Submitted on: 14/07/2014

Libraries at the centre of the debate on copyright and text and data mining: the LIBER experience

Susan Reilly
LIBER, the Association of European Research Libraries
Netherlands

Abstract:

In May 2013, LIBER, the Association of European Research Libraries, coordinated the withdrawal of research and industry end-user representatives from the Licences for Europe stakeholder dialogue on text and data mining (TDM). The reason? Because a dialogue which focused on licencing as a solution for TDM was not in the interest of the end-user or reducing barriers to data driven innovation.

Since then LIBER has worked to bring a broader representation of stakeholders together to discuss barriers to the use of TDM. The European Commission also launched a consultation on the copyright framework, opening the door for libraries to effect real changes in the copyright system in order to accommodate, and foster, advances in research in the digital age.

At the same time publishers such as Elsevier and Nature have launched their own TDM services and policies. The launch of these services are a welcome development, but, if the terms of service are too restrictive, it may be that libraries are signing away the rights of end-users over the long term for access in the short term.

This paper will explore:
The barriers to data driven innovation as identified in our work with stakeholders
The LIBER position on the need for legislation change to accommodate text data mining
Why licencing is not a scalable and long term approach to facilitating text and data mining

Introduction

Data driven innovation is widely acknowledged as the way forward for the European economy. The ability to apply new and more powerful technologies to the huge mass of available digitally encoded information opens up whole new areas of possibility for both research and enterprise.
Tools for text and data mining are at the heart of data driven innovation and will, at the very least, be

essential to help researchers cope with the exponential growth in research output. In optimal conditions, text and data mining will allow us to harness the wave of data in order to solve society's grand challenges and realise the enormous economic potential of this rich resource.
Research libraries are in the midst of the data deluge, digitising collections, collection born digital objects, acting as repositories and publishing platforms for open access electronic articles and journals and, most recently, research data, as well as negotiating access to subscription content. Whether electronic journals, digitised material or raw research data, all of this content are data and it is up to libraries to ensure that this data is accessible and made available in a meaningful way.

It is in the interest of both libraries, in the realisation of their mission to support world class research and to ensure return on investment in digital collections and infrastructure, and researchers in pursuit of excellence, to see text and data mining realise its full potential.

There are, however, certain challenges to the development of text and data mining, including a copyright framework which is unfit for the digital age, a need for more integration of infrastructure and investment in the development of skills.

Background

"*Text and data mining (TDM) is the process of deriving information from machine-read material. It works by copying large quantities of material, extracting the data, and recombining it to identify patterns*."

TDM may be viewed as a secondary use of copyrighted works. There are four stages to the TDM process. First, potentially relevant documents are identified. These documents are then turned into a machine-readable format so that structured data
can be extracted. The useful information is extracted and then mined to discover new knowledge, test hypotheses, and identify new relationships.

Text and data mining more broadly has huge potential economic and social impact.

Looking at the field of medicine, McKinsey Global Institute reported in 2011[1] that effective use of 'big data' in the U.S. healthcare sector could be worth more than $300 billion U.S. a year, two-thirds of which would be in the form of a reduction in national health care expenditure of about 8%. In Europe, the same report estimated that government expenditure could be reduced by €100 billion a year. Text and data mining has already enabled new medical discoveries by linking existing drugs with new medical applications and by uncovering previously unsuspected linkages between proteins, genes, pathways and diseases.

Text and data mining of the web is being used by many technology companies to develop new products and services. They include IBM, SAS and many European-based Small and Medium Enterprises (SMEs). Many businesses currently use text and data mining to help the corporate sector profile current and future trends on the web, to provide customers (including governments) with text and data mining tools that save time managing their online assets, and to track potentially dangerous social upheaval.

Whilst in countries, such as the US, where fair use is in operation, TDM may be performed legally, the situation is not so clear in Europe. Although TDM is concerned with the extraction of facts and data, which do not fall under copyright, the process of deriving these facts and data may involve making copies of the original object (e.g. to convert to a machine readable format). What is more, as research today must be transparent and reproducible, it may be necessary to securely store the copies of these objects rather than simply making temporary copies for the purpose of an experiment.
A further mitigating factor in Europe is the implementation of a unique law relating to the protection of database rights. The European Database Directive and the sui genris database right renders the copying of significant portions of a database illegal. This means that even the copying of open access content in open access repositories may be illegal unless the database right has been explicitly waived.

Even if all subscription content were made available under a TDM friendly licence and all open access repositories waived the database right, a question mark still remains over the mining of the open Web. At the moment, it is easier for researchers in Europe to ask their US collaborators to mine openly available content (e.g. Twitter content) than risk the legal minefield presented by the European copyright framework.

Two recent studies have found that the European copyright framework may be having a chilling effect on TDM research in Europe. A recent report by the Lisbon Council pointed out that this could leave Europe at a serious disadvantage as it could create a situation where the majority of patents for text and data mining technology are held by third countries.

Barriers to text and data mining

Libraries representing their users have found themselves at the centre of discussions around the facilitation of text and data mining in Europe. In February 2013 the European Commission launched a stakeholder consultation exercise, called Licences for Europe, to discuss licencing as a solution for text and data mining. Over 60 influential organisations and individuals, representing researchers, science groups and industry, signed a letter to the European Commissionexpressing concerns about the scope of the Licences for Europe process. They did this because they believed that the right to read is the right to mine and that no additional licences for mining should be required once access to the content has been paid for. A licence solution requiring researchers to contact the owners of all of the content they wish to mine for every project they wish to use TDN in was viewed as unscalable. Ultimately the stakeholders representing end users withdrew from Licences for Europe because it was not in their interest to appear to support licencing as a solution for overcoming the barriers to text and data mining to the exclusion of any dialogue around the need for copyright reform.

What Licences for Europe highlighted is that there is a general lack of awareness amongst policy makers about what TDM is and why it is import for research and for innovation. To counter this LIBER[5], the Association of Europe Research Libraries organised a workshop in London to continue the conversation on the need for legislative and other changes related to text and data mining. European policy makers were invited to attend, so that they could explore and understand how text and data mining fits in to the bigger picture of data-driven innovation from both the research and industry perspective.

The workshop drew together practitioners of text and data mining from both research and industry, open access publishers, research infrastructures and legal experts to discuss their practices and the barriers they perceived to text and data mining.

Although the participants came from a diverse range of backgrounds there was a great deal of agreement over what the barriers to text and data mining are:

Copyright:

Logically TDM should not fall under the scope of copyright law as it is not concerned with the artistic expression of ideas which copyright regulates, but with the extraction and analysis of facts and data which are specifically excluded from regulation by IPRs in international laws. However, because TDM can involve the making of copies of content for the extracting of data it does fall under European Copyright law.

We cannot discriminate between humans and robots in the digital age. TDM is a form of reading, whether or not it is performed by a human or a machine. The right to read should be the right to mine.

"Should 'fair use' become a fundamental user right?" is a question that European policy makers should be asking. It is clear that a specific exception is needed for text and data mining, but the evolving nature of digital content and technologies will necessitate a more flexible IP framework in Europe to reduce barriers to innovation and ensure competitiveness. Is it practical to expect busy parliaments to legislate on each and every technological activity to create a new limitation and exception in the Copyright and Database Directive? A 'fair use' type exception, in addition to existing specific exceptions, would be a welcome and logical solution.

Copyright law in Europe must be harmonised. That the Information Society Directive is currently implemented differently in different European countries creates a sense of uncertainty for users. Any future exception for text and data mining should be mandatory and must be implemented in the same way across Europe and should not differentiate between commercial and non-commercial research. The definition of non-commercial is extremely problematic and, if interpreted narrowly, could be defined according to funding source or to the means of dissemination e.g. is a blog with advertising commercial?

Competitiveness:

Other regions in the world are ahead of Europe in the use of text and data mining because the have laws in place that allow it e.g. the US has fair use. Also, Europe has an additional legal barrier to text and data mining in the form of the Database Directive. Because it prevents extraction and/or re-utilisation of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of a database the sui generis Database right not only precludes the mining of database content, but, as Europe is the only region in the world to have such a right in place, reduces European competitiveness in data driven innovation. This means that, for instance, a researcher who mines a freely available, and possible publicly funded, database of content that is available under a creative commons licence may still be infringing IP law.

Licencing:

Given the heterogeneity of the content mined by both researchers and commercial companies, the variety and heterogeneity of needs, aims, and projects, and the fact that content is collected automatically from a multitude of resources, licensing can not scale and is an unfit solution for this purpose.

Awareness:

There is a lack of awareness generally about what TDM is. The current low demand to mine content from researchers is due to lack of awareness as well as the prohibitive amount of effort that is consumed by negotiating its legalities. It will become increasingly necessary for researchers to employ text and data mining to ensure the quality and accuracy of their research. Awareness raising amongst researchers about the potential and benefits of TDM as well as the related IP issues needs to occur. There is also some confusion in relation to the use of personal data. There needs to be guidance for researchers and industry on how to deploy mining ethically and responsibly.

Evidence:

Although there is a need to build up an evidence base or create 'facts on the ground' for the need to make legislative change to accommodate TDM, such an evidence base could contravene current copyright law and few are willing to take this risk. In order to show the power and societal usefulness of text and data mining, we need to do text and data mining at scale over content, but such activity largely contravenes current copyright law.

Solutions

In the discussions about the negative impact that the lack of legal clarity is having on the uptake of TDM in Europe several solutions to have been put forward by legal analysts, library and research organisations, and rights holders.

1. Licencing

One of the outcomes from Licences for Europe was a joint statement by STM publishers committing to providing licence solutions that would permit TDM. There are several drawbacks to this proposed solution. Even if the terms of these licences were to meet the needs of researchers wishing to perform TDM e.g. by permitting automated crawling of content and long-term deposit of copies in a secure repository, this would not solve the problems researchers in Europe face in relation to mining the open Web.

From the library perspective, the negotiation of licences seems a time comsuimg and unscalable solution as it would involve negotiating with thousands of publishers on an individual basis and assumes that all publishers will agree to the same terms.

Recent developments have shown that the licence terms that publishers are offering are not fair and provide an insufficient level of access for the purpose of text and data mining. In January, Elsevier launched its new policy for TDM. The policy states that researchers may only mine content to which they have legal access to via an API. This restricts the content which researchers can mine as only the text and not figures and images are available via the API. Within the institutional licence, automated crawling of the content is explicitly prohibited, as is deposit of copies in a secure repository, therefore experiments are not fully reproducible. Furthermore, conditions are imposed on how the researchers may make the results of their research available.

One of the most objectionable conditions of the policy is that it requires the researcher to register and to sign a click-though licence to access the API. The terms of the licence may change at any time and place an unfair burden of liability on the researcher, especially given that the institution has already signed an institutional licence for access to the content. Such is the concern about this policy that several library and

research organisation have signed an open letter to Elsevier asking them to withdraw the policy.

2. Changes to the copyright framework

An exception in copyright law for the performance of data analysis would create the legal clarity that is currently lacking within the European copyright framework. Such an exception for TDM has been adopted by the UK in June 2014. Perhaps because it has been developed under the constraints of the current European Information Society Directive, the UK exception only permits TDM for non-commercial purposes. Given that even within the context of research institutions, what can and cannot be defined as non-commercial is ambiguous (universities receive funding from several sources) the UK exception does not achieve the ideal level of legal certainty. The UK exception does have one very beneficial element in that it prevents override by contracts. This means that an institution or individual cannot sign away their right to mine under licence. Any new exception at European level must also have such an element, as the legal certainty which it will have been implemented to achieve should not be undermined by licence conditions. It also does not make sense to make such an exception non-commercial. Apart from the fact that funding from research comes from both commercial and non-commercial sources, a non-commercial clause would act as a disincentive to those who would seek to develop and subsequently patent TDM technologies.

Another proposed solution is that, rather than implement a specific exception for TDM, strategic reform of copyright law, which makes the distinction between the role of copyright in protecting the author of expressive works and where it prevents innovation, via an interpretive instrument. Such an instrument would be beneficial over the long term as it would negate the need for further reform to accommodate new technologies.

Looking forward

The current copyright regime in Europe is not fit for this digital age and is certainly precluding European research from progressing at the same pace as global counterparts e.g. digital humanities researchers in the US are already mining the millions of digitised books available in the Hathi Trust. This is made possible because the principle of fair use applies to the US. There is no equivalent framework in Europe.

Anecdotally it seems that researchers are mining content without being aware or considering if what they are doing is legal as there is an assumption that if the content is available to read then it is available to mine. The technical process involved in mining, however, means that European researchers may be inadvertently infringing copyright law by mining in copyright content. To address the need for increased awareness and an evidence base for policies to support text and data mining, researchers should actively promote their text and data mining projects.

Policies for TDM similar to that of Elsevier may become more common. Crossref has already developed a platform which will allow access to content for TDM via an API. It has been designed so that each publisher can require that the researcher sign a different click-though licence in order to mine their content. Whether the expectation that a researcher can agree to terms from a multitude of licences and still operate with legal certainty is fair remains to be seen.

Libraries have an active role to play in increasing and facilitating text and data mining. They can increase awareness amongst researchers about the benefits of TDM and of their rights. They can also encourage the use of open licences that allow TDM e.g. the Creative Commons licence 4.0 which include a waiver of the database right and provide tools and infrastructure for mining. They should also be mindful of researcher requirements for TDM when negotiating licences for access to content and, in jurisdictions where the right to mine content already exists, avoid clauses in licences which prevent the exercise of this

right or unfairly limit the mining activities of the researcher. Lastly, they can also help to facilitate the development of protocols and best practice for TDM, which would help to ensure that researchers deploy robots responsibly and without having a negative impact on the functionality of publisher infrastructure.

The European Commission is conducting a review of the copyright framework during 2014 with a view to potentially beginning a process of reform in 2015. This means that it is important that libraries and researchers voice their concerns about the barriers they face when accessing content for the purpose of mining. At the very least a specific exception allowing the copying of legally acquired content for the purpose of extracting facts and data, so long as the output cannot act as a substitute for the original content should satisfy both researchers who wish to deploy innovative research techniques and rights holders who wish to protect their assets.

HYPERLINK "http://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining" http://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining
Text and Data Mining: Its importance and the need for change in Europe, **HYPERLINK "http://www.libereurope.eu/news/tdm"** http://www.libereurope.eu/news/tdm

HYPERLINK "http://www.nature.com/news/tensions-grow-as-data-mining-discussions-fall-apart-1.13130" http://www.nature.com/news/tensions-grow-as-data-mining-discussions-fall-apart-1.13130

HYPERLINK "http://ec.europa.eu/internal_market/copyright/docs/licences-for-europe/131113_ten-pledges_en.pdf" http://ec.europa.eu/internal_market/copyright/docs/licences-for-europe/131113_ten-pledges_en.pdf

HYPERLINK "http://libereurope.eu/news/european-research-organisations-call-on-elsevier-to-withdraw-tdm-policy/" http://libereurope.eu/news/european-research-organisations-call-on-elsevier-to-withdraw-tdm-policy/

HYPERLINK "http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf" http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf