**Text-Mining and Libraries: Summary of a Conversation with Publishers**

Bernard F. Reilly, Jr.

*As noted in the CRL report in the October 2012 issue of* The Charleston Advisor: *“The growing application of text mining techniques and technologies in many fields of research has implications that are beginning to be felt by libraries.”[[1]](#endnote-2)*

Text mining is generally defined as the automated processing of large amounts of structured digital textural content, for purposes of information retrieval, extraction, interpretation, and analysis. This practice usually entails one of two scenarios: the downloading of large amounts of digital text to local media environments, where special tools are employed to manipulate, annotate, and process the text; or the use of special applications to act upon the content within the publisher or host system and generate new information or works from this content.

Modern scientists and scholars now employ proprietary and open-source software and tools to process and make sense of the oceans of content at their disposal, and the application of text-mining techniques and technologies is growing in many fields of research. Recognizing the implications of this activity for academic libraries, the Center for Research Libraries recently held a webinar on “Text Mining: Opportunities and Challenges.” The webinar was part of CRL’s [Global Resources Forum](http://www.crl.edu/forum), a set of activities, events, and resources that support informed, strategic decision-making on library investment in digital collections and services.

To date, much text-mining activity has focused on large open access corpora of digital text, such as Google Books, PubMed Central, even WikiLeaks. But interest in mining proprietary publisher content is growing. The CRL webinar featured experts from publishing and the academic library world, who discussed recent trends in text mining, how publishers and libraries are responding to those challenges, and what new services are envisioned or in the pipeline.

* David Magier, Associate University Librarian for Collection Development at Princeton University, spoke about “Recent Trends in Text Mining and Library Services: the Research Library Perspective.” Magier surveyed a number of open-access text-mining projects, and discussed what role libraries can play to support these activities. He stressed the need for librarians to know more about the types of research being done.
* Judson Dunham, Senior Product Manager, Elsevier, discussed recent efforts to mine the scientific literature from the perspective of the journal publisher. He described new research using computer-assisted processing and analysis of text and metadata in the journal content published and hosted by Elsevier, primarily in the fields of biomedical science and chemistry. Dunham outlined the challenges and larger issues these practices pose for database publishers.
* Ray Abruzzi, Director of Strategic Planning, Learning and Research Solutions, Cengage Learning, described Gale’s recent experience with researchers’uses of Gale databases such as Eighteenth Century Collections Online (ECCO). He related how Gale made ECCO page images, OCR text, and metadata available for mining by scholarly projects like NINES and 18thConnect. In return for their cooperation, Gale expects to obtain corrected OCR and annotations of the content that might then be incorporated to improve the database.
* Ann Okerson, Senior Advisor on Electronic Strategies, Center for Research Libraries, was commentator for the session. She pointed to the long history of text-mining in linguistics and related disciplines, and the growing number of users in government intelligence, business, social media, news, finance, and today’s legal arena. Okerson identified issues that “text mining in the Cloud” raises for libraries in obtaining electronic access to digital resources, and discussed how those issues are now surfacing in policy discussions at the national level, in the UK in particular.

From the standpoint of the presenters, the principal issues connected with supporting text mining are twofold: technical and legal.

*Technical Challenges*

Supporting text-mining projects to date has involved direct interaction between publishers and the scholars and scholarly communities. The publishers’ experiences suggest that this approach will not scale to accommodate future demand. Meeting that demand will require publishers to offer a different kind of database product: as Gale’s Ray Abruzzi put it, a product that is “machine-facing” rather than “human-facing.” One webinar attendee asked whether publishers will pass the costs of the required tools and special services developed on to the libraries in the form of higher database costs.

Standards and standard formats for content output, such as stripped-down, structured text, could facilitate the new uses, as research topics often span materials on multiple publisher platforms. However, at this stage it is difficult to create such standards, given the lack of uniformity among the many text-mining tools being deployed by researchers. Publishers will probably have to continue to accommodate myriad research practices.

It was also noted that intensive mining of online databases and journal content can place significant stress on a publisher’s network resources. Judson Dunham noted that “storing and redistributing . . . text-mining results will require long-term infrastructure, hosting and other services” that are yet to be built. “A service offering a reliable, persistent, managed, open venue for the hosting of content-mining results” he suggested, “would be valuable.”

*Legal Issues*

Copyright law is not at all clear in its application to the mining and computer-assisted analysis of copyrighted content. For that reason, publishers have been conservative in granting users rights for such uses. The rights obtained by publishers themselves from the sources or rights holders of the content they publish are often limited and do not explicitly address the necessary processing applications. (In some cases, these sources include libraries.) Once obtained by the publisher, those rights would need to be re-granted to users through license agreements. It was noted that the right to mine digital texts, even when not withheld, can be rendered moot by barriers embedded in the publisher’s interface and platform technology.

Conclusions

The Elsevier and Gale presentations about the projects supported were quite informative, but presenters agreed on the need for a better understanding of the text-mining now being done. David Magier and Ann Okerson both noted the importance of librarians becoming more aware of local projects and of offering support and expertise to facilitate those efforts. This would better enable librarians to determine the types of new rights that need to be addressed in license agreements.

In her closing remarks, Ann Okerson noted that the conversation on this topic between the community and publishers is extremely important at this stage. The potential exists to identify common interests through such dialogue, and collaboration will be required across publishing, libraries, and the research community. CRL will keep this conversation alive.

1. Bernard F. Reilly, Jr., “CRL Reports. When Machines Do Research, Part 2: Text-Mining and Libraries”, *The Charleston Advisor* October, 2012 14 (2) http://charleston.publisher.ingentaconnect.com/content/charleston/chadv/2012/00000014/00000002/art00022 [↑](#endnote-ref-2)