Responsible Content Mining

Maximilian Haeussler, Jennifer Molloy,
Peter Murray-Rust and Charles Oppenheim

June 16, 2015

Abstract

The prospect of widespread content mining of the scholarly literature is emerging, driven by the promise of increased permissions due to copyright reform in countries such as the UK and the support of some publishers, particularly those that publish Open Access journals. In parallel, the growing software toolset for mining, and the availability of ontologies such as DBPedia mean that many scientists can start to mine the literature with relatively few technical barriers.

We believe that content mining can be carried out in a responsible, legal manner causing no technical issues for any parties. In addition, ethical concerns including the need for formal accreditation and citation can be addressed, with the further possibility of machine-supported metrics. This chapter sets out some approaches to act as guidelines for those starting mining activities.

1 Introduction to Content Mining

Content mining refers to automated searching, indexing and analysis of the digital scholarly literature by software. Typically this would involve searching for particular objects to extract, e.g. chemical structures, particular types of images, mathematical formulae, datasets or accession numbers for specific databases. At other times, the aim is to use natural language processing to understand the structure of an article and create semantic links to other content. This chapter aims to provide a practical introduction to responsible use of content mining technologies. Such practical advice was highlighted as necessary in the JISC report 'The Value and Benefits of Text Mining to UK Further and Higher Education' [McDonald et al., 2012]:

> "Recommendation 5: Advice and guidance should be developed to help researchers get started with text mining. This should include: when permission is needed; what to request; how best to explain intended work and how to describe the benefits to research and copyright owners."

We aim to address the points above alongside technical guidance on choosing and configuring software for content mining that will behave responsibly on the web and in tracking licensing and attributions throughout a text mining project (Figure 1).

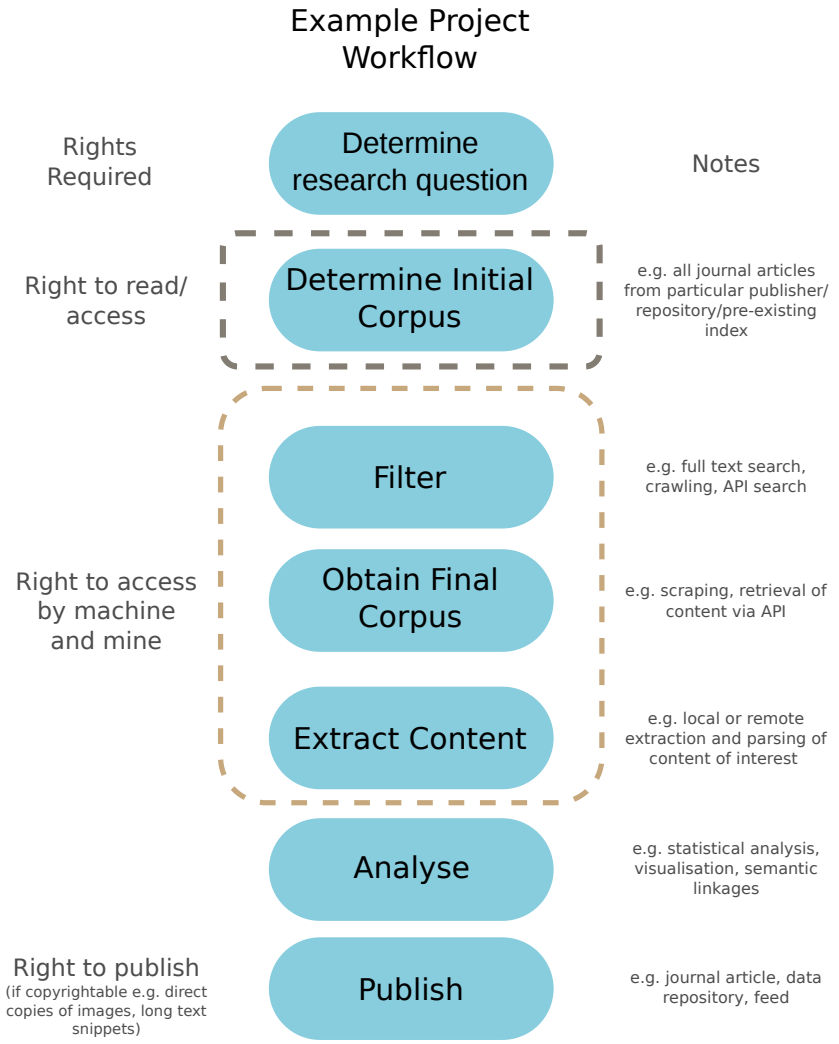

Figure 1: An exemplar workflow for a TDM project detailing the rights required at each stage and the type of activity undertaken.

2 Obtaining Permission to Content Mine

Responsible content mining entails operating within the law; currently, the legality of applying content mining technology to published articles falls under three legal areas, copyright, database rights and contract law. Permissions granted to universities and other research institutions under publisher licence agreements may deal with the right to view, download, text mine and publish the resulting analysis separately. Many researchers in the text mining community assert that 'the right to read is the right to mine' [Murray-Rust et al., 2012], but until this right is confirmed by changes in publisher policies and/or by changes to legislation, text miners have no choice but to make individual approaches to obtain permissions, often on a publisher by publisher, if not journal by journal basis.

This time-consuming burden places huge restrictions on the availability of content to mine and on the time available to perform a mining based study. For example, the Wellcome Trust found that it is illegal to mine 87% of articles in UK PubMedCentral, the UK's main medical research database [McDonald et al., 2012]. It also calculated the cost of obtaining permissions to mine articles containing the word 'malaria' in the title, which would total around £3000 of researcher time for the 187 journals covered. Moving to a full text search for malaria would generate enough journal hits that making the relevant requests would take 60% of a working year [McDonald et al., 2012]. Researchers such as Max Hauseller and Casey Bergman have tracked their efforts at obtaining permissions for the text2genome project enabling automatic annotation of human genome regions with relevant papers, demonstrating the sheer volume of contacts that must be made across journals and time taken, with 27 positive responses since 2009 from a total of 46 requests[1]. It is therefore not surprising that the vast majority of text mining studies undertaken so far have been based on Open Access materials. However, inevitably, this means that such studies are limited, as in all subject areas there will be either a significant minority, or even a majority of texts that are behind paywalls and subject to licence permissions.

The following section explains to the potential content miner how the law applies to content mining, maps the current landscape of publisher policies and offers practical guidance on requesting permissions for mining.

2.1 Copyright law as applied to content mining

Copyright protects the creative expression of intellectual output and in general, prohibits copying, publicly distributing, and/or adapting the original without the permission of the copyright holder. Small amounts of data are not generally considered a creative expression, and so are not protected by copyright, although in the EU and in the US, collections of data where there is clear evidence of creativity, for example in the fact that the collection of data involved has involved decision-making on what is to be included and what is not, are protected by copyright. Databases are discussed further in section 2.2 below.

The Agreement on Trade Related Aspects of Intellectual Property Rights [TRIPS], Article 9(2) states:

> "Copyright protection shall extend to expressions and not to ideas, procedures, methods of operation or mathematical concepts as such."

[1] Available at: `http://text.soe.ucsc.edu/progress.html`. Accessed 18 September 2014

Copyright does not apply to single words or short sentences, but anything from a long sentence to an entire article or book will be subject to copyright. The original owner of the copyright is the author unless the work was created as part of their employee duties, when the default legal position is that the employer owns the copyright. However, by custom and practice, most universities and public sector research institutes do not enforce their rights, and leave the copyright with the researchers. Private sector research bodies, such as the R&D Departments in pharmaceutical industries, are less likely to be so generous. However, typically, when submitting outputs to a publisher, the original copyright owner agrees to assign their copyright to the publisher (i.e., to pass over ownership). At that point the researcher has no further say in how the publisher chooses to exercise the copyright, and indeed, may find that they themselves are unable to access their own outputs without permission. Not all publishers require such assignments, and even those that do will often back down when challenged by the researcher who has submitted the manuscript and will be content with a licence to publish, leaving the ownership with the researcher. But it is often difficult to ascertain this fact amongst a publisher's electronic offerings.

In order to undertake content mining, copying and/or adaption and/or digitisation of the original work is required. These initial acts, even if that copy is never made public, is a potential infringement of copyright. Infringement takes place whenever a third party (i.e., not the copyright owner) copies, or adapts all or a 'substantial part' of a copyright work without express permission. 'Substantial part' means 'what is important in the work', and arguably any content mining activity will involve a substantial part of the original work. Incidentally, the fact that large scholarly publishers own the rights to the vast majority of the materials on their services means they can carry out text and data mining on a large scale.

To review the implications for potential content miners: the acts involved in content mining involve potential copyright infringement. The only way one can be sure that one is not infringing copyright is *either* if the copyright owner has given explicit or implicit permission for third parties to copy, *or* if the actions fall under an exception to copyright. Exceptions to copyright are to be found in all countries' copyright laws, but differ in detail. In essence they say that certain actions will never infringe copyright, as long as the ground rules imposed by law are followed, i.e., there is no need to ask for permission to undertake these actions. Bear in mind that any exception relates to the country where the copying is taking place. Thus, for example, the fact that the materials to be mined are owned by a US corporation but are maintained in Canada is irrelevant. What is relevant is where the instruction to carry out the mining comes from. Thus, there may be exceptions to copyright which will lessen or remove copyright issues in your jurisdiction:

1. Copyright exemptions for content mining (Japan, UK since Oct 2014)

   This is the most important of the exceptions. Japan was until 2014 the only country with a copyright exception specifically for content mining purposes, which is a provision in Article 47 of The Japan Copyright Act [McDonald et al., 2012]:

   > "For the purpose of information analysis ('information analysis' means to extract information, concerned with languages,

> sounds, images or other elements constituting such information, from many works or other much information, and to make a comparison, a classification or other statistical analysis of such information; the same shall apply hereinafter in this Article) by using a computer, it shall be permissible to make recording on a memory, or to make adaptation (including a recording of a derivative work created by such adaptation), of a work, to the extent deemed necessary"

The UK Government implemented a change in its law [The Copyright and Rights in Performances , Research, Education, Libraries and Archives], for non-commercial content mining from October 2014 following the recommendations of the Hargreaves Report [Hargreaves, 2011]. The change in the law makes text and data mining permissible without going through any bureaucracy or paying any fees for those with "lawful access" to the original materials, as long as the mining is for a "non-commercial purpose". In practice, lawful access means a licence of some kind, either paid for e.g. access via a University subscription or free e.g. via Creative Commons licensing. Importantly, the new UK legislation makes any contractual term that purports to restrict one's ability to take advantage of the new exception null and void.

In addition to explicit copyright exemptions, some Nordic countries have laws that facilitate text mining through university libraries [LACA].

2. Fair use clauses (US and Israel)

   The 'Fair use' exception in the US is generally considered sufficient to allow for content mining, although researchers may be reluctant to test the law, which is not well defined for these activities. The difference between "fair use" in the US and "fair dealing" in the UK is that the US law does not specify the purpose of the copying, whereas fair dealing is very specific. This means that whilst fair use appears to be generous, one can only be certain by testing a case in Court – and indeed, the US is famous for the number of Court cases which do test the limits of fair use in various circumstances.

   In terms of digitisation and TDM, there is limited case law but notable recent developments include the American Author's Guild vs Hathi Trust[2], who had created a searchable digital library for universities and research libraries [Kishor, 2014]. The presiding U.S. District Judge Harold Baer decided that this did fall under fair use:

   > "I cannot imagine a definition of fair use that would not encompass the transformative uses made by Defendants' MDP and would require that I terminate this invaluable contribution to the progress of science and cultivation of the arts that at the same time effectuates the ideals espoused by the Americans with Disabilities Act."

   This verdict was upheld by the second circuit on appeal[3]. No court cases

[2] Authors Guild, Inc. v. HathiTrust, 902 F. Supp. 2d 445 - Dist. Court, SD New York, 2012

[3] Authors Guild, Inc. v. HathiTrust, Case No. 12-4547-cv (2d Cir. 2014)

directly concerning academic content mining are known to the authors.

3. Non-commercial research (EU)

   In addition to the specific exception for text mining, EU member states permit copying for non-commercial research or private study. Whilst not explicitly including content mining in the broad definition involved, the legislation may be sufficiently broadly worded that researchers in EU member states may wish to consider undertaking content mining on the basis of their exception.

At the time of going to press, Eire was considering introducing a similar mining exception to that of the UK, and at the same time, the European Union itself is thinking about the possibility of introducing a Directive on the topic. If passed, which would be a lengthy process, this would impose an obligation on all EU member states to introduce similar legislation in their national copyright laws. The chances of such an exception being adopted by the World Intellectual Property Organisation, the UN special agency with responsibility for copyright laws world-wide, in the foreseeable future are very low.

To conclude on exceptions, checking local copyright law and considering collaborating with partners in countries with more permissive copyright laws is strongly recommended.

The extent to which the results of content mining analysis are protectable under copyright is a grey area; the majority of outputs will be facts or reconstructions of data, e.g., chemical structures will be re-rendered rather than a direct copy of the publisher's image being republished. However, in some cases one may wish to re-publish images or excerpts of text, in which case further consideration of copyright implications will be necessary.

2.2 Database rights as applied to content mining

In Europe, the 1996 EU Database Directive [Directive 96/9/EC] grants so-called *sui generis* rights to those who make a substantial investment in a database through collecting, verifying or presenting the contents. This results in a somewhat complex set of rights associated with collections of data. Copyright, with all the rules associated with it as explained above, applies to a database only if there is sufficient creativity involved in the presentation or arrangement of the data collection. Quite separately, database rights apply if the substantial investment described above has been expended. Note that "collecting" does NOT mean "creating from scratch", but rather 'collecting from somewhere else". Thus, if someone undertakes a series of experiments to find the melting points of compounds that have never been synthesised before, that collection of data, counter-intuitively, will not enjoy database right. But if the person had scanned the literature and collected a set of melting point data from a variety of sources, that collection would enjoy database rights. Note that an EU-based database is perfectly capable of enjoying both copyright and database rights. As copyright is a much stronger right than database right in terms of length of protection and the rights conferred upon the owner, if a database does enjoy both copyright and database rights, then the database right protection can be ignored by both owner and users. Finally, a very few databases enjoy no protection at all if they are not subject to either copyright or database right.

In the case where the database only has database rights, extraction or reuse of the entire collection or a substantial subset would require permission. Continuous extraction and reuse of database components, e.g., running a continuous query via an API to extract updated records, also requires permission. The copyright exceptions described in Section 2.1 above do not apply to a database, which only has database rights.

It should be noted that individual items of data do not enjoy any protection at all and can be reproduced freely.

2.3 Contractual restrictions on content mining

As has been already noted, in the absence of the use of an appropriate exception to copyright, the only way a content miner can lawfully undertake their research is by getting permission to do so from the copyright owner. In some cases, e.g., Open Access materials, the Creative Commons licence permits, at no charge, content mining subject to certain conditions, depending on the particular Creative Commons licence adopted. Details of these licence terms can be found at the Creative Commons web site[4]. In the case of scholarly content available from a commercial supplier on a subscription basis, charges will usually be made unless mining permissions are bundled into the subscription already, and there will be strict conditions to follow. Restrictions placed on subscribers to content via publisher licence agreements are usually restrictive, with many, for example, banning the use of all automated search and index software (Table 2.3). However, most Open Access publishers have permissive policies in terms of use of their sites for text mining in addition to permissive licensing (Table 2).

Since the changes to UK copyright law, certain publishers, including Elsevier, have announced they are willing to let researchers undertake text mining of their materials so long as the researcher uses an API developed by the publisher and so long as the researcher signs a contract which defines what they can do in terms of mining and what they can do with the results. Such a contractual term can under the new law be ignored, but what if, despite that, the publisher restricts the ability of the researcher to content mine, e.g., by deliberately slowing the mining procedure for those who haven't signed up? UK copyright law includes a provision for making a complaint about technical restrictions preventing a bona fide user from enjoying an exception to copyright, but the complaints procedure is extremely convoluted and is hardly used for that reason.

Unfortunately, some user licence agreements are specific to institutions and may not be available online, so one would need to check with one's institutional library if permission has already been granted before approaching publishers individually. To add to the complications, the publisher may not own the copyright to every item in it's collection, and not all publishers have developed clear content mining policies, and so may take a long time to respond to a request for permission to mine.

[4] Available at:`http://creativecommons.org/licenses/`. Accessed on 18 September 2014.
[5] Available at: `http://www.biomedcentral.com/about/datamining`
[6] Available at: `http://elifesciences.org/terms-and-conditions-of-use`
[7] Available at: `http://f1000research.com/about/legal/termsandconditions`
[8] Available at: `http://www.hindawi.com/corpus/`
[9] Available at: `http://blogs.plos.org/opens/2014/03/09/best-practice-enabling-content-mining/`

Publisher	Mining explicitly prohibited?	API Only?	Quote from standard licence agreement
ACS	Yes	NA	Grantee acknowledges that ACS may prevent Grantee, its Authorized Users and Other Users from using, implementing, or authorizing use of any computerized or automated tool or application to search, index, test, or otherwise obtain information from ACS Products (including without limitation any "spidering" or web crawler application) that has a detrimental impact on the use of the services under this Agreement.[a]
BMJ	Unclear	NA	Systematic downloading is prohibited.[b]
Elsevier	No	Yes	Researchers can text mine subscribed content on ScienceDirect for non-commercial purposes, via the ScienceDirect API's. Text and data mining enabling clauses for non-commercial purposes will be included in all new ScienceDirect subscription agreements and upon renewal for existing customers. Librarians interested in adding the TDM clause to their existing agreement prior to renewal are able to request a simple contract e-amendment via their Elsevier Account Manager.[c]
JSTOR	Yes	Yes[d]	Institutions and users may not: (d) undertake any activity such as the use of computer programs that automatically download or export Content, commonly known as web robots, spiders, crawlers, wanderers or accelerators that may interfere with, disrupt or otherwise burden the JSTOR server(s) or any third-party server(s) being used or accessed in connection with JSTOR.[e]
Nature	Yes	NA	1) Author deposited manuscripts: Users may view, print, copy, download and text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use. 2) NPG Material: the Licensee warrants that it will not:... (j) make mass, automated or systematic extractions from or hard copy storage of the Licensed Material.[f]
OUP	Yes	NA	The Licensee and Authorised Users may not: 2.3.2 systematically make printed or electronic copies of multiple portions of the LicensedWork(s) for any purpose.[g]
Royal Society	No	No	The Licensee may: Use Text and Data Mining (TDM) technologies to derive information from the Licensed Materials meaning: Download, extract and index information from the Licensed Materials to which the Authorized User has access under this License. Where required, mount, load and integrate the results on a server used for the Authorized User's text-mining system and evaluate and interpret the Text and Data Mining Output for access and use by Authorized Users. The Authorized User shall ensure compliance with Publisher's Usage policies. Text and data mining may be undertaken on either locally loaded Licensed Materials or as mutually agreed. Electronic copies of the Licensed Materials may be locally stored for this purpose only during the lifetime of any TDM project.[h]
Springer	No	No	Individual researchers are encouraged to download subscription and open access content for TDM purposes directly from the SpringerLink platform. No registration or API key is required. Full-text content can be accessed easily and programmatically at friendly URLs based on the content's Digital Object Identifier (DOI).[i]
Taylor and Francis	Yes	NA	Licensee must not: 8.3.2 use automated retrieval devices (such as so called web robots, wanderers, crawlers, spiders or similar devices).[j]
Wiley-Blackwell	Unclear	NA	Except as provided above or in any applicable Open Access License(s), Authorized Users may not copy, distribute, transmit or otherwise reproduce, sell or resell material from Electronic Products (s); store such material in any form or medium in a retrieval system; download and/or store an entire issue of a Electronic Product or its equivalent; or transmit such material, directly or indirectly, for use in any paid service such as document delivery or list serve, or for use by any information brokerage or for systematic distribution, whether or not for commercial or non-profit use or for a fee or free of charge.[k]

Table 1: Permissions for machine access to content in licensing agreements from a selection of major 'traditional' publishers and where applicable whether that access is provided via publisher API only. Correct as of 18 September 2014.

[a] Available at: `http://pubs.acs.org/userimages/ContentEditor/1367593694540/ACS_Institutional_Access_Agreement_Academic.pdf`

[b] Available at: `http://www.bmj.com/company/single-institution-license/`

[c] Available at: `http://www.elsevier.com/about/policies/content-mining-policies`

[d] Data mining for research is available only via a JSTOR provided user interface with fixed tools, not an API. Available at: `http://about.jstor.org/service/data-for-research`

[e] Available at: `http://www.jstor.org/page/info/about/policies/terms.jsp`

[f] Available at: `http://www.nature.com/libraries/site_licenses/license_agreements.html`

[g] Available at: `http://www.oxfordjournals.org/en/help/instsitelicence.pdf`

[h] Available at: `http://royalsocietypublishing.org/text-data-mining` and `http://royalsocietypublishing.org/text-data-mining`

[i] Available at: `http://www.springer.com/gb/rights-permissions/springer-s-text-and-data-mining-policy/29056`

[j] Available at: `http://www.tandf.co.uk/libsite/pdf/licensingInfo/TermsAndConditions.pdf`

[k] Available at: `http://onlinelibrary.wiley.com/termsAndConditions`

Publisher	Mining explicitly prohibited?	API Only?
Biomed Central	No[5]	No
eLife	No[6]	No
F1000 Research	No[7]	No
Hindawi	No[8]	No
PLOS	No[9]	No

Table 2: Permissions for machine access to content in licensing agreements from a selection of major Open Access publishers and where applicable whether that access is provided via publisher API only. Correct as of 18 September 2014.

It may therefore be worth considering collaborating with researchers at institutions with more permissive licences, or with researchers based in countries with clear exceptions for content mining embedded into their copyright law.

2.4 Practical advice for obtaining permissions

Where permission does need to be obtained from the copyright owner, the following points are worth considering:

1. Obtaining individual permissions from relevant publishers is a time-consuming and often unsuccessful process.

2. Services are in development to streamline the process e.g., in the UK, an interesting initiative called The Copyright Hub[10] was launched in summer 2013 to streamline licence permission requests. However, progress on the Hub has been painfully slow, very few copyright owners are connected to it, and content mining does not appear to have been considered as a possible use of copyright materials.

3. We recommend that any agreement with a publisher to content mine should only permit restrictions on the amount of, or speed of content download that are "in the reasonable opinion of the publisher" necessary to protect its commercial interests and technical performance. It should not restrict the licensee to use the publisher's API, but should permit any reasonably efficient API to undertake the work. If the publisher refuses to negotiate on these points, the researcher should seriously consider alternative approaches.

4. Open Access content which is licensed as CC-BY-NC or an equivalent, or more permissive licence can be used without risk of copyright infringement but might still fall under contractual agreements with regard to reasonable restrictions on the use of publisher servers and download of materials.

5. Explaining your intended work well may increase the chance of permission being granted.

[10] Available at: `http://www.copyrighthub.co.uk/get-permission`. Accessed on 18 September 2014

Case Studies: Obtaining permission for text mining

Text-mining at the University of British Columbia
Heather Piwowar wanted programmatic access to Elsevier journals for her research on how scientists use, reuse and cite data. During 2012 she extensively documented her attempt to gain permission, which started with a Tweet that got picked up by Elsevier's Director of Universal Access and went on to involve meetings with six Elsevier employees and Piwowar's institutional librarians [Piwowar, 2012a]. Permission was eventually granted for the purposes of:

- Direct analysis for research
- Selection of excerpts for citizen science
- Calculating statistics on the usage of research objects for open dissemination in research tools.

[Piwowar, 2012b]

Importantly, this agreement was made available for others to read unlike many subscription publisher terms of agreement which are subject to non-disclosure agreements. Initially permission was granted only for Piwowar's research but this was soon extended to cover the whole university. While a success in several respects, Piwowar considers this process to be unscaleable across the many hundreds of research projects which could benefit from access to the literature for text mining. In addition Elsevier is only one of many publishers with content that could strengthen data citation research results.

Piwowar advises researchers to talk to librarians at their institutions to find out about license terms and conditions and the process for their negotiation. She also encourages librarians to take a pro-active stance on negotiating for text mining rights and providing researchers with information and training on its potential [SPARC].

text2genome project
Maximilian Haeussler, Casey Bergman and the text2genome project have documented their progress[a] in obtaining permission to mine full-text articles in order to annotate the UCSC Genome Browser with links out to relevant publications. So far 27 positive responses have been received from a total of 46 requests since 2009 and crawling is underway on the available corpus. Bergman reports that:

> "it takes six months to two years per publisher to get a final written agreement to mine and release extracted content, and often when pushed these deals fall through on the issue of releasing extracted content." [SPARC].

The documentation of approaches made to publishers includes copies of default end user license agreements and UCSC library agreements where available. Interestingly, in at least one case the library license agreement

allows text mining but permission has still been refused. This demonstrates the importance of checking permissions but also the extent to which copyright exemptions as enacted in the UK and Japan can save researchers valuable time in a process which is currently fraught with uncertainty and requires extensive communication and follow-up with individual publishers.

[a] Available at: `http://text.soe.ucsc.edu/progress.html`. Accessed 18 September 2014

3 Responsible Crawling

The rest of this chapter assumes that the content miner has legal permission to mine for the purposes of their research and now discusses the resposibilities of exercising your right to mine.

3.1 Understanding the impact of crawling

Gathering content for the purposes of content mining requires finding and copying articles and other content of interest on the web. The software agent performing this task is known as a crawler and will visit a list of seed URLs, followed by additional URLs harvested during the crawling process, e.g., all those on a journal issue's contents page. It is important that web crawling is perfomed ethically to minimise any costs, disruption and privacy or security concerns (reviewed in Thelwall and Stuart [2006]).

Crawlers can send multiple requests per second to the same server and potentially download large files, particularly if the content of interest is in PDF format. It is therefore not surprising that publishers are concerned about the potential impact of uninhibited crawling activities on the performance of their servers, which already deal with multiple requests from subscribers. Crawlers require considerable bandwidth and poorly written crawlers even more so, which is one of several reasons that some publishers, such as Elsevier, attempt to require content miners to use their own approved APIs, as well as imposing maxima on what may be downloaded. Download speed should not be a major problem for big international publishers, who have to deal with dozens of requests per second during the whole year. Smaller publishers, for which one single crawler can affect the performance of the website, have more of a problem.

Another issue are the access logs that publishers keep for libraries. For each subscription and journal, the librarians can find out from the publisher how many articles were accessed by users and in which journals. A crawler that retrieves all articles from a journal can significantly alter the access statistics. We suggest that anyone wishing to content mine use a defined useragent string in all HTTP requests such that publishers can distinguish crawlers from normal requests.

The "useragent" of a web request is the name of the internet browser that is transmitted when a page is requested from a web server. An example is "Mozilla/5.0 (iPad; U)" for a web request from an IPad. Google's crawlers are usually excluded from any access logs, as they can be identified by IP address or their "useragent". We suggest using a useragent string that clearly indicates that the web request comes from a crawler like "GoogleBot" or "TDMCrawler"

and adding in parenthesis contact information of a person that can stop the crawler in case problems occur.

There exists a robots exclusion protocol, which indicates which parts of a web server must not be accessed by crawlers, but as publishers want their content to be indexed by search engines such as Google, these are most often not implemented.

Crawling OA Content: Journal Perspectives

> "As an OA publisher we are certainly happy for people to use our published content for text mining purposes. As you mentioned, there is some concern about the load that this may place on our web servers, so we are in the process of setting up an FTP site where researchers will be able to download the XMLs of all of our published articles for text mining purposes, which should help reduce the load on our servers.
>
> Practically speaking, the impact of text mining on our servers has not yet become a problem, and I'm sure that the load from search engine spiders is a lot higher than from researchers trying to text mine our content. Also, given that we host all of published articles using Amazon's Simple Storage Service (S3) my guess is that it would take quite a few researchers doing text mining on our content at the same time to cause any real problems."

Paul Peters, Head of Business Development, Hindawi Publishing Corporation [McDonald et al., 2012]

> "We haven't had any problem with server load performance from robots text mining the journal sites. Previously, there were occasions that site performance would degrade from what appeared to be out-of-control scripts hitting a single article. In that case, we would block the IP of the script. But since we implemented the new software, we haven't seen this problem come up. There are no restrictions for text mining our content other than respecting the crawl- delay in robots.txt (currently set to 30 seconds) and fetching content from one journal at a time."

Public Library of Science (PLOS) [McDonald et al., 2012]

3.2 Respecting crawler limits

Publishers will often impose crawling limits to protect their technological infrastructure and presently these can extend to complete blocks on any crawling software (Table 3). Going above crawling limits is likely to lead to publisher action to block your IP address and potentially cut off institutional access if they feel the service is being abused. In most cases, delays of 10-20 seconds between two consecutive requests should be sufficient to avoid overloading the publisher's webserver. Respecting current crawler limits is severely limiting to text miners,

as they are typically set low to discourage such usage of their content

Publisher or Platform	Crawler Restrictions	Reference
Highwire	Weekday 10 s between requests, Weekend 5 s. No crawling during US East Coast work hours.	UCSC
Wiley	1 s. IP address needs to be authorised by publisher	UCSC
Elsevier	Need to sign license and go through Elsevier ConSyn. 20 s between requests	UCSC
Springer	Require FTP access and 20 s between requests	UCSC
PLOS	30 s between requests	Neylon [2014]
Biomed Central	1 s between requests	Neylon [2014]
eLife	10 s between requests	Neylon [2014]

Table 3: Example crawler limits for a range of publishers from which UCSC has obtained permission to text mine or which allow crawling by default (e.g. Open Access publishers PLOS, Biomed Central and eLife) and have published limits.

3.3 Choosing and configuring crawler software

Writing code that respects crawler limits but remains on a site without causing problems for the server is challenging. There are existing pieces of software and we recommend that new text miners make use of these before attempting to write their own crawler code. Often publishers will block the IP address of crawlers that place too high a load on their servers so the following guidelines should be followed for responsible crawling and mining.

1. Email each publisher to notify it of the proposed crawling activity, including the crawler IP address and the date(s) and time(s) when crawling is most likely to take place.
2. Attempt, wherever possible, to crawl publisher websites outside the working hours of the publisher timezone.
3. Keep delays to 10-20 seconds between requests.
4. Set the useragent to TDMCrawler, adding contact and project information.

4 Publication of Results

Publishing the results of a content mining project requires decisions regarding the availability, presentation and format of the mined data. As per all research

outputs, the limitations of your methods should be described and potential error rates presented.

4.1 Access and Licensing

Many research funders now mandate, or are introducing requirements for open publication of research data, and it is considered good practice to store the output of your analysis publicly using recognised repositories or other storage with persistent identifiers and distributed copies, ideally adhering to guidelines such as the Panton Principles for Open Data in Science [Murray-Rust et al., 2010]. Data licensing is a complex area, but it is generally considered that as facts are not copyrightable and there can be legal interoperability issues between differently licensed datasets, a copyright waiver or public domain licence such as the CC0 waiver[11] or the Public Domain Dedication License[12] are most appropriate for scientific data.

One area where text mining may raise copyright problems is where one wishes to publish verbatim results, either because this is the unit of data you are interested in or it is necessary for contextualisation or illustrative purposes. In this case, we recommend publishing no more than around 200 words in the case of text, although it must be stressed that what is considered to be copying outside what the country's copyright law limits for an exception for research to apply varies from case to case, and so there is no set number of words that are "safe". For this reason, it is safest to discuss the matter with the copyright owner(s) of the original works that had been mined before submitting the results of your mining activity for publication.

If a publisher takes issue with your publication of content derived from mining their articles, it is helpful to provide a contact address and notice and take-down procedure clearly stated on your website or repository entry and to document their objections and obtain legal advice where possible. The Jorum notice and takedown policy[13], which is available under a Creative Commons licence and can be edited to suit your requirements, is recommended.

4.2 Downstream Use

The majority of the end-user licencse agreements, which allow content mining and the copyright exemptions proposed exceptions adopted by the UK government apply only to non-commercial research, despite several arguments for not making this distinction [Hagedorn et al., 2011, Klimpel, 2012]. It is unclear to what extent 'non-commercial' applies to typical downstream uses of the data arising from academic mining projects. In the UK, the Intellectual Property Office, the Government agency responsible for handling copyright law, understandably refuses to give an opinion on the matter, as this is something for Courts to decide. There has been some research on what the general population defines as non-commercial (e.g. Commons [2008]) but confusion remains [Klimpel, 2012].

[11] Available at: `http://creativecommons.org/publicdomain/zero/1.0/`. Accessed on 18 September 2014

[12] Available at: `http://opendatacommons.org/licenses/pddl/`. Accessed on 18 September 2014.

[13] Available at: `http://www.jorum.ac.uk/policies/jorum-notice-and-takedown-policy`. Accessed on 18 September 2014.

Some downstream activities, such as making money from the sales of reports of the work, are clearly 'commercial', but others, such as using the results in a lecture to students, are unlikely to be. What is important to note is that the restriction to non-commercial refers to the researcher's plans at the time of copying. Thus, if completely unexpectedly, a commercially valuable result is obtained, there is not a problem with the original mining activity. If, on the other hand, it could reasonably be surmised that something commercially valuable would result, then the mining activity will be deemed to have been 'commercial'.

5 Citation and Acknowledgement

Responsible use of text mining technology includes citing the articles and datasets mined as per community norms for reuse of scholarship. This raises considerable technical issues in the text mining workflow as to how to record each source and maintain the association of mined content to source throughout the analysis, especially as some mined resources will may not be used in the analysis.

There are no current established norms for citation in content mining projects and it is clear that applying the norms of lower throughput scholarship, where an article may reference the work of 100 other works or less, will in many cases be unpragmatic and extremely technically challenging in the face of thousands of potential sources.

Ironically, many tools which allow better tracking of articles and datasets as well as the generation of alternative metrics rely on programmatic access to the literature themselves [Piwowar, 2012a]. We encourage all potential content miners to explore their options for attribution and where possible to assist the research community in building tools for this purpose.

6 Proposed best practise guidelines for content mining

We present a list of principles to which we hope both text miners and publishers can ascribe, ensuring that both parties are able to make use of text mining technology to conduct fruitful research without detriment to content providers. We hope that increased uptake of content mining through greater clarity of the rights of researchers (The right to read is the right to mine) and their responsibilities undertaking these analyses (The Content Mine) help unlock the potential of content mining as a technique to discover more about the world by using the knowledge we have already accumulated in the scholarly literature to its full potential.

Responsible Content Mining Code

1. **Don't break the law**
   (a) Honour copyright as you understand it and consult about current interpretations in your jurisdiction.
   (b) If there is no copyright exemption for content mining in your

country, consult your institutional librarians for the terms of your licensing contracts with publishers that do not explicitly permit mining.

(c) Be aware of additional legal permissions required for mining with intended commercial use of results.

2. **Don't break servers or services**

(a) Set acceptable delays between each crawl.

(b) Try not to recrawl and use public repositories of crawled or submitted materials where they exist and allow this.

(c) Avoid corrupting content in the crawling process.

3. **Be visible and polite**

(a) Use a defined useragent string in all HTTP requests that clearly identifies you as a crawler and provide contact details.

(b) If you are using any subscription material inform your library and the publisher of your proposed crawling.

4. **Work with other content miners**

(a) Consult publicly online about current good practice before starting.

(b) Use *de facto* standard tools (only write your own if there's a gap).

5. **Give credit where credit is due**

(a) Credit original producers of mined research outputs whenever possible, as per community norms for the reuse of scholarship.

The right to read is the right to mine

Principle 1: Right of Legitimate Accessors to Mine

We assert that there is no legal, ethical or moral reason to refuse to allow legitimate accessors of research content (OA or otherwise) to use machines to analyse the published output of the research community. Researchers expect to access and process the full content of the research literature with their computer programs and should be able to use their machines as they use their eyes. The right to read is the right to mine

Principle 2: Lightweight Processing Terms and Conditions

Mining by legitimate subscribers should not be prohibited by contractual or other legal barriers. Publishers should add clarifying language in theirsubscription agreements that content is available for information mining by download or by remote access. Where access is through researcher-provided tools, no further cost should be required. Publishers should always

explain to subscribers in countries that have implemented an exception to copyright for text and data mining that this exception exists, and should also ensure that in those countries they will not attempt to side-step the exception by adding terms or conditions, or technical barriers, to restrict what subscribers are entitled to do under the law. Users and providers should encourage machine- processing.

Principle 3: Technical restrictions

Bona fide content mining should not be restricted by unreasonable or unjustified technical restrictions imposed by publisher servers.

Principle 4: Agree what is commercial and what is non-commercial

Publishers should make clear by means of use cases linked to their licences what sorts of downstream activities they reasonably consider to be commercial, and what activities they consider to be non-commercial.

Principle 5: Use of Mining Results

Researchers can and will publish facts and excerpts which they discover by reading and processing documents. They expect to disseminate and aggregate statistical results as facts and context text as fair use excerpts, openly and with no restrictions other than attribution. Publisher efforts to claim rights in the results of mining further retard the advancement of science by making those results less available to the research community.; Such claims should be prohibited. Facts don't belong to anyone.

References

Creative Commons. Defining 'noncommercial': a study of how the online population understands 'noncommercial use'. *Creative Commons Wiki*, 2008.

Directive 96/9/EC. Directive 96/9/ec of the european parliament and of the council of 11 march 1996 on the legal protection of databases. `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML`, 1996. Accessed: 18 Sep 2014.

Gregor Hagedorn, Daniel Mietchen, Robert A Morris, Donat Agosti, Lyubomir Penev, Walter G Berendsohn, and Donald Hobern. Creative commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information. *ZooKeys*, (150):127, 2011.

Ian Hargreaves. Digital opportunity: a review of intellectual property and growth: an independent report. 2011.

Puneet Kishor. Legal implications of text and data mining (tdm). Presented at Open Knowldge Festival 2014, Berlin, 15 - 17 July 2014, 2014. Accessed: 18 Sep 2014.

Paul Klimpel. Consequences, risks, and side-effects of the license module non-commercial – nc. Technical report, 2012. Accessed: 18 Sep 2014.

LACA. Independent review of intellectual property and growth response by laca: the libraries and archives copyright alliance march 2011. Technical report, 2011. Accessed: 18 Sep 2014.

D McDonald, I McNicoll, G Weir, T Reimer, J Redfearn, N Jacobs, and R Bruce. The value and benefits of text mining. Technical report, 2012.

P Murray-Rust, J Molloy, and D Cabell. Open content mining. In *The First OpenForum Academy Conference Proceedings*, pages 57–64. OpenForum Europe LTD, 2012.

Peter Murray-Rust, Cameron Neylon, Rufus Pollock, and John Wilbanks. Panton principles: principles for open data in science. *Panton Principles*, 2010.

Cameron Neylon. Best practice in enabling content mining. `http://blogs.plos.org/opens/2014/03/09/best-practice-enabling-content-mining/`, 2014.

H Piwowar. Data citation and text mining. `http://researchremix.wordpress.com/2012/04/19/data-citation-text-mining/`, 2012a. Accessed: 18 Sep 2014.

H Piwowar. Elsevier agrees ubc researchers can text-mine for citizen science, research tools. `http://researchremix.wordpress.com/2012/04/17/elsevier-agrees/`, 2012b. Accessed: 18 Sep 2014.

SPARC. Pushing the frontier of access for text mining: A conversation with heather piwowar on one researcher's attempt to break new ground. `http://www.sparc.arl.org/news/pushing-frontier-access-text-mining-conversation-heather-piwowar-one-researcher%E2%80%99s-attempt-break`, 2012. Accessed: 18 Sep 2014.

The Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014. `http://www.legislation.gov.uk/uksi/2014/1372/regulation/3/made`, 2012. No. 1372. Regulation 3.

Mike Thelwall and David Stuart. Web crawling ethics revisited: Cost, privacy, and denial of service. *Journal of the American Society for Information Science and Technology*, 57(13):1771–1779, 2006.

TRIPS. `http://www.wto.org/english/docs_e/legal_e/legal_e.htm#TRIPs`, 1994. Accessed: 18 Sep 2014.

