

Watermanstraat 98, 7324 AK Apeldoorn, The Netherlands

SPARC Europe Guidance to Members

Publishers' Content-Mining Policies

Elsevier has [announced a new policy](#) on Content Mining (text-mining and data-mining, or TDM). The implementation of the policy requires that libraries sign an agreement with Elsevier that will permit their patrons to use TDM tools on Elsevier's content. Below we set out for libraries the essential issues and offer some guidance on how libraries should respond.

What is TDM (text- and data-mining)?

TDM is a research process whereby computing tools are applied across multiple research articles to analyse their content and create new knowledge by combining information gathered from them. They look for facts, entities and relationships within the text and extract that information for analysis. Importantly, they can analyse information across a broad range of fields and across many articles (hundreds of thousands if necessary), so taking analysis to a different level than that which the human brain can manage.

Although in relative infancy, these technologies hold great promise for the future and are already proving their worth in fields such as pharmaceutical, biomedical and chemical research. The important points about TDM are:

- Researchers carrying out TDM need to mine the contents of large numbers of articles
- These articles may be in multiple fields of research
- These articles may be published in many different journals by many different publishers

What has been the publishers' position on TDM?

Publishers positions do vary, but in general publishers consider that material licensed to libraries is made available for readers to download and read, but not to text-mine. In some cases researchers who have tried to use TDM tools to crawl large numbers of articles from a single publisher have found that their IP address has been blocked: since the IP address used is usually that of their institution, the repercussions are serious.

A number of publishers are developing text-mining tools that they intend to provide to researchers: however, researchers who wish to carry out text-mining generally develop their own tools for the job (some of which are necessarily rather specialised) and would prefer to use those. It is not clear why researchers would expect publishers to provide research tools for them, though it is easier to see why publishers may wish researchers to use their proprietary tools.

In addition, some publishers have said that they would make a charge for allowing text-mining of their content if a commercial product were to result, presumably fearing that what could be created by such activities (new databases) may damage their own business prospects.

What is the position of research libraries?

Libraries consider that when they have licensed access to a publisher's content it should be available for their patrons to mine as well as to read. LIBER has published a useful and informative [factsheet on TDM](#) that sets out the case from the library community's position.

What has Elsevier's position been until now?

Until now, Elsevier has required researchers who wish to mine its content to seek specific permission from the company. This has been granted on a researcher-by-researcher basis. This solution may have helped individual researchers but is not a sustainable one in the face of the growth in text-mining.

What is Elsevier's new position?

Elsevier's [new policy](#) enshrines TDM rights in the standard ScienceDirect subscription agreement for academic libraries. It applies to all new subscription agreements and can be added to existing agreements on request.

Researchers at a subscribing institution wishing to undertake TDM must register with Elsevier's developer's portal, upon which they will receive a key to access the ScienceDirect API. Outputs must be licensed using the Creative Commons CC-BY-NC licence (i.e. the product of the text-mining exercise cannot be used for commercial purposes). The researcher must also add the DOI of the original sources so that authors may get credit and there is a verifiable provenance to the data. The policy allows researchers to re-publish 'snippets' from the original papers, but these must not exceed 200 characters in length.

What else is going on regarding publisher permissions to text-mine?

A number of publishers have agreed to work towards a solution for researchers who wish to mine across a body of articles published by multiple companies. There is a developing initiative from CrossRef, called Prospect, that seeks to address this by providing a common API and a licence framework whereby researchers can agree to terms and conditions from multiple publishers through one portal.

What are the research-based problems with Elsevier's licensing scheme and the one proposed by Prospect?

For researchers wishing to undertake text-mining, the types of licence-based solutions have the following problems:

- Researchers wishing to mine across vast collections of articles (e.g. from much of their library's collection) find having to sign separate agreements with different publishers a barrier to their work. The Prospect initiative goes some way towards resolving this administrative burden
- Research teams collaborating across multiple institutions where mining will take place in each institution will require all the individual libraries involved to be signed-up to Elsevier's licence terms
- Researchers do not always wish to work through an API, but to mine directly the articles themselves. There are new technologies that enable mining even of PDFs of articles, a great step forward technically, but licences such as Elsevier's one will not allow this: researchers must use ScienceDirect's API
- Elsevier is not including images in the content that is minable. Researchers wishing to mine images must request specific permission for this and Elsevier will provide an image retrieval API in return. Elsevier's reason for examining each case individually is that there is some ambiguity about re-use rights for some images. This presumably implies that Elsevier will want to know the purpose for which the researcher will put the resulting mined data: some researchers may not wish to divulge this
- Researchers must use a non-commercial (NC) licence when publishing the resulting information from the text-mining exercise. This may not suit all purposes

What else is going on that affects whether a library should sign a licence?

Governments are taking an interest in this issue. The UK Government intends to act in Spring 2014 on the recommendation in the [report](#) by Professor Ian Hargreaves for the UK's Intellectual Property Office that there is an exemption from copyright law for text-mining for non-commercial purposes.

The European Commission is also exercised by this issue. Last year (2013) it conducted a 'stakeholder dialogue' exercise on its [Licences for Europe](#) initiative under which it was proposed to have a specific European licence for TDM. There was a working group specifically addressing this issue (WG4). The organisations representing the research community, led by library organisations, pulled out of the process in May because their aim was not to have a licence-based solution at all, but instead to have an exemption from copyright for TDM. The exercise has now concluded and it is generally agreed that in its intended form it has no future. The Commission has now invited Professor Hargreaves to chair a group convened for the purpose of examining the economic impact of TDM in scientific research and the barriers to its use. The conclusions are expected at the end of February 2014.

So, should libraries sign a licence with Elsevier?

If you are about to renew or renegotiate your deal with Elsevier, the licence to text-mine will be part of the terms and conditions. You may of course choose to negotiate to omit that clause. If your deal still has some time to run, you can have the licence for your patrons to text-mine included straight away by separate agreement.

If you do sign, your patrons will be able to undertake text-mining under Elsevier's terms. This may not suit all of them, so a consultation with appropriate researchers is in order before the licence is signed. The issues are laid out here and should be fully explained to researchers so that they can take a position on this.

Should you decide not to sign, your patrons will be unable to mine Elsevier's content for the foreseeable future. It is possible, however, that moves by the European Commission and national governments to exempt research activities from copyright law may in time render publisher licences moot.

Other reading:

Trouble at the text mine. *Nature News & Comment*, 8 March 2012: <http://www.nature.com/news/trouble-at-the-text-mine-1.10184>

Elsevier opens its papers to text-mining. *Nature News & Comment*, 3 February 2014: <http://www.nature.com/news/elsevier-opens-its-papers-to-text-mining-1.14659>