The mining landscape: A researcher's perspective

Chris HJ Hartgerink

03 June, 2016

The scientific literature is the largest source of knowledge; additionally it harnesses information we have not even discovered within it. Researchers are actively working on extracting these discoveries by conducting meta-analyses, systematic reviews, narrative reviews, archival studies, and more. By extracting that same information from it using (semi-)automated tools, that entire process can be made more efficient and reliable. This (semi-)automation, Text- and Data Mining (TDM; or Content Mining) can help generate new insights.

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, with Elsevier making up approximately 15 million pieces of scientific output. Even if just .1% of all outputs are relevant to any specific researchers that amounts to 15,000 papers — much more than any human can be expected to fully parse for a single project.

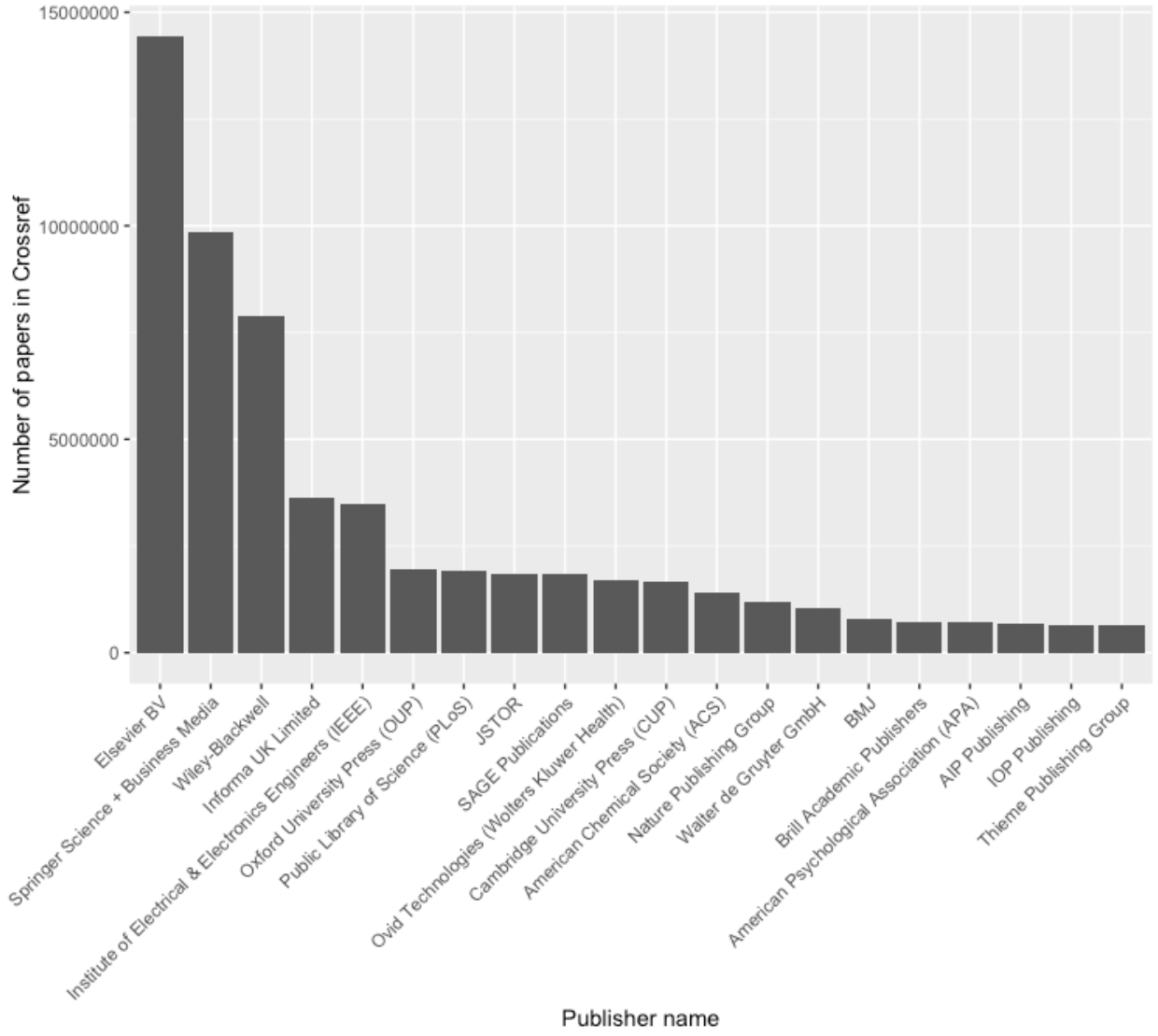

Figure 1. Publications per publisher available in CrossRef. *Image credit: Richard Smith-Unna.*

TDM was not feasible before the digital revolution in the 1990s and has become more feasible in the decades following it.

Despite that TDM has become technically feasible, its practical feasibility is hampered by copyright legislation in most countries.

As we see in Figure 1, the five largest publishers include the majority of all scientific output.

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

https://svpow.com/2012/01/13/the-obscene-profits-of-commercial-scholarly-publishers/

http://libguides.usc.edu/textmining/databases

http://tdmsupport.crossref.org/researchers/

https://github.com/CrossRef/rest-api-doc/blob/master/rest_api.md

http://tdmsupport.crossref.org/

http://www.ubiquitypress.com/site/merch/

https://apps.crossref.org/home/

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