The mining landscape: A troubled researcher's perspective

Chris HJ Hartgerink

04 June, 2016

The scientific literature is the largest source of knowledge; additionally it harnesses information within it that has yet to be discovered. Researchers are actively working on extracting this value by conducting, for example, statistical meta-analyses, narrative literature reviews, and archival studies. By extracting that same information from it using (semi-)automated tools, the entire process can be made more efficient and reliable. This (semi-)automation with Text- and Data Mining (TDM; or Content Mining) can serve as an invaluable research tool.

Moreover, the scientific literature is becoming so vast, researchers cannot manually digest all these findings properly. 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.

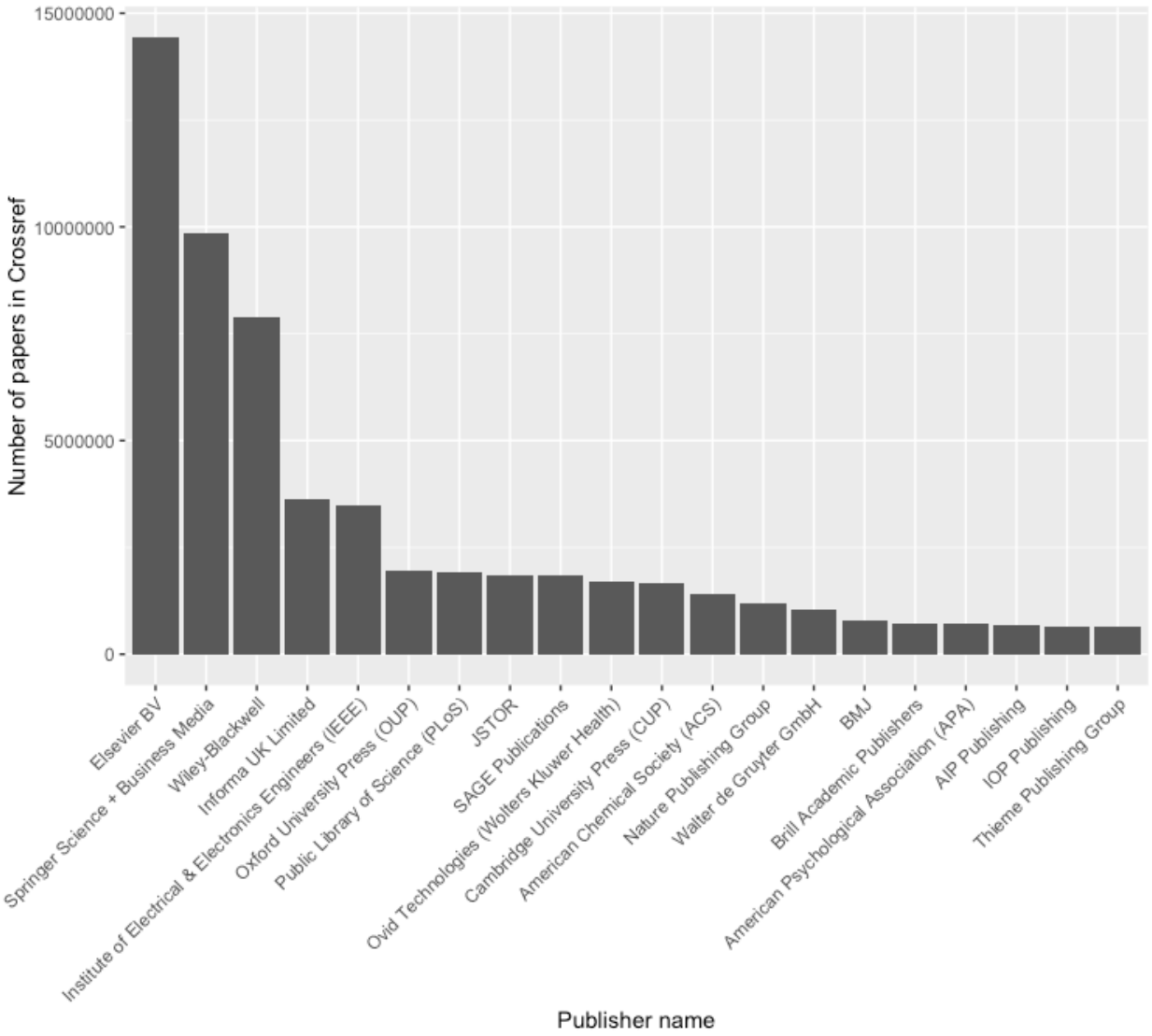

Figure 1. Publications per publisher available in CrossRef. *Image credit: Richard Smith-Unna.*

This new research tool, TDM, was not feasible before the use of computer processing and has become more and more feasible in recent decades. As years passed, computing power increased such that larger corpora of text could be analyzed for more extensive amounts of information. Today, the technical possibilities are in rapid development and uptake has been on the increase as well, albeit minor.

Despite that TDM has become technically feasible due to the digital revolution, its practical feasibility remains problematic due to copyright. Copyright in the analog world meant that you bought a copy and were then restricted in redistributing it. Copyright in the digital world means you buy access, but you inherently need to copy to make use of that access (Doctorow, 2014). As such, creating many copies of that which you have bought access to, is frequently seen as copyright infringement (as we will see later in this article). Considering the use of TDM is inherently tied with computers, digital copies are essential to not only analyzing the literature, which is where researchers can seemingly become copyright infringers, in the eyes of publishers.

In this article, I will outline my attempts at using TDM using two cases and a systematic investigation into the mining policies of 31 publishers. The first case I outline is a previous, completed, project where we manually collected ~30,000 research articles for TDM without problems. The second case is a project where I collected ~300,000 research articles automatically, with problems (the original sample was ~900,000

large). Finally, I investigated the publisher's website terms and conditions, their official TDM policy, and the availability in the CrossRef TDM interface. With these examples, I aim to provide an outline of the potential of TDM, the limitations of TDM, and an inventory of its practical feasibility.

Case 1:

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.

Doctorow, C. (2014). *Information doesn't want to be free: Laws for the internet age.* San Francisco, CA: McSweeney's.