To boldy go…. Text and data mining: the role of librarians

Jane Secker, Chris Morrison, Neil Stewart and Laurence Horton, July 2016

Introduction

The relatively new exception to copyright law that we enjoy in the UK, permitting text and data mining (TDM) for the purposes of non-commercial research, offers much potential to further knowledge and make scientific and medical breakthroughs. Importantly, the new exception states that any contractual clause which purports to restrict this exception is automatically null and void. Librarians who manage electronic resources and datasets, can assist researchers greatly. However, in order to do this they need a robust understanding of the law and be assertive in their relationships with publishers when negotiating or interpreting licence agreements. This article examines the ways in which librarians can facilitate the work of researchers who want to use TDM. It also argues librarians need to encourage researchers to exploit the new copyright exceptions as key partners in the research process.

Background

As Seth Godin told us back in 2011, librarians are no longer the gatekeepers to knowledge, they are gate-openers[1]; the walls have come tumbling down and researchers can get access to vast amounts of data and knowledge without ever needing the step into a library. However, as librarians we are well aware that open access and data sharing do not currently deliver all the resources an academic might require to undertake research in almost all disciplines. Although debate continues to rage over how to optimise and sustain systems of scholarly communication, advocates for libraries and research successfully argued for changes to the law to allow researchers to make full use of information they have already paid for access to. In 2014 amendments to UK copyright legislation following the Hargreaves Review of intellectual property law, removed legal barriers for researchers who wish to perform text and data mining (TDM) for non-commercial research purposes, providing they have legitimate access to the materials they wish to mine. This change holds the promise of being able to answer entirely new research questions that were previously not amenable to enquiry.

Exploiting copyright exceptions requires librarians to be bold about the advice they give researchers, to encourage them to use the facilities such as TDM. This may at times mean challenging or ultimately contravening the terms and conditions of a licence with an electronic resource supplier. However, it is possible to do this responsibly through considered institutional approaches and working with colleagues to share best practice.

Data Sharing: the panacea or holy grail

The concept of data sharing is gaining a lot of attention currently as being a way of potentially offering solutions to many of the world's problems, from transport to healthcare. Data sharing is strongly linked to, but not synonymous with, the concept of open data which is defined as data or content that can be "freely used, modified, and shared by

[1] http://sethgodin.typepad.com/seths_blog/2011/05/the-future-of-the-library.html

anyone for any purpose" (Opendefinition.org)[2]. It is not always possible for all data to be open for ethical, practical or commercial reasons, but the principles behind data sharing are to encourage re-use of research outputs as widely as possible. Just as we increasingly have open access to publicly funded research outputs, advocates of the open data movement such as Nigel Shadbolt, who keynoted at the CILIP 2016 conference, believe that open data will eventually become the norm. However, many researchers recognise that much of the world's data is not open or available for sharing, but owned by commercial organisations and kept behind paywalls. Licensing and copyright are potentially restricting our ability to use data in new and interesting ways which is why the text and data mining exception is so important.

Content Mine (http://contentmine.org/), is a project led by Peter Murray-Rust that is doing a lot of work to highlight the TDM issue. They strongly advocate that publishers who restrict researcher's ability to mine through technical protection measures should be challenged. Content Mine have been leading the way with a number of TDM projects, for example a project to mine the research literature for information about the Zika virus. They offer software, which allows you to interrogate the research literature and process it to search for keywords or phrases. They also run workshops and training which include more information about copyright law and the TDM exception.

Comment [CM1]: Please can these sections be put into separate boxes, so not in the main body of the article?

What might researchers want to mine?

TDM is generally performed on corpora of text or data in electronic form. For example, a body of chemistry research literature might be mined by running TDM software on it to discover then create a database of molecular structures, a valuable resource for molecular chemists, crystallographers and other scientists. Another example is the 'Robots Reading Vogue' project at Yale[3] which uses the ProQuest licensed archives of Vogue magazine to provide fascinating insights into the changing use of imagery and language in fashion throughout the 20th century.

What problems might researchers encounter?

There are various problems associated with TDM where researchers might call on librarians for advice. These include:

- Reaching a download limit (often arbitrary and relating to contractual stipulations made by the publisher) for papers from a particular publisher or database, at which point access to a resource might be cut off.
- Being served with an "unusual behaviour" report from publishers, for example when systematically downloading large amounts of material.
- Encountering Digital Rights Management (DRM) technologies that prevents them using TDM methods, which then require a request for permission to have the DRM removed.

In the first instance, librarians can be deal with these issues on a case-by-case basis, but as the library builds up more expertise in this area they might consider writing some FAQs or developing toolkits to assist researchers.

[2] http://opendefinition.org/
[3] http://dh.library.yale.edu/projects/vogue/

How else can librarians help?

There are a number of additional things that librarians can do to assist with TDM, including:

- Highlighting the relevance and application of the TDM exception as part of any copyright training they offer to researchers and advertising TDM support as a service. This could be promoted alongside related research support services.
- Encouraging the development of partnerships with academic colleagues to work on TDM projects, through highlighting institutional collections that might be suitable for mining.
- Being clear about licensing and terms & conditions under which resources are made available to colleagues at your institution, and ensuring they fully understand the copyright exception that permits TDM for non-commercial research regardless of contractual stipulations.
- Being firm with publishers and other content suppliers in the case of "unusual behaviour" reports or DRM blocks to protect the legitimate interests of researchers, as defined by law.
- Advising on next steps should local support not be enough, up to and including appealing to the Intellectual Property Office (IPO) if a rights-holder will not remove DRM locks. Information on making a complaint can be found on the IPO's website[4].
- Being aware of data protection issues: TDM is not exempted from data protection law, so content identifying a living individual must be processed in compliance with the Data Protection Act (1998).
- Being mindful of clauses in any new licence agreements for resources that might restrict TDM activities.

TDM, Copyright Risk and Anxiety

Librarians understandably might feel anxious about sanctioning or undertaking action that is legally ambiguous[5]. The nature of UK copyright exceptions is that they are defences to accusations of infringement rather than rights, and this puts the onus of responsibility on the person doing or facilitating the copying to ensure it is legal. Similarly, the hard won changes to the law which mean that contracts cannot override exceptions are fundamental to equal access to information in today's digitally connected world. However, these require reviewing a legal contract drawn up by a well-funded commercial organization and deciding to ignore parts of it.

It is essential that information professionals are supported by senior managers to take a measured, yet assertive approach to use of TDM. This requires a mature attitude to risk, which balances the institution's reputation and continuity of its business against its ultimate mission. Allowing a situation where the default position is to acquiesce to the demands of commercial publishers because of a lack of institutional capacity (for example not having a properly resourced copyright support provision) potentially creates a bottleneck in the system. In 2015 The Publishers Licensing Society undertook research which suggested that the TDM exception had brought little value to researchers and that commercial publishers

[4] See https://www.gov.uk/government/publications/technological-protection-measures-tpms-complaints-process

[5] http://www.slideshare.net/seckerj/copyright-literacy-in-the-uk-understanding-library-and-information-professionals

had responded to the challenge by creating innovative licensing solutions[6]. These findings seem premature and the research question and methodology also appears to be framed to reinforce their members' commercial interests rather than seek an objective view. However libraries should ask themselves what role they are playing (or not playing) in changing the status quo and advancing the research agenda.

The Libraries and Archives Copyright Alliance (LACA) and Universities UK have been collecting evidence of problems researchers run into when trying to use the TDM exception and in September 2015 LACA appealed to the IPO on behalf of a UK academic to have DRM removed from a site he wished to mine. In this instance the academic was trying to mine a publicly accessible website, which used CAPTCHA technology that prevented him from downloading more than a few records at a time. The case was eventually referred to the IPO who, after some deliberation, decided the case was not within scope of the exception. They argued the exception did not apply to work "made available to the public on agreed contractual terms in such a way that members of the public may access them from a place and at a time individually chosen by them." You can read more about the case on the LACA website[7], however it seems likely then that this would apply to any web-based resource. Although the outcome of the application was not successful, it is hoped that with continued referral of similar cases the library community will be able to create an environment where publishers respect the new exception and researchers are more easily able to access the information they need. As Lauren Smith said in the closing keynote of this year's CILIP conference (in the context of public library closures), it is important to keep fighting battles even if you think it's unlikely you will win. Without further applications to Government to address this issue on a case-by-case basis there will be no possibility to change the status quo.

Conclusion

Researchers will increasingly look to librarians for assistance with TDM, and they should be ready and willing to assist with these projects, and actively promoting the opportunities that technology provides. Librarians are well place to help with rights issues and to create case studies which can influence the legislative reform required to use TDM for the wider benefit of society. The Hague Declaration[8] is a powerful statement, which advocates removal of legal barriers to knowledge creation in the digital age and the declaration web site includes a number of case studies from across Europe. The future of the UK's involvement in efforts to further harmonise TDM laws across the EU may now be in a state of legislative limbo following Britain's EU referendum result but there are still moves afoot to bring member countries in line with the recent UK legislation. This may still allow use of TDM in multi-national research partnerships across Europe, something not currently permitted where licences are not available or practical to acquire.

Ultimately, librarians should be mindful of the right to mine, and recognise the important role they can play in liberating data and knowledge. Their continuing mission: to explore

---

[6] http://www.pls.org.uk/news-events/n-tdm-august-15/

[7] http://www.cilip.org.uk/sites/default/files/documents/notice_of_complaint_to_the_secretary_of_state_-_test_case_1.pdf

[8] http://thehaguedeclaration.com/

strange new worlds, to support research and discovery of new scientific breakthroughs, to boldly go where no one has gone before.

Further reading
Copyright User website: http://copyrightuser.org/topics/text-and-data-mining/

JISC Guide to Text and Data Mining: https://www.jisc.ac.uk/guides/text-and-data-mining-copyright-exception.