EDiMA positioning on text and data mining

Harnessing data driven innovation is key to the success, future growth and competiveness of European companies and markets. Discussions on "text and data-mining" and copyright take place in this context. In this paper, EDIMA highlights the impact of the issue far beyond the scientific publishing sector, which is often the focal point of discussions. We have highlighted some insights into the broader impact of the copyright review in a way that would elaborate the impact that current and future knowledge discovery tools would have on the European economy. EDIMA recommends that it is clearly established that text and data mining is not subject to copyright.

As highlighted in the European Commission Communication '*Towards a thriving data-driven economy*' (2014)[1] trends in big data technologies and services offer 'enormous potential in various fields, ranging from health, food security, climate and resource efficiency to energy, intelligent transport systems and smart cities, which Europe cannot afford to miss.' Yet, text and data mining (TDM) is one area within the global data-driven economy that Europe has yet to exploit to its full potential.

TDM is the process in which data analytics are used to discover knowledge, for example by establishing connections or patterns between texts and other types of content, using large amounts of text or data. TDM reduces exponentially the time it takes to source and correlate relevant data and has an enormous potential to foster innovation and bring about European economic and societal.

EDiMA would like to highlight the role that TDM will continue to play in the realisation of the Digital Single Market for Europe. Through maintaining that facts and ideas cannot be protected by copyright and the activity of TDM should not be able to be controlled by the copyright owner, whether for commercial or non-commercial purposes, will allow EU companies, researchers, institutions and public sectors to continue to drive efficiency and quality, while remaining competitive globally.

TDM is knowledge discovery, and knowledge cannot be subject to copyright

Depending on how TDM is defined or interpreted, it is likely not to require additional or specific copyright protection. No-one is challenging the notion that access to the material being mined (whether copyright subject matter or not) needs to be secured lawfully in the first place. More surprising is the suggestion that all TDM activities should require a specific licence 'on top'.[2]

Many consider that "the right to read is the right to mine": TDM is the functional equivalent of human researchers reading and analysing text and data - an activity that has never been an infringement of copyright – through computer analysis. The results of such TDM is new knowledge, not a reproduction of the work. In this sense, it is a non-expressive use and presents no legally cognisable conflict with the statutory rights or interests of the copyright holders.

---

[1] COM(2014)442 final,

[2] Lawful access may not require a licence, for instance for public domain material, non-copyright protected subject matter, material outside of the scope of copyright (e.g. court judgements in certain countries), or subject to an exception.

Looking at it another way - TDM is a process that extracts information or knowledge from vast amounts of data, as opposed to extracting the data itself.[3] It is better understood as "knowledge discovery", extracting new knowledge for data sets. This means it is a non-consumptive use: it does not undermine the market for the original work or reduce incentives for its creation in the first place. It also means copyright protection is not justified: knowledge and discoveries are not subject to copyright. Copyright protection only covers 'expression' and not ideas, procedures, methods of operation or mathematical concepts.[4] A monopoly on information or ideas would be extremely burdensome on society as a whole, as well as being contrary to international law.

There is nevertheless some debate as to how the ubiquitous reproduction right might also inadvertently put TDM activities under strain and lead to over-reliance on the exception for transient reproductions. Such a narrow interpretation of the exception, however, would go against the spirit of copyright law, as there is no new communication to the public of the work in the context of TDM.[5] It would have consequences – now well documented – for research, but also for growth, innovation and competitiveness, which have so far not been given sufficient attention.

The current state of play, and the need to move beyond

Current discussions are naturally linked to the initiative of some Member States to introduce an exception for TDM.[6] Such national initiatives need to be taken into consideration. In the UK, legislation has been enacted, with a focus on the relationship between publishers (so-called "STM" publishers) and research institutions. There preparatory work was careful to indicate that there were doubts as to whether TDM required a specific licence. Further, the provisions enacted also focus on preventing contractual clauses which would overrule the benefits of the limitation.

In relation to research activities, evidence clearly suggests that access to TDM increases the productivity and efficiency of research and, critically, unlocks 'hidden' information and leads to an improved research and evidence base.[7] There is also clear evidence that the use of TDM by researchers in Europe is currently lower than in the US and Asia, with a knock-on negative effect on research outputs. This phenomenon impacts research in a broad range of fields – more so in computer science but also in social sciences such as economics, management, international business, innovation studies, etc.[8] The report from the European Commission's Expert Group on TDM found this to be a serious handicap to Europe's competitiveness and lead to the possible loss of talent and investment to more favourable research locations.[9]

It is also worth noting that the benefits of TDM also greatly contribute to the operational efficiency of European governments, for example, saving government administrations over €100 billion in

[3] Jiawei Han, Micheline Kamber, 'Data mining: concepts and techniques' (3rd ed, 2011).
[4] See e.g. TRIPs Article (2), WIPO Guide to the Berne Convention: Article 2(1)). In addition, under EU law, facts are incapable of being 'original' in the EU sense (as described e.g. in C-5/08 *Infopaq*; In the UK, "Copyright is not intended to prevent use of facts for research, and this exception is intended to remove the block on reuse of materials for research using these tools", "Ian Hargreaves, Digital Opportunity: A review of Intellectual Property and Growth' (2011)
[5] See e.g. Article L. 122-3 French Intellectual Property Code, which provides that a reproduction is the material fixation of a work by means which allow, indirectly, its communication to the public.
[6] https://www.gov.uk/exceptions-to-copyright. The Irish government is also considering a copyright review which would cover TDM.
[7] The Value and Benefits of Text Mining. JISC, 2012. http://www.jisc.ac.uk/sites/default/files/value-text-mining.pdf
[8] S. Fillipov, Mapping Text and Data Mining in academic and research communities in Europe (Lisbon Council, 2014), available at http://www.lisboncouncil.net//index.php?option=com_downloads&id=1034
[9] Standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining. DG Research and Innovation, 2014. http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf

operational efficiency improvements alone by using data more effectively, and can lead to €195 billion of potential annual value to Europe's public sector.[10]

However, we seek to highlight that discussions at the European level have moved beyond this initial focus. Accordingly, it is critical to also take into account the broader impact that reviewing TDM activities in a copyright context may have, beyond the specifics of STM publishing. In particular, the impact on innovation and the Internet should be taken into account.

Text and data mining: an innovation, growth and competitiveness boon

"Data mining" or knowledge discovery – the combination of large amounts of information with improved computer engineering technologies – provides a raft of new opportunities for innovation, growth and competitiveness in Europe.[11] It directly impacts economic operators whose primary activity focuses on computer and software engineering applied to data: from start ups such as Sweden's Recorded Future to more established products such as those developed by Autonomy in the UK or SAP in Germany. This sector is at the forefront of highly competitive, cutting edge developments in technology and engineering, from machine learning to advanced data analytics.

The economic impact of TDM extends beyond those economic operators and underpins the growth and competitiveness of vast sways of the European economy that are harnessing digital technologies, in a manner that cuts across their entire activity and across all sectors, whether services or industry. As the examples below suggest, the impact of TDM is cross-cutting, from gathering intelligence to develop new products, to improving the production and marketing of existing products.

For instance, new technologies for speech recognition, subtitling, translating (including the European Commission's "Machine Translation Service"),[12] software analytics (as developed e.g. by German-based SAP or UK based Autonomy), etc., rely on large amounts of data as input, including but not limited to materials found on the Internet. These technologies underpin the development of applications in life sciences (i.e. bioinformatics, genome research),[13] humanities (social sciences),[14] health care (e.g. from new applications for existing drugs to pharmacovigilance),[15] journalism (data-journalism and visualisations) and many other markets and applications.[16]

These technologies and products would all be jeopardised should TDM be subject to further copyright restrictions. London-based Shazam, which reached global scale by developing fingerprinting technology to allow its users to identify songs (and subsequently purchase them on third party stores), is one example. It is worth noting that Shazam has extended its technology to identify television shows and ads, Swedish start up Recorded Future which analyses data to develop predictive analysis tools, is another example. Varying interpretations of what distinguishes commercial versus non-commercial use of TDM will continue to create barriers for EU innovation. Although some EU member states have

---

[10] See e.g. McKinsey Global Institute's (MGI) report on data driven innovation, The next frontier for innovation, competition, and productivity (2011),

[11] See e.g. McKinsey Global Institute's (MGI) report on data driven innovation, The next frontier for innovation, competition, and productivity (2011),

[12] Based on based on statistical machine translation, see http://ec.europa.eu/information_society/newsroom/cf/dae/itemdetail.cfm?item_id=14322

[13] E.g. the EU's 10-year €1.2bn Human Brain Project.

[14] Brief of Digital Humanities and Law Scholars as Amici Curiae in Authors Guild v. Hathitrust (2013). http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2274832%20; see also the mining the closed-captioning of broadcasts for the purposes of research and journalism, http://www.knightfoundation.org/blogs/knightblog/2014/1/7/internet-archives-virtual-reading-room-empowers-data-mining-societal-scale/

[15] See e.g. Weeber et al. 'Generating Hypotheses by Discovering Implicit Associations in the Literature', J Am Med Inform Assoc. 10 (2003); Sophia Ananiadou et al., 'Text Mining and its Potential Applications in Systems Biology', 24 Trends in Biotechnology 571 (2006)

[16] From start-ups such as German based Linguee (which won an award from the German Federal Ministry of Economics and Technology) or Swedish company Urkund to IBM's Watson.

already enacted legislation[17] their caution is entirely based on their interpretation of the current exception for research purposes. The vast majority of EU law exceptions and limitations do not distinguish between commercial and non-commercial activities. When they do, it is in the context of provisions that target a particular type of use (e.g. private use) or user (i.e. museums, hospital). This is for good reason, as an outright distinction between commercial and non-commercial would be unworkable (as well as unjustified). Consider products developed commercially to serve non-profit activities (use of "TurnItIn" to detect plagiarism in the education sector), the creation of infographics from different sources in a classroom or in a business presentation, data-journalism for a commercial news organisation, drugs being developed by a non-profit, a university that seeks to exploit its work commercially or a pharmaceutical company.

Recommendations

The problem lies in the assertion that TDM might infringe copyright. It is important to establish, clearly and unambiguously, that TDM is not currently subject to copyright protection, and should not be. In particular, facts and ideas cannot be protected by copyright.

The distinction between commercial and non-commercial uses is a falsehood. It is in the public interest that facts and ideas remain outside the scope of copyright, whether for commercial purposes or not, just as it follows that it is in the public interest for TDM to remain outside the scope of copyright.

Any review of copyright which touches upon TDM should exercise great care not to threaten the innovation it is driving. Inaccurate scoping of legislation can trigger inadvertent effects that would significantly jeopardise innovation and growth in Europe.

[17] https://www.gov.uk/exceptions-to-copyright