharmaceutical research industry,[14] and the STM Association's model Clause, which builds on the PDR agreement.[15] While librarians may agree that these documents are not entirely relevant for non-profit, academic purposes, we respect the efforts to date.

II. **Supporting our Researchers.** Reilly observes: "Building the right kinds of conditions and terms for computer-assisted processing and analysis of commercial database content is going to be difficult without a clear sense of the practices in this kind of research activity. Librarians can play a critical role in this process but only if they fully understand the practices of their constituents and integrate that understanding into their licensing and resource development work."[16]

Absolutely. While many issues to be resolved are purely technical, a number are "customer service" issues. The person who wants to do data mining without becoming a tekkie miner will have trouble getting the best results out of even any pre-packaged service. So, there will be demand not just for mining, but also for steady help in getting to the point where mining can be most effective. These seem to me to be librarian opportunities. Knowing our communities, helping them understand what data mining can do for them, and connecting them to the tools and resources they need: that's surely our job.

And, right now, TDM gives us an opportunity in some cases to "get ahead" of our scholars and students, to feature talks, web information, instruction or orientation sessions about TDM and the new doors that it opens for study and discoveries. Librarians are excellent at reaching out and organizing such activities.

There's also an important role for the data-mining-savvy non-technologist who can sit with the end user and interview him or her; then propose a strategy; then work with that user on a pilot mining exercise -- perhaps help to run the queries against 1 percent or 5 percent of the whole set and see if the results look promising. If they're not quite right, then the librarian and the user can talk some more about what they're looking for and what seems possible, revise the queries and try again until the full-scale project can be mounted. Doesn't this sound similar to what reference librarians already do? Especially those we've already appointed to be data or GIS or digital humanities librarians?

[13] See the *LIBLICENSE* Model License link, for this and other model licenses around the world: http://liblicense.crl.edu/licensing-information/model-license/

[14] See the Release and language at: http://www.stm-assoc.org/2012_09_12_PDR_ALPSP_STM_Text_Mining_Press_Release.pdf

[15] For the STM language and commentary, see: http://www.stm-assoc.org/text-and-data-mining-stm-statement-sample-licence/

[16] Reilly, *op. cit.*, p. 76.

Within the research community, expertise in knowing what's possible *and* in helping novice users take advantage of it has been very much a part of our librarian profession for a long time. To transfer it to the world of TDM will be hard work, but it looks like a natural to me.

IN SUMMARY

The reported numbers of requests for data mining are extremely small. My own instincts say there's something incomplete with how we're collecting those numbers, that explicit demand is quite a bit higher, and potential demand is VERY much higher. I surmise that there are interested researchers who are simply doing their own work in ways that don't get on the radar. Let me just say again that in my view demand will rise much higher very soon.

BTW, I should add that librarians do not want to see a future where researchers (and libraries) must depend on costly publisher tools and services, in addition to the large sums we are already paying for e-resources, and that this could perhaps be part of contract language.

Another comment along these lines: Open Access content has, I believe, a real advantage with respect to TDM -- at least with respect to the thorny issues of seeking permissions across numerous sources and databases. This hassle-free aspect is something that the current efforts of publishers cannot easily replicate.

Meanwhile:

- Libraries can become more aware of campus needs and offer support/expertise
- Libraries can encourage publishers in licensing and researcher support and offer to co-develop some license principles and services
- Collaboration is required across publishing, libraries, the research community
- Librarians can participate in facilitating such activities

And who knows what we will someday be able to learn that we do not now suspect, even from this paper presented at the IFLA World Congress!

**Thanks to Bernard Reilly, President of the Center for Research Libraries, for his advice on this paper, as well as the CRL's leadership in this new area of research and scholarship.*