TDM

Chris HJ Hartgerink

01 June, 2016

The scientific literature is the greatest source of knowledge mankind has ever known, but it can be built upon by mining information from it. With the fourth industrial revolution upon us, the information revolution, the scientific literatures aggregate information is ripe for the pickings. Information can be mined from text- and data in previous research articles in order to create new knowledge. 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

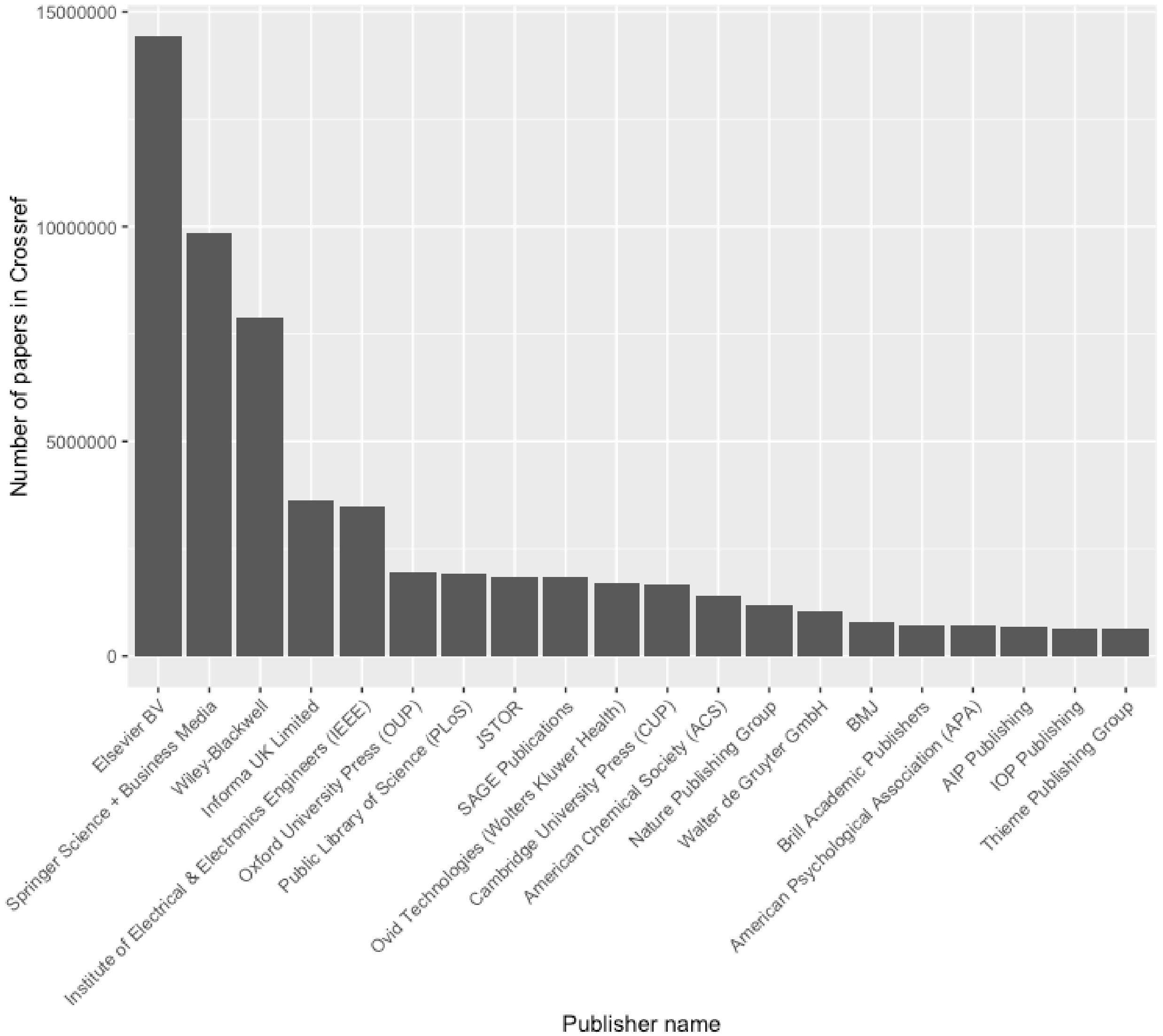

Figure 1. Publications per publisher available in CrossRef. *Image credit: Richard Smith-Unna.*

The researchers wanting to apply TDM methods will have to face navigating the publisher's landscape and copyright's landscape.

This case study focuses on my personal experiences while Text- and Data Mining, incorporating a

Copyright

Copyright landscape

Copyright exception

TDM

Format

HTML PDF XML txt

Elsevier

Block

API

Lega