Briefing Paper

Text and Data Mining and the Need for a Science-friendly EU Copyright Reform

April 2015

SCIENCE EUROPE
Shaping the future of research

Abstract

The steadily-growing amount of digitally-available research data and publications enables researchers to search and analyse these sources with the help of special software. The application of such text and data (content) mining techniques (TDM) is not limited to research. In fact, most users of the internet use them on a daily basis via companies offering search engine services. The use of TDM techniques beyond those employed by search engines is already of great importance in some research fields (for example bio-genetics, linguistics) and interest in these technologies is growing rapidly. The publishing industry – including academic publishers – strives to benefit by developing TDM services, but, in doing so, hinders the ability of third parties (subscribers) to mine their content. This has led the research community advocating a reform of copyright law to ensure that legally-accessed content can be freely mined without additional permission and cost. After years of debate, the European Commission and European Parliament appear to be ready to amend the European Union (EU) Copyright Directive. This paper gives a brief overview of the legal regulations under which TDM practices fall and the issues arising from increasing use of licensing by publishers.

Contents

1 ‘Text and Data Mining’ or ‘Content Mining’?

The term ‘text and data mining’ (TDM) is slightly misleading as it refers to a wide range of content, for example data, audio-visual material, texts and corresponding metadata, on which a wide range of computerised searching, analysis and integration techniques are used. Therefore the term ‘content mining’ is more appropriate than ‘text and data mining’. Nonetheless in this paper the term ‘text and data mining’ will be used because it is established and recognised; however, it is intended to encompass the wider meaning of the term ‘content mining’.

2 Benefits of TDM Techniques

TDM techniques are indispensable for researchers, who need to access a huge number of publications and a huge amount of research data, even just in their own fields of research. Because of this, the term TDM is often used in the context of ‘Big Data’. Obviously, large amounts of data cannot be analysed without computers. However, in many respects this is also true for smaller data volumes, for instance searching for a term in a newly-published issue of a journal.

TDM techniques are not only used to speed up processes that were previously done manually; they also enable new methods, which are sometimes referred to as ‘data-driven science’. For example, specially-developed software tools for specific semantic analyses can be used as part of TDM.

Thus, TDM is regarded as a driver for improving the performance of all sciences, including social sciences, arts and humanities. This applies both to attaining insight that cannot be gained without the use of TDM techniques, and to the acceleration of research processes by building on previous work.

Interest in the use of TDM techniques is not limited to researchers. More powerful (in the sense of an extended ability to gain insight) and more efficient research processes promise to yield more innovation that will ensure economic growth.

Increased amounts of freely-available or legally-accessible content will foster the growth of new service providers to help researchers, and others, to mine content. Legal hurdles, which currently enable the publishing industry to block these newcomers to the market, impede innovation and increase prices; this is not in the interest of the European economies.

3 Uses of TDM Techniques: Indirect or Autonomous

Virtually all researchers use TDM techniques indirectly via search engines or bibliographic databases, often without being aware that these are TDM services. As mentioned above, TDM can go far beyond the use of search engines. An increasing number of researchers autonomously operate TDM techniques and therefore need extensive access to research data and publications under conditions allowing autonomous TDM.

4 Pre-conditions for TDM

The use of TDM has several pre-conditions. Firstly, and obviously, there is the need to be able to physically access and, depending on the technologies that are to be used, locally store the content that is to be mined. In many cases the benefit of TDM is linked to the amount of content that can be mined. Often it is necessary to collect content from a large number of different sources.

Physical access needs to be accompanied by appropriate legal conditions. This paper focuses on copyright law but other legal aspects, such as data protection law, may also apply.

The ability of researchers to mine copyright-protected works – especially research publications – is greatly hindered by many publishers. Therefore, researchers often mine abstracts rather than whole works. It can be extremely difficult for researchers to ascertain what uses are legal due to the fact that copyright law is very arcane. In addition, content that is to be mined often stems from several sources with different copyright conditions.

Depending on the quality of the content and on the techniques that are to be used, refinement and normalisation of the collected content is necessary before the actual mining process can start. Content may need to be converted into a common format. Variations in terminology need to be identified in order to perform successful searches. Content from various sources may be brought together in one file. The content may be treated in several processes, each involving TDM techniques that depend on the outcome of previous mining. The outcome of the consecutive analyses will form part of the metadata[1] of that set of content. For example, the output of a Google search is the result of a search through previously-produced indexes which are based on information harvested from websites. The efficiency of these searches is increased by the use of corresponding algorithms.

5 Legal Claims of Rights Holders

Rights holders, such as academic publishers, base their claim to have a right to grant or refuse the mining of their works on copyright law (5.1), EU database protection (5.2) and licenses, for example as part of subscription contracts (5.3).

The term 'copyright' in this paper refers to European copyright law which is based on two directives:

- Directive 1996/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases (Official Journal L 077, 27/03/1996 P. 20-28).
- Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (Official Journal L 167, 22/06/2001 P. 10-19).

Both directives are applicable via the national law of the EU Member States. European copyright law is not fully harmonised by these two directives as they leave scope for national lawmakers to form their law. This paper does not address the differences amongst the national copyright laws of the EU Member States.

5.1 Copyright Law

The claim of rights holders to control TDM based on copyright law requires explanation, because facts and figures, the object of content mining, do not enjoy copyright protection. Similarly, the act of reading, on which content mining is based, is not protected by copyright.

Copyright law grants authors control over a number of concretely-listed forms of exploitation rights, such as the reproduction right, the right of communication to the public of works, and the right of making available to the public other subject-matter and the distribution right (Ref. Art. 1 to 4 Dir. 2001/29/EC, see Art. 1 below). TDM is not an exploitation right protected by any national copyright law of EU Member States or multilateral copyright treaty.

Directive 2001/29/EC
Chapter I 'Objective and Scope'
Article 1 'Scope'

1. This Directive concerns the legal protection of copyright and related rights in the framework of the internal market, with particular emphasis on the information society.

2. Except in the cases referred to in Article 11, this Directive shall leave intact and shall in no way affect existing community provisions relating to:
   (a) the legal protection of computer programs;
   (b) rental right, lending right and certain rights related to copyright in the field of intellectual property;
   (c) copyright and related rights applicable to broadcasting of programmes by satellite and cable re-transmission;
   (d) the term of protection of copyright and certain related rights;
   (e) the legal protection of databases.

Rights holders base their quest to control the TDM of their works on the reproduction right which grants rights holders "the exclusive right to authorise or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part" (Ref. Art. 2 Dir. 2001/29/EC). They claim that this right is infringed by electronic copies that are created as part of the content mining process. Based on this notion, researchers intending to mine works protected by copyright would need either a corresponding license from the right holders of the works or an appropriate copyright exception.

While the EU copyright directive (2001/29/EC) does protect the reproduction right it also includes a mandatory exception with regards to temporary acts of reproduction; EU Member States have to include this exception in their national copyright law (See below Art. 5 Sec. 1 Dir. 2001/29/EC).

Directive 2001/29/EC
Chapter II 'Rights and Exceptions'
Article 5 'Exceptions and Limitations'

1. Temporary acts of reproduction referred to in Article 2, which are transient or incidental [and] an integral and essential part of a technological process and whose sole purpose is to enable:
   (a) a transmission in a network between third parties by an intermediary, or
   (b) a lawful use
   of a work or other subject-matter to be made, and which have no independent economic significance, shall be exempted from the reproduction right provided for in Article 2.

Based on Art. 5 Sec. 1 Dir. 2001/29/EC, the mining of legally-accessed, copyright-protected works which does not involve the making of permanent copies arguably does not infringe upon the right of reproduction. However, this assertion is only true if the license, on which access to the content is based, does not contain language restricting or ruling out storage or mining of the content. Provisions in subscription contracts limiting the amount of content that can be downloaded per online session can be another hurdle.

Mining of copyright-protected works which involves the making of permanent reproductions is not covered by Art. 5 Sec. 1 Dir. 2001/29/EC. All other exceptions listed in Art. 5 Dir. 2001/29/EC are not mandatory and do not explicitly refer to TDM. Currently, only the copyright law of the UK contains an exception for TDM as part of a privileged use of protected works for non-commercial research[2].

Operators of search engines are in the same position as researchers in respect to their ability to mine copyright-protected works. The fact that rights holders tolerate or explicitly license search engines to harvest and mine their content reflects their interest in traffic to their websites and their dependence on the search engines for this.

5.2 EU Database Protection

Content to be mined is often accessed via a database. In the EU, databases are protected by the Directive 1996/9/EC, which stipulates that only small fractions of the content of databases may be extracted without the consent of the database owner. This rule applies independently of possible copyright protection of the database content. According to this Directive, downloading a substantial part of the database content needs to be authorised by the database owner, even if none of the content is under copyright.

This special protection of databases via an ancillary copyright is a European peculiarity. For researchers, it constitutes another hurdle on top of copyright protection restricting access to content.

5.3 Licenses

Electronic access to research publications, especially access to big publication sets, is typically based on subscription contracts. These contracts can contain language restricting TDM. Additionally, systematic download of the subscribed content is often strictly limited. Recently, publishers have started to introduce licenses regulating the mining of subscribed content more specifically. These licenses are a combination of permissions and prohibitions. The subscribing party risks relinquishing leeway granted by statutory law when accepting such a license.

In practical terms, the necessity of obtaining a license not only implies additional costs but also the risks that licenses are not granted, or use of the result of mining is restricted. Furthermore, the possibility of obtaining a license is dependent on information about whether content that is to be included in a content mining endeavour is protected by copyright and who the copyright owner is. In many cases it will be impossible or disproportionately expensive to answer these questions and obtain the license.

Licenses, coupled with subscription contracts, limit TDM to the subscribed content. Limitations like this are at odds with research endeavours for which broader collections need to be mined. In many cases the researcher will be interested in mining content that is owned by a number of rights holders. It is unrealistic to negotiate licensing terms with more than a few rights holders, and even this is often too time-consuming to be feasible. Successful negotiations with several rights holders will probably result in different licenses and this poses additional problems. Any license-based approach will seriously hinder the use of TDM. For small projects (limited duration and budget), it is too complicated to start the negotiating process; for big research projects, which aim to mine content from a large number of rights holders, there is no realistic chance of identifying all rights holders, let alone successfully concluding negotiation of licensing terms.

The licensing approach[3] adopted by some academic publishers obliges licensees to perform the actual mining on servers controlled by the publisher and to use software installed by the publisher. This not only limits the ability of the researcher to mine, it also exposes his interests

and algorithms to the publisher. In addition, the publisher regulates how the researcher can share and publish the results of that mining. Finally, from a purely scientific point of view, such a licensing approach is unacceptable because the results cannot be reproduced by the author of the original research or by his or her peers. Reproducibility is a major issue for two main reasons: access to the content is limited by the duration of the subscription period, and the composition of content may change as the publisher may sell or acquire journals.

5.4 Exceptions

In countries with a copyright law that includes a fair use or fair dealing provision[4], the requirements for a TDM exception can be accommodated by case law. The current interpretation of the fair use doctrine in the USA and Canada is assumed by many to cover TDM for non-commercial research. Japan and the UK both have a special statutory exception for TDM. The UK is the only EU Member State whose copyright law includes an explicit exception for content mining.

Besides a lack of similar exceptions in the copyright law of other EU Member States, the Copyright Directive does little to harmonise the national copyright laws. This not only leads to competitive inequality, but also poses a major barrier for international co-operation, thus contradicting the notion of a European Research Area. A mandatory exception for TDM including commercial use should be added as part of an amendment to the European Copyright Directive.

The inclusion of commercial use is important as many research projects considered to be non-commercial by the participating researchers may be categorised as commercial by lawyers. Also, there is a continuum from public research with no commercial purposes at all to public research with clear commercial purpose; setting a boundary would be artificial and make the positioning of a given research activity very difficult. Besides, research organisations are increasingly encouraged by national authorities and European institutions to establish synergies and partnership with the private sector. Limiting an exception for TDM would thwart this very aim. The development of new content mining services by commercial companies is in the interest of the research community and society in general. Limiting a TDM exception to non-commercial use will force the affected companies to move their activities to other countries outside Europe.

6 Difficulty of Assessing the Legality of Content Mining

The legal circumstances in respect of TDM are so complex that assessing the legality of the use of particular mining techniques on specific content is impossible, in practical terms, for researchers and it is unfair and unreasonable to impose such a demand on them. However, abstention from TDM is not an acceptable option from the perspective of researchers, or even society; researchers want to avoid missing out on the results which can be achieved through employing these techniques, and additionally their ‘competitors’ (for example in the UK, the USA and Japan), are able to use these technologies due to the differing legal situation.

Therefore, it seems likely that in many cases researchers use content mining techniques without fully understanding the potential legal consequences.

7 Necessary Advocacy for a Science-friendly Copyright Law

The academic community in Europe is advocating amendments to the Copyright and Database Directives that will improve and harmonise the current legal situation with respect to TDM. Rights holders have responded to this demand by claiming that copyright reform is not necessary because they will enable content mining based on standardised licenses.

Many individual research organisations will most likely have to accept these new licenses as part of the subscription contracts to journals even though they are opposing them in principle. Nevertheless, it is important that they continue to advocate for a science-friendly copyright law.

The promotion of open access publishing is another policy option because it releases content from the control of the publishers. However, the desired accessibility of research publications – including the right to mine – cannot solely be realised by open access publishing in the sense of only attaching the right license to individual publications. In addition, an infrastructure is needed to ensure the physical accessibility of the works, since licenses do not include an obligation to make the works available.

8 Suggested Way Forward for Research Organisations

Research organisations would be advised to:

Avoid requesting their researchers to abstain from mining

An abstention from mining would be contrary to the interests of individual researchers in particular and the scientific endeavour in general.

Avoid implementing greater control of employees in order to avoid nuisance/ liability of interference complaints

The risk of complaints seems low and can easily be managed. So far, no cases are known in which allegedly illegal content mining by researchers has led to legal consequences. Sporadic difficulties have occurred because of extensive downloading of articles in breach of publishers' agreed licenses (such as quotas), and, as a result, access to a database has been temporarily disabled. These quota overruns do not necessarily occur as a result of content mining. Usually these cases can be resolved informally.

Empower employees through training and provision of legal advice

Such training would aim to:

- Improve the ability of researchers to make informed decisions; and
- Develop within the organisations a robust capacity/mandate for negotiation of subscription contracts without restrictive regulations on content mining and for the support of advocacy activities.

Refuse to sign TDM-licenses as part of subscription contracts

In general, TDM-licenses, typically as part of subscription contract, should not be accepted. At least, it should be ensured that in contract negotiations with publishers any licenses on content mining are tested carefully, otherwise own regulatory proposals are preferred. The probability of the enforcement of acceptable regulations depends not only on the negotiating mandate of the representatives of scientific organisations, but also on support within their organisation. This support will be much greater if researchers were trained on the issue.

Advocate a future research-friendly copyright law at national and European level

The current EU Copyright Directive enables rights holders to control mining of their works to the detriment of research and innovative entrepreneurs. Without advocacy for change in the interest of research and innovation this situation will not change. In fact, it may become worse, as illustrated by the new ancillary copyright introduced in Germany and Spain[5]. Research organisations should therefore maintain or enhance existing activities at national and European level, advocating the necessity of broad mandatory exceptions for TDM and education and research in general.

9 Sources

Filippov, Sergey (27.V.2014): 'Mapping Text and Data Mining in Academic and Research Communities in Europe', Lisbon Council special briefing: Text and data mining, 16/2014.
www.lisboncouncil.net//index.php?option=com_downloads&id=1034

Hargreaves, Ian et al. (2014): 'Standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining', Brussels, European Commission.
www.ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf

Notes and References

[1] Metadata are explanatory information. Most commonly, metadata give very basic information about a work (e.g. author name, title). Librarians use standardised sets of metadata (e.g. 'Dublin Core') but metadata can be very complex and do not necessarily reflect certain standards. In fact, if research data are published as supporting information for an article one could also view the article as metadata for this particular set of research data. In the context of content mining these metadata could be a list of variations of a term used in a set of works which are to be mined or information about the structure of the work.

[2] The Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014 (UK legislative text):
3. Research, private study and text and data analysis for non-commercial research
29 A Copies for text and data analysis for non-commercial research:
(1) The making of a copy of a work by a person who has lawful access to the work does not infringe copyright in the work provided that—http://www.legislation.gov.uk/uksi/2014/1372/regulation/3/made
(a) the copy is made in order that a person who has lawful access to the work may carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose, and
(b) the copy is accompanied by a sufficient acknowledgement (unless this would be impossible for reasons of practicality or otherwise).
(2) Where a copy of a work has been made under this section, copyright in the work is infringed if—
(a) the copy is transferred to any other person, except where the transfer is authorised by the copyright owner, or
(b) the copy is used for any purpose other than that mentioned in subsection (1)(a), except where the use is authorised by the copyright owner.
(3) If a copy made under this section is subsequently dealt with—
(a) it is to be treated as an infringing copy for the purposes of that dealing, and
(b) if that dealing infringes copyright, it is to be treated as an infringing copy for all subsequent purposes.
(4) In subsection (3) "dealt with" means sold or let for hire, or offered or exposed for sale or hire.
(5) To the extent that a term of a contract purports to prevent or restrict the making of a copy which, by virtue of this section, would not infringe copyright, that term is unenforceable.
http://www.legislation.gov.uk/uksi/2014/1372/regulation/3/made

[3] Response to Elsevier's text and data mining policy: a LIBER discussion paper. 28 March 2014. http://libereurope.eu/wp-content/uploads/2014/04/TDMdiscussionpaper-final1.pdf

[4] Fair use is a doctrine that permits limited use of copyrighted material without acquiring permission from the rights holders. Examples of fair use include commentary, search engines, criticism, parody, news reporting, research, teaching, library archiving and scholarship.
Fair dealing is an enumerated set of possible defences against an action for infringement of an exclusive right of copyright. Unlike the doctrine of fair use, fair dealing cannot apply to any act which does not fall within one of these categories.

[5] In Germany and Spain, ancillary copyright laws have been introduced protecting online content of newspapers. These legislations were passed with the intention to force companies who operate search engines to share profits if they display search result acquired from newspaper websites. As of now these legislations have not met their initial aim and at the same time have created a legal barrier for start-ups.

Colophon

April 2015

'Text and Data Mining and the Need for a Science-friendly EU Copyright Reform': D/2015/13.324/1
Author: Science Europe Working Group on Research Data
Editor: Christoph Bruch

For further information please contact the Science Europe Office:
office@scienceeurope.org

Science Europe is a non-profit organisation based in Brussels representing major Research Funding and Research Performing Organisations across Europe.

More information on its mission and activities is provided at: www.scienceeurope.org.

To contact Science Europe, email office@scienceeurope.org.

Science Europe
Rue de la Science 14
1040 Brussels
Belgium

Tel +32 (0)2 226 03 00
Fax +32 (0)2 226 03 01
office@scienceeurope.org
www.scienceeurope.org

SCIENCE
EUROPE
Shaping the future of research