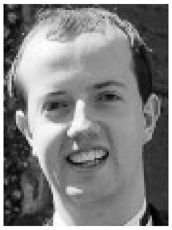
The right to read is the right to mine: Text and data mining copyright exceptions introduced in the UK.

blogs.lse.ac.uk/impactofsocialsciences/2014/06/04/the-right-to-read-is-the-right-to-mine-tdm/

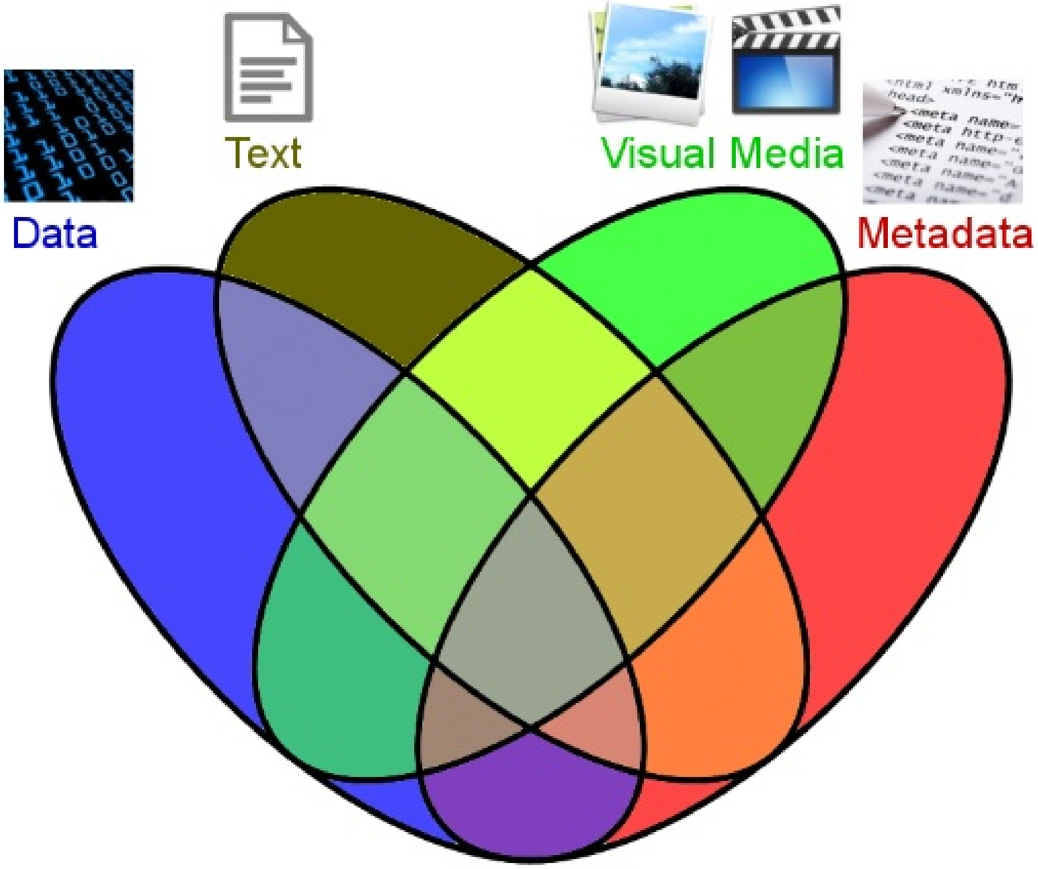
6/4/2014

New copyright exceptions to text and data mining for non-commercial research have recently come into effect and this is welcome news for UK researchers and research, argues **Ross Mounce***. Here he provides a brief overview of the past issues discouraging text and data mining and the what the future holds now that these exceptions have been introduced. But despite legal barriers being removed, many technical barriers still remain. Furthermore it remains to be decided what formally constitutes 'non-commercial' research.*



After eight long years including not one but two expert-led reviews of intellectual property; new copyright exceptions, some of which in particular will enable and empower UK academic research came into force on June 1st 2014. All disciplines are set to benefit from this : the humanities, the social sciences, science, technology and medicine.

Of particular interest to myself and other researchers is the 'Exception for copying of works for use by text and data analytics'. In order to understand why this is so important, let me take you back to how it was before the copyright exception came into force (and how the legal situation *still is* for researchers in most other European countries):



Content Mining: mining one or more types of media for information; media as data.

The situation before the copyright exception

Before this exception came into force in the UK, for subscription-access content, you'd essentially have to ask *permission* from the publisher, before you started analysing. If you proceeded, *without permission*, to download electronic copies of 'their' copyrighted materials (see author's note) *en masse* for analysis, you would be infringing 'their' copyright – it would be illegal, and they could take legal action against you, even if your analysis was undertaken for non-commercial academic research purposes. Depending on the exact subscription-access agreement held with your institution (of which your institution may not be able to disclose the details of, because of confidentiality clauses!), the publisher could even ask for additional fees to be paid to cover this 'additional' type of usage if it is not covered in the subscription agreement. Many agreements did & and still do explicitly prohibit text and data mining.

If one did ask for permission, the process was complex and lengthy, involving many employees, and much bureaucracy, for each publisher. That's *if* the publisher agrees to give permission. In a study by the Publishing Research Consortium it was found that "only 35% of the respondents [publishers] state that permission is granted in the majority or in 100% of the cases for all requests" (p. 106) – and that sample of publishers included open access publishers that *by definition*, allow mining. Thus publishers can and have denied permission for content mining research on 'their' works.

The situation *after* the introduction of the UK copyright exception for TDM

After June 1st 2014, for research conducted in the UK, under the jurisdiction of UK law, for 'non-commercial' research purposes (more of which later…), the new copyright exception overrides anything in subscription contracts that prohibits content mining. As Peter Murray-Rust puts it: The Right To Read Is The Right To Mine, and provided you are in the UK, and doing 'non-commercial' research that is now true, and legal. This provides welcome and useful protection for researchers against litigious publishers.

> *No researcher, doing 'non-commercial' research in the UK, needs to agree to, nor abide by the terms of any text and data mining 'licence' that publishers may wish to impose upon researchers.*

Schemes such as CrossRef's text and data mining services will be heavily advertised to researchers by the major publishers, in order to try and control the way in which researchers do content mining both through legal means (the licencing) and technical means (the API). The use of such services entails agreeing to detailed and lengthy licencing agreements, which many probably won't read. If you do read the full terms and conditions you'll find them disappointingly limiting which is why organisations such as LIBER have publicly criticized these terms.

Even with some legal barriers now removed, technical barriers remain

Despite legal barriers being removed, non-trivial technical barriers still remain which can be problematic for content mining. Most websites for instance have rate limits. If you are detected attempting to crawl or scrape too many pages (i.e. research articles) within too short a time-span, your access to that website may be blocked. Publishers such as BioMed Central (BMC) have a crawl rate limit of one article per second which is an acceptable rate limit for researchers. Through Elsevier's text-mining API there's a limit of 10,000 articles per week which is equivalent to a rate limit of 1 article every 60 seconds. At that rate limit it would take ~21 years to go through all 11 million articles that Elsevier control access-to through their Science Direct platform – not really feasible! The rate limit imposed is entirely artificial – researchers with good internet connections can crawl many articles per second if they were allowed to. The publisher sets the rate 'allowed' and even despite this new copyright exception, to get the rate-limit changed a researcher would still have to beg permission from the publisher, which the publisher is fully within their rights to either grant or not.

Open Access publishers tend to be exceedingly helpful to content miners: BMC, Hindawi, and MDPI to name but a few, make available whole content dumps (i.e. everything they publish), openly available to download by anyone for any purpose which greatly facilitates content mining. For biomedical researchers, the PubMed Central Open Access Subset and Europe PMC also allow downloading of full-text dumps but these are limited to CC BY papers only (another reason of many why CC BY is the preferred licence of open access publishers).

Other less-helpful publishers sometimes pay money to employ external firms like Atypon to populate their websites with booby-trapped links that block access to the entire subscribing institution if clicked. These links called 'spider-trap' links, inevitably end-up doing more harm than good as in the recent #ACSgate debacle whereby over 200 institutions had their access to one publisher's content blocked by people innocently clicking on a DOI-like URL that was openly available on the publisher's website.

Image modified from Flickr user Dan Perry / CC BY 2.0

Why do these publishers dislike crawling and scraping so much? Scraping the web is a normal, legitimate activity for researchers; even a recent European Commission report says so:

> *'Scraping' the World-wide web for data is today a familiar activity for the digitally literate researcher. (p. 11)*

With over 50 million scholarly articles out there, and millions being published each and every year in popular fields like biomedical science; content mining is fast-becoming a necessity. Human-eyes can only read so much. Computers, and computational techniques to help us comprehensively and rigourously mine the literature are a boon for research. One expert report on the state of content mining argues that "European academics are falling behind their Asian and North American counterparts" – this new copyright exception will thus help the competitiveness of UK research in the global sphere.

The only nagging question remaining to address is: what is 'non-commercial'?

I won't pretend to give a convincing answer to this. I simply don't know, and I can see it being a potentially difficult sticking point for many.

For my own research on extracting data from evolutionary tree figures (phylogeny), I can feel safe that this subject and use-case might not readily be definiable as 'commercial' but for other researchers I can imagine it may not be so easy to safely & surely classify their research as 'non-commercial'. Indeed a recent court in case in Germany seemed to indicate that 'non-commercial' use was only safely equivalent to personal-use. The consequences, risks and side-effects of 'non-commercial' remain largely untested in case law and can prevent much more usage than you might think. Will publishers be eager to sue academic researchers for what they perceive to be commercial mining? I hope not, but sadly it would not surprise me if they did.

Author's Note: *I feel pained to discuss the copyright owned by publishers, over work written by academics hence the inverted commas when discussing 'their' copyright. Part of the reason academia got into this copyright-pickle in the first place is that we allowed publishers (and still do for some!), to take copyright away from the authors with completely unnecessary copyright transfer agreements (CTA's). Publishers do NOT need a CTA to publish your work, so don't sign them! You can instead retain your copyright over your work, and just give them a non-exclusive license to publish. Keep your copyright!*

Disclaimer & Warning: None of this article constitutes formal, vetted legal advice and should not be relied on or treated as a substitute for specific advice relevant to particular circumstances. Academic publishers and even societies can, and *do* take legal action against research-related activities, if they feel so inclined.

About the Author

Dr Ross Mounce is a BBSRC-funded postdoc at the University of Bath, working on the PLUTo project to liberate phyloinformatic data from the literature. He is working with The Content Mine team to encourage the adoption and use of content mining tools and techniques, including giving a workshop at this year's Open Knowledge Festival 2014 (Berlin, July). As a keen advocate for open scholarship, you can also find him at OpenCon 2014 (Washington D.C., November) – the student and early career researcher conference on Open Access, Open Education and Open Data.

Featured Image credit: DeclanTM (Flickr, CC BY)

Note: This article gives the views of the authors, and not the position of the Impact of Social Science blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.