chem-bla-ics

Elsevier's new text mining initiative is a step sideways

Egon Willighagen

Published February 15, 2014

Citation
Willighagen, E. (n.d.). In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/5e32g-08a89

Keywords
Publishing, Textmining

Copyright
Copyright © Egon Willighagen 2014. Distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Elsevier's new ideas on text mining are getting a lot attention now. Sadly, they get it wrong, again. On the bright side, all other publishers, which are expected to follow this year, can learn from this mistake.

Because if done right, the publishers can even help forward science, despite crippling progress. That sound harsh, and surely they have done a lot of good for science. In fact, we would not be where we are now without the publishers. But things have changed. With the internet anyone can be publisher. We see this with blogs, we see this with Lulu.com. And, unlike some misinformed people think, this is independent from peer review. Publishers were important because they provide a channel to disseminate knowledge. But paper publishing is no longer the most efficient way. In fact, in terms of value, paper has been overtaken for some years now.

And we need more added value. Not the shipping of the knowledge, but keeping up is the issue. And there too, publishing is inefficient: human language is nice for sharing ideas and concepts, but it fails at disseminating raw facts: measured data. Anyone who has tried creating a data set to find patterns knows this: extracting the information is a lot of effort, mostly caused by the broken paper publishing model. This is most apparent in some research domain where data repositories exist, but sadly this applies to a small minority of data types.

Now, text mining seems in that sense the wrong question: why trying to recover knowledge that should have gone into repositories in the first places. I agree. However, we cannot just throw away all the knowledge kept in these papers, and certainly not as long as people keep insisting on seeing only papers as scientific success. We are slowly seeing this improve, but only very slowly. Things that were apparent to me as a student 20 years ago, are the things that scholars are still struggling with today. Depressing indeed, but it does help you grow a good sense of patience.

And now, Elsevier wants to make a step forward, wants to be leading in science dissemination again. And they come up with an intermediate solution between actual knowledge dissemination and profit: they come up with a license-model, increasing their monopoly on knowledge and trying to lure the scientist into a non-commercial license. From a money-making perspective this is what society expects from them. From someone who likes to see societal problems solves, this is disappointing. They had a great opportunity to lead the field.

Now, is all bad? Not at all. It's a step, but not the step I would have liked to see. It will be a success: because the CC-BY-NC data that will come out of it, will be part of the web of knowledge. No one will care about the NC part, except all those SMEs in Europe that work on products to help society which will find it much harder to collaborate with other companies, because they cannot share the knowledge the created from analyzing the literature (does Elsevier want a monopoly in this analysis?).

Nor will many in the academic community complain. Surely, those that have worried about this, they will. But the scholar at universities do not care about NC licenses. After all, universities are not commercial. Asking a student to pay 30 thousand euro for a year is surely not commercial. That is the consensus. But I note that this consensus has not be tried in court, and I am looking

forward to the day it will happen. Elsevier will likely not challenge this, and silently accept this situation. Just like Microsoft never made a big deal out of people copying office versions of their operating system for at home: you do not bit the hand that feeds you (too hard). You rather go after others, like Academia.edu. It will not be scholar Elsevier will enforce the NC on, and it will not be large companies either: if any, it will be the SMEs. Support them, and do not agree with the license.

Well, it was a nice opportunity for Elsevier. I only see my choice to sign The Cost of Knowledge reaffirmed.

The choice of the NC clause is totally useless in any context of dissemination. I call for Elsevier to at least add this option, if they are serious about improving: text mining is provided to subscribers, via a decent API, adhering to:

1. Facts extracted from literature are licensed CCZero and attribution is paid (facts are copyright free in most parts of the world)
2. Output can contain "snippets" of the original text under international "fair use" concepts, and licensed as CC-BY

Any scientist is expected to attribute the source of information in the first place, and it is kind of sad Elsevier is on such bad foot with their audience that they feel this must be enforced via a contract, but that is not a problem. I also see no reason to deviate from international law about "fair use"; I do understand this is probably an ill defined concept, but 200 characters seems pretty limited to me, as facts can be spread of sentences longer than this.

I know that many will disagree on the CCZero license, and many will feel awkward about giving away data. It has value, right? It's your property, right? I am not going to argue against that. But personally I do not understand how it aligns with the idea of scientific dissemination. Holding back knowledge as part of making knowledge available? How exactly does that make sense? Importantly, just like with software, Open is not the same as Without-Cost! Hosting and sharing Open Data also costs money (particularly, if it is 1 TB of data). Those are different concepts.

However, I also stress that the scholars have a great responsibility hear: I call for all Elsevier journal editorial boards to not accept this deal either. In fact, all editorial boards have great say in this: it's them who make a journal valuable. I also call all scholars to be aware the consequences of selling away your copyright. That is a choice in the current era. There are plenty of means to disseminate your science *without* (much) cost, and APC is a flawed argument.

The current step by Elsevier, after all the effort from many, is not a step forward, it's a step sideways. Elsevier, I know you can do better. Are you willing?

I am willing, and have been supporting science by making data available as CCZero. However, I also am happy if others are not ready for this, or have other reasons not to. It is not always under their control. For example, I have heard stories where data has been used by politicians as small change to get industry to test their products for safety. I also accept that getting

funding as a scholar is hard work, often not paid for, and that it is hard to give away your only security of a future career. Then again, we all know what data is valuable, has already given its value, or is of no use to you anymore. And this latter case I ask you to consider to make data available: data of no use to you anymore, but that could be valuable to others. Make it available, and get cited, and get value out of it, you would not have received when it sat on some hard disk, and probably is lost in five years.

I also fully understand this is my opinion. Thus, not all data I make available is CCZero: I fully respect copyright and license from others; in fact, I often feel I do much more than scientists which object to Open licenses, which just take data as their own as they please. That is why I insist often on clear copyright and license information. Because if missing, default (local) law applies.

If you want to read more analysis, please refer to the following posts:

1. Elsevier opens its papers to text-mining
2. #elsevier's TDM Terms (TaC): Can they force us to copyright data? (2)
3. Nature's recent "news" article on Text and Data Mining was unacceptable [redacted]; I ask them to renounce licensing.
4. "Dear Peter,", Richard van Noorden
5. Reply to Richard van Noorden