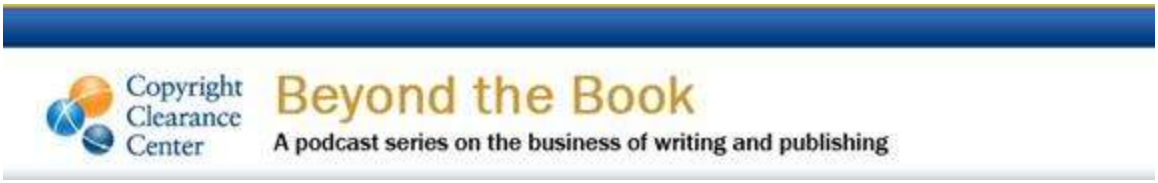

BTB #251: Workin' in the Content Mine

A report from the Publishing Research Consortium notes that content mining is increasingly of value to a broad group of users. Amsterdam-based researchers who prepared the PRC report Lead CCC's **Chris Kenneally** into the content mine

- **Eefke Smit** is the Director of Standards and Technology for International Association of Scientific, Technical & Medical Publishers, and she coordinates the activities of the STM Future Lab Committee.

- **Maurits Van Der Graaf**, is the owner of Pleiade Management & Consultancy, with expertise in publishing, public education, and library and documentation institutes.

For podcast release
Monday, August 29, 2011

KENNEALLY: Content mining is shorthand for automated information extraction and relationship analysis. A report from the Publishing Research Consortium notes that content mining is increasingly of value to a broad group of users from information scientists who use it for sophisticated indexing and clustering of related documents to drug discovery researchers who use mining techniques to discover new disease treatments or new application areas for existing medicines, and even among marketers who use text analysis for sentiment analysis of large crowds by measuring positive and negative expressions around topics as they occur on Facebook or Twitter.

Hello and welcome, everyone, to *Beyond the Book*, a podcast series on the business of writing and publishing from the nonprofit Copyright Clearance Center. My name is Christopher Kenneally.

Leading us into the content mine today are two Amsterdam-based researchers who prepared the report. Eefke Smit is director of standards and technology for the International Association of Scientific, Technical and Medical Publishers and she coordinates the activities of the STM Future Lab Committee. In addition to her STM work, Eefke works as an independent consultant in new business development, e-publishing and innovations. Welcome to *Beyond the Book*, Eefke Smit.

SMIT: Thank you, Chris, for inviting me. This is really an honor. Thank you.

KENNEALLY: Well, we're very glad to have you. And also joining us from Amsterdam today, Maurits van der Graaf is the owner of Pleiade Management and Consulting. He has held various management positions in publishing, in library and documentation institutes and in public education. Welcome also to *Beyond the Book*, Maurits van der Graaf.

VAN DER GRAAF: Glad to be on the show, Chris. Thank you.

KENNEALLY: It's a pleasure to have you both join us today on the line from Amsterdam, a report that has global implications and I suppose a place to start – and I try to give people at the opening there a very high-level notion of what content mining is all about, but tell us – Eefke, we'll start with you – why is content mining tipped to be increasingly important in publishing and research?

SMIT: That is a very good point, Chris, because in the 29 interviews that we did for this study, followed up by a survey that yielded responses from around 190 publishers, we posed a similar question. And content mining is a technique or a technology that knows optimists and pessimists and it is actually the group of optimists that seem to get the floor nowadays, and probably rightly so, because a few things are now falling together that can give content mining a really big boost.

As you said, content mining is a technology to go through vast amounts of content in an automated way, and people do that to extract information from it or to do deeper analysis. And in the past 10 years, a vast digital corpus has become available and it's growing all the time. So simply speaking, there is a lot to mine through that wasn't there 10 years ago. So that is one very simple aspect that plays a role.

Secondly, the tools are improving really fast. There's much more software, it's much easier to handle, and also, the computer capacity that you need to go through vast digital corpuses is less of a problem nowadays.

KENNEALLY: What are the areas that attract the most interest as far as content mining goes? And are we seeing a shift at all from one group of study to another?

SMIT: Very much so, because originally, this has very much been an area that was applied, for example, in drugs research, and in drugs research, there's even a clear logic why people have been doing it. Drugs research is a very expensive process and you need a lot of information. People working on new drugs, they cannot afford to miss a side effect that was in the literature somewhere. The consequences for a pharma company can be enormous about that.

So, in the pharma industry, people have always been fine-combing the literature to find any and all kind of relationships that you can have there. But also, since drugs research is trying to be more and more efficient all the time, for example, they use content mining to find new applications for existing drugs. For example, if you make good analysis of all the side effects, then some tracks of side effects can actually bring you into new areas where the drug can also be applied.

But aside from those really research and high scientific areas, there have been more and more application areas in the world around us. For example, in business administration, a lot of content mining takes place for competitive analysis. You can go through patent requests to see what other companies are working on. There's talk that in the legal area

there's much more content mining now. The *New York Times* had an article a few months ago under the tempting headline, Expensive Lawyers Being Replaced by Cheap Software, and the cheap software can go through all the case law.

And also in social areas, as you've just mentioned, people use content mining to do so-called sentiment analysis on Facebook and for example, here in Holland, and if it exists here, it will surely exist in the U.S. as well. In our newspapers you can see which words or which topics were tweeted most during – on a certain day so that you can see what people are involved in and what they talk about between each other.

And those are all very simple applications for content mining.

KENNEALLY: It seems to be just the right moment to prepare the kind of report that you've been working on, and Maurits van der Graaf, let me turn to you and ask you to tell us about how the report was prepared, what kind of research you did, who you spoke to, who you surveyed and just generally, more about the expressions of interest they had or that were revealed through your research for content mining.

VAN DER GRAAF: Eefke and me did the work, so to say. The study was set up in two phases. We did a qualitative study with 29 interviews with experts in academia, research libraries, (inaudible) publishers and those interviews were carried out by Eefke. And the next step, the second part of the study, we made an elaborate questionnaire together, of course, based on those interviews and that was sent – and I was responsible for the survey – to many people in the publishing world and we tried to get as many respondents as possible.

KENNEALLY: And give us a general sense. Eefke was saying it sort of divides between great excitement and sort of – I don't know – perhaps the opposite of that. But would you have a sense right now of a mood among publishers as far as the potential for their business that content mining offers?

VAN DER GRAAF: Eefke said earlier that she encountered quite a number of skeptics in the interviews. But we asked them – we translated the optimism and pessimism in a number of statements in the questionnaire and asked if people agreed or not agreed with it. And from those statements, you can see that the publishers generally are rather optimistic about the next steps in content mining and they see it's increasing and growing in a number of applications and in significance.

KENNEALLY: Well, obviously for it to grow as a business, the publishers have to be open to it in the first place and the requests for mining come from a variety of sources. Can you tell us about that? Where do the third-party requests originate and how common are such content mining requests today?

VAN DER GRAAF: Well, that was a rather surprising result from the study, I think. The third parties are corporate customers, for instance, the pharmaceutical companies, as Eefke told

you earlier. But there are research groups in academia that are concerned with text mining or with open access. And you have the publishers themselves, secondary publishers, abstract and indexing services, that are also important, responsible for demands to all the publishers.

But the frequency of the number of mining requests that are received by the publishers was rather low, in our view. We found that 77% of the publishers from the survey reported such requests, but only 21% reported more than 10 requests per year, so it seems to be quite widespread, but frequency is rather limited.

KENNEALLY: Well, we're probably still at a very early stage in all of this. And Eefke, if I can ask you about publishers' responses to such requests. Are they happy to grant them? Under what circumstances do they find a request more interesting than others? Tell us how they react.

SMIT: There's an interesting dividing line between the publishers, because there's a first percentage between 20 and 30% of publishers who don't even require permission. They grant any mining re – well, they don't even grant it. Mining is allowed in their content. And a large part of them are open-access publishers and the other ones are usually small publishers. But that needs the extra comment that we also know that the smaller the publisher is, the lesser requests they get anyway. So they allow it, but they would also never get a lot of requests.

If we look at the other publishers, then a very high percentage, even over 95%, wants information about the intent and the purpose of the mining, and as soon as they have that, a good share of them will generally grant permission. Up to 90% will grant permission in 100%, in the majority of the cases or in some cases. And roughly 35% will do so in the majority or in 100% of the cases. So generally, they're quite open to that. This percentage is even higher if the requests are made for a real research purpose.

And there's only one criteria that we found in the whole list where publishers are far more reluctant, and 90% tends to decline mining requests if it's for a navigational product that could compete or replace their own content.

So in general, publishers are fine with content mining if it doesn't substitute their own services, so to say.

KENNEALLY: Right. So a question for Maurits, then. Does your research show us any potential for a business model in content mining as far as publishers consider it, or is this just a way to help serve their customers and to help keep the customer bound to them?

VAN DER GRAAF: I'm not sure if the word business model will apply to it, but yes, we asked the question. Eefke asked all the experts what are the hurdles and what could be the solutions, a cross-publisher solution for it, and they came up with five possible cross-

publisher solutions to facilitate text mining, content mining. Of course, we asked those solutions in the survey and asked what people thought of it.

There were three that were most popular among the respondents of the survey, standardization of content format, of the AP platform standards, semantic tagging terms. Of all the aspects of text mining, content mining, that was the most popular among the respondents and probably that's quite difficult to achieve, I think.

The other idea is that there should be one content mining platform special for content mining and all the publishers deliver their content to it. That was quite popular among all the respondents but we had also so-called expert respondents, people who said that they know a little bit more about content mining themselves. They have quite a good knowledge about it, and then they scored much lower and that's probably because that's really difficult to achieve, such a one-platform solution.

And then another popular solution and that I think we feel that is maybe the way forward for the publishing community is commonly agreed access terms for text mining for research purposes, that if all the access terms are more or less are the same for research purposes across all the publishers, that might help facilitate content mining a lot.

KENNEALLY: A fascinating response there, and it seems to me that like so much of the promise of data in the online world, the potential is there, the dream is there, but the reality is something else entirely. I wonder for both of you, when we come back five years from now, can you give us an idea of where you think content mining will be, what its place will be in publishing? Will it have changed any way that we conduct research? Will it change the way publishers work? Use your crystal balls for us for just a moment and tell us what this is all going to really mean in practice.

SMIT: Well, if I can give that a try. My crystal ball is sometimes a bit clouded, but here and there are a few rays of light. I think my expectation is that five years from now, content mining will be a very normal thing that everybody does everywhere and that it has replaced a bit of how we go through literature now.

Take the – as an analogy, take the difference, the way we use information by what search has brought us in the last, let's say, five to 10 years. Search really became popular around the turn of the century and since then, we've been digesting information in such a different way from before that. And I think that that is what content mining will do to information, which means several different things.

First of all, it will make it much easier for the users and the customers of publishers to make optimal use of the information that the publisher is supplying to them. They can much more easily and much more customized find the stuff in there that they really need. The needle in the haystack? You'll have it. You don't have to search for it anymore.

KENNEALLY: Right. And Maurits, as a publishing consultant, are you advising clients or thinking about advising clients to get into content mining now because the promise is there for real benefit five years from now?

VAN DER GRAAF: Yes, I think I would do so if they asked me. I think I agree with Eefke that it will take off from now on and the semantic Web is a popular term which is related to semantic tagging of information, and I think the semantic tagging part of content mining also will – I would say that will skyrocket between now and five years.

KENNEALLY: All right. We've been chatting with Eefke Smit and Maurits van der Graaf about a report on content mining, but before we close the program though, I wonder if either one of you could tackle a pretty easy question, but I think useful to the audience, which is about that group, the Publishing Research Consortium, which underwrote the report. What is the PRC exactly?

SMIT: The PRC is a consortium of several trade buddies in the publishing industry and they assemble publishers from many different areas, from the scientific, technical and medical background, but also general publishers' associations in different countries. Some of the big publishers support it.

And what the PRC typically does is – you'll find it on their website. They publish four, five, sometimes up to 10 times a year reports on new developments that happen in the publishing industry, like content mining.

KENNEALLY: Well, we can thank them for this report and we want to thank both of you for joining us today on *Beyond the Book*. Eefke Smit is the director of standards and technology for the International Association of Scientific, Technical and Medical Publishers, also known as STM, and she coordinates the activities of the STM Future Lab Committee. Eefke Smit, thank you for joining us today on *Beyond the Book*.

SMIT: Thank you, Chris.

KENNEALLY: And we've also had on the line from Amsterdam Maurits van der Graaf, the owner of Pleiade Management and Consultancy and a gentleman who has held various management positions in publishing, library and documentation institutes, and thank you for joining us as well, Maurits.

VAN DER GRAAF: Thank you very much for an interesting discussion.

KENNEALLY: Well, we appreciate it and look forward to having you back in five years to let us know about content mining in 2016.

VAN DER GRAAF: We'll hold you to that.

KENNEALLY: That's right. OK.

Beyond the Book is produced by Copyright Clearance Center, a global rights broker for the world's most sought-after materials including millions of books and e-books, journals, newspapers, magazines and blogs. You can follow *Beyond the Book* on Twitter, like *Beyond the Book* on Facebook and subscribe to the free podcast series on iTunes or at our website, copyright.com/beyond the book.

Our engineer is Jeremy Brieske of Burst Marketing. My name is Christopher Kenneally. For all of us at Copyright Clearance Center, thanks for listening to *Beyond the Book.*

END OF PODCAST