TDM

Chris HJ Hartgerink

02 June, 2016

The scientific literature is the largest source of knowledge and it harnesses information we have not even discovered within it. By extracting information from it using automated tools, Text- and Data Mining (TDM; also called Content Mining), new insights can be generated. This was not possible before the digital revolution, but it is now. Nonetheless, actually implementing these tools without pushback has proven difficult because of outdated copyright and the oligopoly of scientific publishers.

Moreover, the scientific literature is becoming so large, researchers cannot be expected to manually digest all these findings. For example, Figure 1 depicts the largest publishers available in the CrossRef database, which clearly shows the size of the knowledge base.

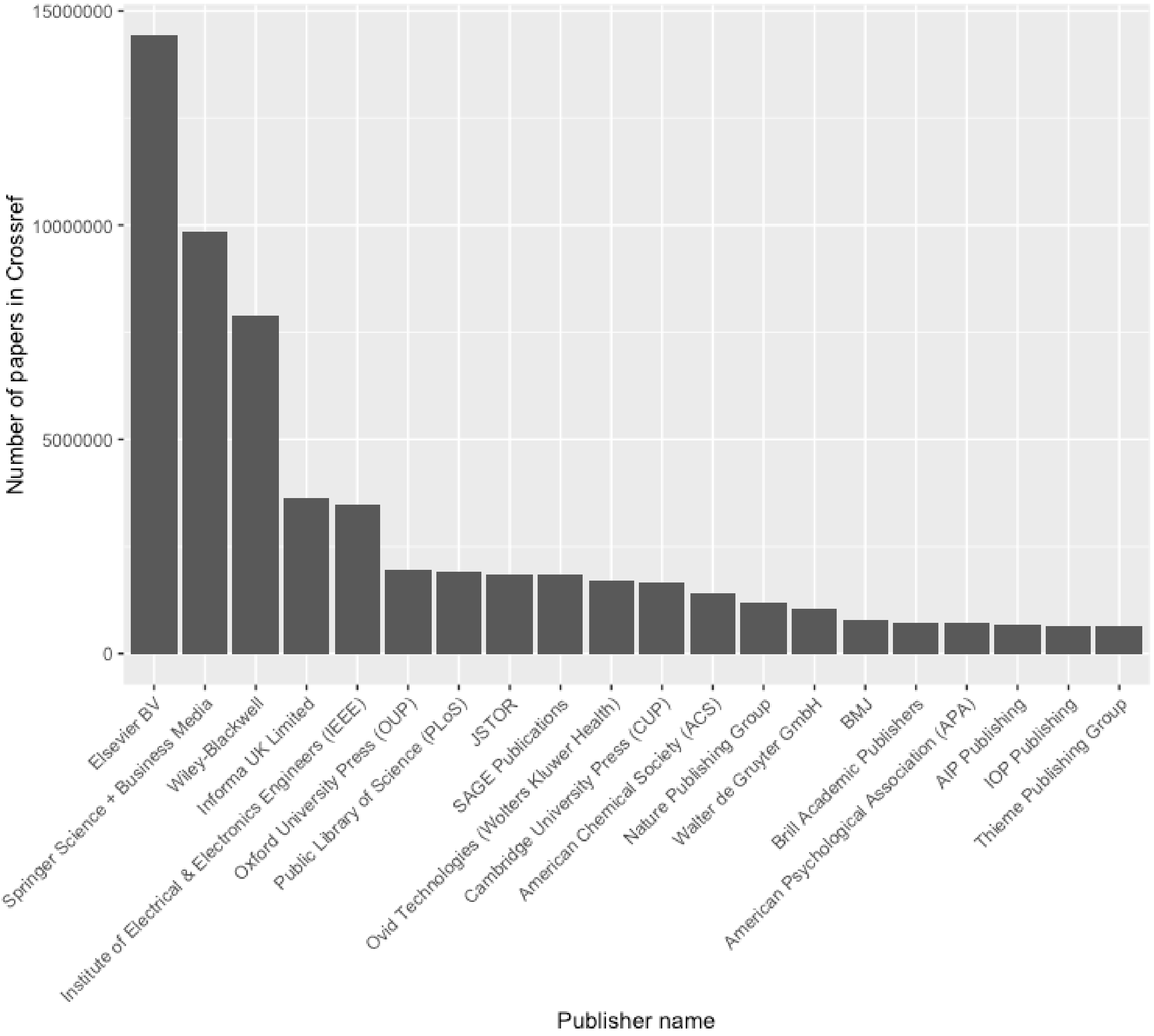

Figure 1. Publications per publisher available in CrossRef. *Image credit: Richard Smith-Unna.*

This methodology, Text- and Data Mining (TDM), contains many fruitful possibilities, including an automated hypothesis finder [@malhotra2013], semi-automated meta-analysis procedures , statistical error detection , and in my case, detecting potential data fabrication in the sciences. This list is hardly extensive, and the potential to innovate is great.

However, researchers wanting to apply these methods face tremendous problems because in order to analyse these articles, they first need to be downloaded.

Downloading does not need to be difficult, considering that an automated download for all scientific literature could ideally require only one line of code. However, the infrastructure to do this has to be built, and the downloads have to be facilitated by the publishers. For example, the publisher PeerJ encourages TDM based on its corpus and places no restrictions on the downloading of articles. As a result, one needs only the following code to download the entire corpus

```
curl -s peerj.com/articles/updat... | csvcut -c url | tail -n +2 | while read -r u; do
curl -H "Accept:application/jats+xml" -O -J -L "$u"; done
```

.publisher's entire corpus can be downloaded with a single line of code., if

https://hypothes.is/stream?q=user:chjh

https://www.copyright.com/business/xmlformining-2/

http://refinder.org/

https://github.com/ropensci/rcrossref

https://github.com/ropensci/fulltext

You do research, find an interesting issue from a journal; publisher forbids you to download entire issue. Yes, this happens.