An analytical review of text and data mining practices and approaches in Europe

Policy recommendations in view of the upcoming copyright legislative proposal

Authors – Diana COCORU & Mirko BOEHM

May 1, 2016

Openforum europe
open, competitive choice for IT users

BACKGROUND

In 2012, the European Commission embarked upon a revision of the 2001 Copyright Directive. Since then, several public consultations have been conducted, structured stakeholder dialogues have taken place, the Commission has published its Digital Single Market Strategy, and the European Parliament has adopted its 'own initiative' report on copyright reform as well as having put a Working Group in place specifically to deal with copyright. Now, the Commission is planning to publish its legislative copyright proposals in the autumn of 2016. Throughout the initiatives enumerated above, there has been (and still is) much controversy around TDM (text and data mining) – i.e., what exceptions and options should be allowed?; under which conditions?; etc.

METHODOLOGY

This paper is based on extensive desk research, including most of the benchmark reports, such as the European Commission funded Expert Group Report (2014), the study by De Wolf and Partners (2014), the UK IPO's 'Exceptions to Copyright' brief (October 2014), as well as numerous other reports, position papers, articles and blog posts[1]. The initial findings have been discussed at the Round Table that OFE organised in October 2015, the conclusions of which are available in the follow-up White Paper. The desk research and Round Table discussion have been complemented by a series of interviews with academics, researchers, start-ups, and more established companies (including publishers and infrastructure providers)[2].

[1] A comprehensive list can be provided upon request.

[2] The interviews were conducted between September 2015 and February 2016, with the following experts (in alphabetical order): Geoffrey Bilder (CrossRef), Vivian Chan (Sparrho), Elizabeth Crossick (RELX), Lucie Guibault (IViR), Prof. Ian Hargreaves, Rachael Lammey (CrossRef), Thomas Margoni (Openminted), Peter Murray-Rust (Content Mine), Cameron Neylon (Public Library of Science), Julia Reda (MEP), Tim Stok (RELX), Kalliopi Spyridaki (SAS).

SUMMARY

TDM can be approached using different regulatory options, such as an explicit right in the national or European law, an optional exception, a mandatory exception or fair use of the copyrighted material (which is the case in the US). The current status in the EU is that TDM is one of the exceptions provided for in the InfoSoc Directive. This led to Member States implementing it in a very fragmented way, or in some cases choosing not to implement it at all, creating significant legal uncertainty for researchers and other stakeholders.

In this paper, OFE presents the vast majority of the various arguments and approaches relating to TDM in Europe, as well as providing recommendations addressing various identified challenges.

For each of the following points, this paper first presents the context, followed by one or more direct recommendations:

> the negative impact on innovation of strong copyright protection
> fragmentation in the application of the exception at national level
> use of terminology
> the basis for allowing any exception
> why is a licence needed?
> the legal regime applicable to the immediate output of TDM
> non-financial costs of allowing non-commercial TDM through imposed licensing
> who should benefit from the added value of TDM?
> whether commercial TDM requires a different regulatory framework, and what other alternatives might apply?

- striking the right balance
- what - and whom - to take into account in terms of transaction costs?
- why distinguish between commercial and non-commercial TDM?
- the argument of fair compensation

> how the respective fears of miners and of publishers could influence the technical arrangements to enable TDM, as well as the impact of choice of data handover point.

Finally, following analysis of the listed context items and the presentation of our associated recommendations for each, we present our overall conclusions.

THE NEGATIVE IMPACT ON INNOVATION OF STRONG COPYRIGHT PROTECTION

A June 2015 research paper[3] entitled *'Is Europe Falling Behind in Data Mining? Copyright's Impact on Data Mining in Academic Research'* discusses how copyright affects data mining by academic researchers. Stronger rule-of-law is associated with less data mining research. While the Hargreaves report published in 2014 contains indications of such an effect (whilst noting that more empirical evidence was needed), this 2015 report is the first time that an empirical study has borne out a significant negative association between copyright protection and innovation. There is a strong argument (advanced by publishers) that creating statutory TDM exceptions to copyright protection would undermine the investment incentives which ensure that high-quality content is available. In response to those arguments, libraries point to the fact that associated copyright licence terms tend to be highly complex, and can take months or even years to complete. Licences are often subject to the laws of other jurisdictions, and - in most European countries - they can override the flexibility that exceptions are intended to provide. Libraries argue that an exception for TDM can act as an investment incentive. By implementing the exception for TDM proposed by the Hargreaves review of the UK's copyright frameworks, the UK government has made a clear statement that legal clarity around activities such as TDM are expected to spur innovation and growth. Following the statutory implementation in 2014 of this UK exception, tools to support TDM and improve the quality of content have already begun to emerge., and researchers in the UK have developed their own openly available tools for converting text files into structured standardised formats. Making licensing a condition precedent to TDM undermines investment in TDM by diverting capital and other resources away from important TDM research and development activity into licence negotiation, compliance and monitoring activities.

Associated copyright licence terms tend to be highly complex, and can take months or even years to complete.

According to the June 2015 paper cited above, researchers conclude that the number of published research articles is a reasonable indicator of the level of innovation by academic researchers; and (as as far as data mining research is concerned) copyright seems to have a negative net effect on the level of innovation[4]. Results suggest that in the case of academic research and data mining, the adverse consequences of copyright protection on the creation of new information goods outweigh the benefits. One key consideration is that data mining research often draws on a large number of input works to which others hold the copyright. Moreover, academic researchers are put at a disadvantage, as typically they collaborate across many organisations, whereas large companies represent a single entity or enterprise with access to bulk data.

[3] Handke, Christian and Guibault, Lucie and Vallbé, Joan-Josep, Is Europe Falling Behind in Data Mining? Copyright's Impact on Data Mining in Academic Research (June 7, 2015). Available at SSRN: http://ssrn.com/abstract=2608513 or http://dx.doi.org/10.2139/ssrn.2608513

[4] Ibid., p. 21

The main identified results are that: researchers in EU Member States with strong copyright protection publish significantly fewer articles on data mining; and the specific copyright system within the EU appears to inhibit data mining by academic researchers. In particular, researchers in major US and Asian economies engage in data mining much more than researchers in large European economies such as France or Germany.

Furthermore, copyright law should have stronger effects where it is associated with effective enforcement. Among EU Member States with strong copyright law, reports find that countries with a weaker rule of law (e.g., Portugal, Greece and Spain) publish significantly more data mining articles as a proportion of their total research output than countries with a stronger rule of law (e.g., Germany, the Netherlands and Sweden).

Recommendation

Europe should aim to enable analytical mining through openness and accessibility of content through supportive frameworks.

FRAGMENTATION IN THE APPLICATION OF THE EXCEPTION AT NATIONAL LEVEL

Allowing for an optional exception, as provided in the InfoSoc Directive (which currently has been approached in a variety of different ways at national level[5]), has led to legal uncertainty and to opportunities being missed.

Recommendation

The Commission needs to provide coherence and harmonisation for TDM across Europe, through a regulatory intervention proportional to the benefits of TDM and the costs of non-intervention.

USE OF TERMINOLOGY

Some use the term "text and data mining", others prefer the term "content mining". The reason for preferring the latter is because although typically some text will be included in any scientific paper, most of the rest of such a paper will tend to consist of diagrams, mathematical equations,

[5] The optional exception for research has not been (fully) implemented in all 28 Member States. It was not at all implemented in Greece, Finland, Netherlands, Sweden. It was partially implemented in Austria, Bulgaria, Czech Republic, Denmark, Spain, Ireland and Slovakia. It was implemented at least in Belgium, Cyprus, Germany, Estonia, France, Croatia, Hungary, Italia, Lithuania, Latvia, Luxembourg, Malta, Poland, Romania and the UK (source: DG RTD, http://fr.slideshare.net/OpenAccessEC/a-researchfriendly-copyright-environment-in-the-digital-age-a-european-perspective-52165041, slide 18)

facts and tables. That is often the most valuable part of the paper for a researcher. and using "text and data" instead of "content" to describe what is to be mined could limit the input to be mined, for no justifiable reason.

Recommendation

Any legislative proposal should take into account the full spectrum of mining input: text, diagrams, mathematical equations, tables, graphic work, images and other facts. Using the term "content mining" rather than "text and data mining" seems to be more accurate, both for academic researchers as well as for commercial actors. Legal provisions should be drafted in a way that matches this evolution of mining activities, covering much more than mere text.

THE BASIS FOR ALLOWING ANY EXCEPTION

While some use art. 5(1) (temporary acts of reproduction), in support of their arguments to support the basis of a TDM exception, others use art. 5(3) (scientific research for non-commercial purposes) – e.g., the recently enacted UK exception. These two different approaches in effect divide the structure of the supporting argumentation in two directions, as follows:

Any legislative proposal should take into account the full spectrum of mining input: text, diagrams, mathematical equations, tables, graphic work, images and other facts. Using the term "content mining" seems to be more accurate.

A. Those who consider mining as a temporary act of reproduction take the view that no copyright infringement occurs when mining; not only do they view mining as a technical act, but they also consider that any such mining is taking place in order to extract facts (which cannot be protected by copyright); and, applying the same logic, the extraction of facts from works of art should also not be considered a copyright infringement. When viewed from this perspective, the appeal of the popular argument "the right to read is the right to mine" is very clear and understandable.

Stakeholders sharing this point of view consider TDM as being a tool which can be used for any purpose, and which should have nothing to do with copyright. They consider it to be an accident that digital copies, which are essential to the functioning of any digital technology (including TDM), are treated in the same way as physical copies under copyright law. Historically, making physical copying subject to copyright made sense; e.g., if somebody had made 20,000 unauthorised copies of a book and was storing them in their basement, a right-holder would need to wait until the infringer started to distribute any of those copies to others before being able to stop their activities. By contrast, digital technology involves a series of copying activities which are not in any way related to making a copyright-protected work available to others. In addition, if a copyri-

ght-protected work is made accessible on the Internet for anyone to see, an act of digital copying will take place at exactly the same moment as the work is made available to the individual reader. Since a licence is required before the work is published on the Internet, there would appear to be no justification for requiring an additional licence to permit digital copying by the individual who accesses the work over the Internet. Those sharing this point of view underline the importance of understanding that TDM is an essential part of our lives and that requiring permission from a copyright holder in order to do it is "absurd". Moreover, operators of search engines would appear to be in the same position as researchers, in respect to their requirement and ability to mine copyright-protected works. The fact that a large number of right-holders tolerate or explicitly license search engines to harvest and mine their sites' content reflects their interest in promoting traffic to their websites and the pivotal role that search engines can play in this regard. Yet the underlying activity is the same, and the making of digital copies for the purpose of large scale analytics should not result in a different approach being applied.

Recommendation

To address this accident, the mere making of digital copies should be taken out of the scope of copyright protection – instead, the focus should apply on exploitation of protected work in digital form through (e.g.) redistribution of unlicensed copies, unauthorised performances, etc.. Such a revised approach would mean that no licence would be necessary to cover the basic digital copying, since the mining activity would not constitute an infringement of copyright that would in turn require any specific exception.

The mere making of digital copies should be taken out of the scope of copyright protection.

B. Turning to art. 5(3) of the InfoSoc Directive, this scientific research exception is limited to cases which satisfy the requirement for a "non-commercial" purpose; and so the debate has opened another Pandora's Box, which has rightly underlined that this commercial / non-commercial distinction is an artificial one, which should be avoided, for a number of reasons:

> more and more partnerships are developed between research institutions and companies;
> there is no consistency in the definition of commercial versus non-commercial purposes. Some define it according to the entity undertaking the activity, while other look at the activity itself;
> the same distinction was also used in the Creative Commons and the Freedom of Panorama debates, and has proved inefficient.

The Commission should aim to achieve coherence in the legal provisions which it seeks to apply to TDM, with no consideration of 'commercial' versus 'non-commercial' purposes.

Recommendation

The Commission should aim to achieve coherence in the legal provisions which it seeks to apply to TDM, with no consideration of 'commercial' versus 'non-commercial' purposes. Allowing TDM for non-commercial purposes while forbidding it for commercial purposes would create legal uncertainty, from which all affected parties would lose, except for those who can afford to capitalise on the legal uncertainty. Limiting permitted mining activities to 'public interest research organisations' (PIRO) and to 'scientific purposes', as proposed in the Commission's December 2015 Copyright Communication, risks being seen as just another way of reflecting the commercial vs. non-commercial debate, which would leave citizens and scientists who are not PIRO-affiliated in limbo, limiting commercial TDM without any justifiable reason, and not resolving the underlying issue.

WHY IS NEED A LICENCE NEEDED?

Copyright-protected material may be the subject of TDM, but the purpose of engaging in TDM is not to perform or enjoy a work of authorship or art or to make it available to others, but rather to extract facts from it.

While specific facts and data elements are not protected by intellectual property laws[6], the text, documents or databases that are to be mined may be subject to copyright, related rights and/or database rights. The extraction and copying of content to which one already has legal access, and its transformation into a machine-readable format, is claimed to touch on the rights-holder's exclusive reproduction right. Some publishers have relied on their exercise of this exclusive right to justify their requiring a contractual licence to be accepted before granting permission to mine (on top of a licence for access to the underlying minable content). However, when embarking on such analysis (in view of the specific exclusive right concerned here), we should go back to the definition of copyright and the reasons which justify its existence. Intellectual property rights were created to remunerate the author for his/her creative effort and thus give incentives to produce/create more. However, if one creates an intellectual property right to extract information from already existing source input, the question remains what is the original criterion, or the added value from the author? Most argue that TDM is closer to reading than redistributing content (as in communicating the source content in a modified form). It follows that TDM is completely different from the types of activity which are typically restricted by copyright but which may be permitted by licence, such as selling a copy of a work or performing a work in front of an audience. Copyright-protected material may be the subject of TDM, but the purpose of engaging in TDM is not to perform or enjoy a work of authorship or art or to make it available to others, but rather to extract facts from it. However, bare facts are not protected by copyright and by the same logic, the extraction of facts from a work of authorship or art should not constitute a copyright infringement.

6 For these purposes, we do not consider "intellectual property" laws to include trade secrecy and confidentiality laws.

In this context it is also worth mentioning that the District Court of Amsterdam recently ruled[7] that TDM is already allowed, because the fundamental freedom of scientific research would be unreasonably restricted if copyright could be used to prevent TDM.

It lies in the nature of TDM to combine information from a large number of different sources that may have different right-holders, or no right-holders at all. Every licensing solution suffers from the fundamental flaw that it is necessary to arrive at some sort of licensing agreement with all of these different right-holders. Even finding out who these right-holders are can be very time-consuming and create transaction costs that render TDM prohibitively expensive. According to some of our interviewees, only large academic publishers with a substantial share of the academic publishing market are advocating for a licensing model.

A recent survey conducted by the Publishers Licensing Society (PLS) found that "the overall number of publishers receiving requests was small, with only 15% of respondents receiving requests in 2014. In addition, the number of requests received was small, with a total of 91 requests to text and data mine content from all respondents in 2014". Using the results of this survey, it has been argued that imposing a licensing requirement is the solution, rather than an exception. However this argument is flawed, because it does not take into account the discouraging effects that miners currently face: transaction costs (e.g., reading & understanding licence terms and conditions so as to know what is permissible, hiring a lawyer to negotiate terms, and fear of infringing in case of legal uncertainty). Many miners choose either not to engage in, or else just to drop, their projects; it cannot safely be said that a low rate of requests to publishers proves there is a low level of interest in mining; our conclusion is rather that the current legal framework is not conducive to satisfying the research ambitions of European society.

The only reasonable and workable solution is a mandatory, fully harmonised exception at EU level which covers all TDM activities for any purpose and which cannot be overwritten by contract.

Recommendation

We conclude that, short of legalising all digital copying activities and limiting the scope of copyright so that only those activities that make digital content available to new audiences constitute infringements, the only reasonable and workable solution is a mandatory, fully harmonised exception at EU level which covers all TDM activities for any purpose and which cannot be overwritten by contract. A generalised exception for TDM would represent a liberalised approach, allowing everyone to decide what to do with their content. Moreover, the sheer volume and diversity

[7] http://www.communia-association.org/2016/01/08/what-the-diary-of-anne-frank-can-tell-us-about-text-and-data-mining/

The table resulting from mining activities brings the EU database law regime into the debate.

of information that can be utilised for TDM, which extends far beyond already licensed research databases, and which are not viewed in silos, makes a licence-driven solution close to impossible. Many voices echo the fact that licences are not an alternative to a mandatory exception. By giving in to those who demand licensing as the best solution, in the academic publishing sector at least we would risk contributing to further market concentration and causing collateral damage to any TDM activity that does not use academic articles as its source material. In addition, even if the relevant TDM licensing conditions were more permissive, licensing would still not be a universal solution, for the simple reason that it is erroneous to assume that all right-holders are big companies, and that each operates (or will operate) its own appropriate licensing solution. The reality is that anybody can perform TDM. Therefore the rules for TDM need to work for normal citizens and establishments, and these rules must not impose unreasonable burdens, either on users, or on right-holders. Only an exception can in our view achieve that goal.

THE LEGAL REGIME APPLICABLE TO THE IMMEDIATE OUTPUT OF TDM

Typically, the immediate output of TDM activity is a table which summarises an extraction of facts. Bare facts are not capable of being owned, and are not protectable by copyright. Only the presentation or interpretation of facts can reach such a level of sophistication that its expression becomes a work capable of being protected by copyright. In such a case, the copyright to the higher-level (sophisticated) work would lie with the person who has produced the work[8], rather than with the person controlling and providing the data on which the TDM was carried out. By definition, the output of a TDM process represents a completely different information set than that provided by the holders of the rights in the original - and probably diversely-owned - datasets, as argued in the April 2014 report[9] of the Commission's Expert Group, led by Prof. Ian Hargreaves.

That being said, it is relevant to underline that the table resulting from mining activities brings the EU database law regime into the debate. Even if the output of mining is a database, the conclusion of the applicable regime depends on the degree of extraction and re-utilisation. As the decision of the ECJ in response for a preliminary ruling in *Innoweb BV v Wegener ICT Media BV and Wegener Mediaventions BV*[10] points out, the protection offered by the sui generis right under Directive 96/9 is intended to ensure that the person who has taken the initiative and assumed the risk of making a substantial investment in terms of human, technical and/or financial resources in the setting up and operation of a database receives a return on his/her investment by protecting him/her against the unau-

[8] Or his/her employer or assignee

[9] Entitled "Standardisation in the area of innovation and technological development, notably in the field of Text & Data Mining"

[10] Case C-202/12 – see: http://tinyurl.com/zoyrbko

thorised appropriation of the results of that investment. If the degree of extraction and re-utilisation is massive enough, it could constitute infringement. In its preliminary ruling, the Court concluded that the search engine did indeed infringe under the provisions of the local transposition of the Database Directive.

When it comes to TDM, the database resulting from the automated search conducted by the software could not fall under the concept of database enshrined in the Directive 96/9. However, these aspects should also be addressed when regulating what is allowed for TDM purposes. This is in order to avoid any legal uncertainty which could be created when analysing the investment in the software which enables mining, the APIs, and so on. According to the Database Directive, if the maker of the database has qualitatively and/or quantitatively made a substantial investment in either the obtaining, the verification or the presentation of the contents, he/she has the *sui generis* right (enshrined in art. 7) to prevent extraction and/or re-utilisation for a limited time of 15 years. If the TDM activity were to fall under this exclusive right, then all such mining activities would be subject to major obstacles ahead, not for any justifiable reason, but simply because the relevant existing laws have yet to be harmonised so as to take into account new technological uses in the realm of copyright/database law.

Recommendation

As the technological aspects of content mining advance faster than the pace of legal review and reform, we often can observe that 20 year old legislation is no longer adapted to present practices. Any legislative proposal covering TDM activities should at least clarify whether or not the Database Directive laws apply to any output of TDM which is presented in the form of a table, and how the applicable mining software and APIs used are defined in the context of the Database Directive ("qualitative or quantitative substantial investment"). However, it should be not be forgotten that to enforce protective mechanisms which bite on the output of software used to perform TDM would be an ideal way to ensure that Europe continues to deny itself the benefit of this new opportunity to innovate and to develop and market successful new products.

NON-FINANCIAL COSTS OF ALLOWING NON-COMMERCIAL MINING THROUGH IMPOSED LICENSING

When looking at the distinction between commercial and non-commercial purposes, it is also necessary to address the argument used by a number of publishers in the current legal framework, where some specific licence agreements can be seen to have been deployed specifically to address and permit TDM, without associated royalty payment obligations, but nevertheless subject to certain important limitations. The publishers concerned make the point that these licences deal with the legal uncertainty issues. In such cases however, even where no financial costs or payment obligations are imposed by the publisher, a number of hurdles remain in place. For example, some academic publishers may state in their TDM licensing conditions that the output resulting from use of their APIs can only be used for non-commercial purposes under a CC BY-NC license. In other words, a researcher conducting TDM under that licence would not be able to publish his/her results in a commercial journal, or would only be able to do so subject to extensive limitations[11]. Another example is that certain subscription contracts contain express provisions limiting the amount of content that can be downloaded per online session. In addition, certain academic publishers adopt a licensing approach which obliges the licensee to perform the mining activity 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 is highly likely to expose his/her interests and algorithms to the publisher. Moreover, such restrictions preclude the potential for gains in innovativeness displayed by decentralised, self-selecting groups, such as open source communities. Further, some publishers impose constraints which regulate how the researcher can share and publish the results of the mining. From a purely scientific point of view, such a licensing approach is unacceptable, because reproducibility is a major issue for researchers. Reproducibility can be impacted by at least these two factors: 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.

Recommendation

Specific emphasis should be directed on achieving greater awareness of the importance for all researchers of reproducibility, and any updated legal framework needs to take this factor into account. Researchers must be able to share the results of TDM activities, as long as these results are not substitutable for the original copyright work - irrespective of copyright law, database law, or any contractual terms to the contrary. Without this

[11] E.g., whilst any commercialisation of the research findings, inventions or ideas resulting from the TDM effort are owned by the researcher/company, some publishers place a restriction on how the underlying subscribed content (i.e. the source article) is re-used. They underline that in their terms and conditions, there are no restrictions on where and how researchers publish their results, but if they are using the underlying article(s), then the researchers are free to use snippets of up to 200 characters surrounding and excluding the text entity matched or to include bibliographic metadata.

Publishers do not always push for a licensing approach purely out of a desire to maximise the royalties that they receive.

right, legal uncertainty may prevent important research and data-driven innovation, and so put researchers, institutions and innovators at risk.

The Commission should be aware that publishers do not always push for a licensing approach purely out of a desire to maximise the royalties that they receive. It seems that fears of overloading (or even blocking) the publisher's website or other servers as a consequence of having allowed too much traffic, or fears of piracy (stealing content by downloading under a false pretext of doing so for TDM purposes) may motivate certain publishers to impose the use of their own APIs. Although this might be the case in some situations, a balance should be found between the measures imposed by publishers to avoid website/server overload or piracy and the actual needs of the miners (whether they be academic researchers or not). Using technical means to avoid piracy imposes a disproportionate burden for researchers, and it also creates a serious obstacle for research, whereas arguments based on the need of publishers to avoid the risk that their websites or other servers might be overwhelmed is yet to be better evidenced and requires proper substantiation.

WHO SHOULD BENEFIT FROM THE ADDED VALUE OF TDM?

It remains debatable whether the original creators of content on which TDM activity is performed should be rewarded, for a number of reasons:

> Creating value by mining the source corpus is based on the efforts of the miner who translated the immediate mining data into something which makes sense, and adds to the previous general knowledge. The mining activities are valuable only and insofar as someone sees connections or links, and decides to translate data into something which makes sense.

> Profit sharing would not be economically efficient, because it is artificial and it would be very difficult on a case by case basis to assess what constitutes a correct (or even just a fair) percentage of the value generated by the miner through his/her creative work.

> There are practical limitations to sharing revenues with initial authors of the corpus (or their employers or assignees). In TDM, one might be using fragments from millions of sources. The mere fact of identifying and tracking them would present considerate and problematic challenges, and this is hardly a precedent which we should wish to set in the domain of scholarship and academic research.

Further, certain publishers consider that they should be remunerated for allowing mining on the content subscribed to. Although they allow mining for non-commercial purposes without additional *financial* costs, in the case of mining for commercial purposes, these publishers consider that they are entitled to reserve the right to "add more value" through targeted products and services, as an expression of their commercial freedom.

Recommendation

The Commission should not introduce a legal system which directly imposes such an artificial limitation between commercial and non-commercial purposes.

We conclude that whoever is able to turn the mining results into something that is economically useful or monetisable should be free to be rewarded for that activity, and moreover without being required to share their reward with the authors of the initial corpus (i.e., the source used for the mining activity).

As for the argument that publishers should be rewarded, an agreement for providing specific products and services could be concluded, and the owner can put in place specific mechanisms (e.g., a pay-wall) or require specific conditions for access that limit the use depending on the purpose. However, this should remain an opt-in choice and the Commission should not introduce a legal system which directly imposes such an artificial limitation between commercial and non-commercial purposes.

Mining should be allowed without further limitations once access to content has been granted. The undertaken research infers that this can only be achieved by passing an EU-wide mandatory exception for allowing TDM no matter the purpose (be it commercial or non-commercial).

One of the basic aims of the EU is to establish a competitive common market - introducing such an exception would create competition, thus encouraging a more dynamic and thriving market.

One could perhaps envisage commercial agreements being set up, under which one would pay not for the act of mining, but for the performance of the mining system (i.e., the speed or throughput of the machine or system used). However, to the extent that there is no distinction between the charges that apply whether one reads for research or commercial purposes, it appears to be entirely inappropriate to allow such a charging methodology to be applied to data where it does not apply to differentiated usage of the underlying copyright articles.

WHETHER COMMERCIAL TDM REQUIRES A DIFFERENT REGULATORY FRAMEWORK, AND WHAT OTHER ALTERNATIVES MIGHT APPLY?

Striking the right balance

In thinking about copyright, economic policy-makers aim for a welfare-maximising balance between benefits for users and incentives for right-holders. The aim is to strike the right balance between incentivising the production of 'works', whilst avoiding 'deadweight' welfare losses[12],

[12] Copyright confers an exclusive right on the copyright holder to control the copying, distribution and performance (etc.) of independently creative works of authorship and other protected artistic works. Demand is lower when price is higher. The "deadweight welfare loss" is the difference between the price that consumers were willing to pay in the absence of copyright protection and the price that they in fact pay, a price which is being fixed above marginal production costs.

between copyright's long-standing and legitimate role in protecting the rights of authors of 'expressive' works, and copyright's more questionable role in the digital age of presenting a potential barrier to modern research techniques and so to the pursuit of new knowledge.

If copyright owners did not enjoy the exclusive rights granted to them by law, over time one would expect the price of copies of works of authorship (etc.) to approximate slightly more than the marginal cost of making the work available, which in the case of digital information goods may be close to zero. The financial incentives for originators of the material to invest in innovation would then be diminished, and the supply of innovative output would most likely decrease, which would in turn reduce welfare for both consumers and producers.

For an artwork to be created in the absence of copyright law protections would require costly direct bargaining between producers and consumers - these are transaction costs. Exceptions limit the scope (coverage) of copyright, and are economically justified when transaction costs are so high that they would prevent a copyright transaction from taking place. If no efficient and transaction-cost-reducing TDM licensing system can be designed (and it seems that this could be the case, based on our desk research and interview sessions[13]), then it would appear preferable to legalise unauthorised use by means of a TDM exception. Without such an exception, in these circumstances, TDM would either not occur or would occur on a significantly diminished scale, thereby generating "deadweight loss" for society: welfare losses that benefit neither the producer nor the consumer[14].

What - and whom - to take into account in terms of transaction costs?

TDM is likely to generate positive externalities, similar to the externalities associated with research spending in general. The outcome of research may increase productivity for a large number of agents and firms, and stimulate GDP growth, thereby benefiting many people. These benefits are not accounted for in the negotiations between a copyright holder and a researcher. Instead, such bargaining will tend primarily to be a function of the copyright holder's private benefits and the researcher's research budget, so that any spill-over effects on other people's welfare are not accounted for. If we take the example of TDM for medical research,

---

[13] TDM licensing would involve low transaction costs if it involved only one copyright holder, say a single journal publisher or database owner, and one user. The two parties could negotiate a deal directly. However, mining covers hundreds of articles, from different publishers, thus the direct single negotiation is not feasible.

[14] http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf#view=fit&pagemode=none

how could all potential beneficiaries be involved in a negotiation with the copyright holder about accessing a medical database for TDM purposes? With incomplete information on potential (future) users of the data, rights holders cannot price discriminate accurately. The result is that copyright holders are not able to appropriate all of the value of the works to which they hold the rights, and will instead tend to maximise their private returns without consideration of the wider social benefits and externalities[15].

The claimed benefits arising from TDM are typically argued under the assumption of "cross-pollination", by linking and processing previously disjunct datasets. This relies on unencumbered relationships between a large number of providers and processors – which concentration defeats.

Why distinguish between commercial and non-commercial TDM?

If creating a generic TDM exception is the best way forward, as explained above, what remains to be discussed is whether that exception should deal only with non-commercial/research activities, or whether instead of seeking to make any distinction between commercial and non-commercial/research TDM, it should not instead be expressed in terms of a set of principles, thus leaving the management of any financial rewards to take place further down the chain.

Indeed, many consider that the justification for a research exception rests on the assumption that research results benefit society at large (i.e., the public). At the same time, many seem to assume that all research results are freely available. If the availability of any research exception for TDM could depend on the potential benefit for society, this could allow removal of the focus on the commercial / non-commercial question. This would mean that a research exception for TDM and open access to scientific results could go hand in hand.

The potential risk posed by 'commercial' research does not reside in the legal status or private motives of the researchers or their organisation, so much as in the potential for the original copyright owner to suffer sales displacements or losses: it is an economic risk. Excluding research by private companies is not a good criterion on which to gauge or reduce that economic risk. Academic research may also lead to the development of commercial products at a later stage. For example, much university research in bio-medical, genetic and natural science could well result in commercial products. University research necessarily rivals and competes with privately-financed research. However, that does not imply that the output of TDM activities, whether privately or publicly financed, would

[15] Idem

substitute for the revenue that copyright holders derive from the data on which the TDM activities were carried out[16].

It is both artificial and counter-productive to distinguish between commercial and non-commercial/research regimes at the very beginning of a TDM project.

TDM occurs because there is a legal framework which encourages the kind of exploration needed for useful and valuable information to be uncovered through TDM. The extraction of facts is not always an automatic promise of the value of the activity, thus imposing financial or other costs at the beginning of the process could prove counter-productive. As mentioned above, the work produced in a non-commercial research lab in a university might easily subsequently be used by a pharmaceutical or business analytics company for commercial purposes. Another example is when a company invests in research efforts which might not come up with anything meaningful for the scientific world and the company stops the project.

Given the lack of any clear definition of the 'non-commercial' aspect (is it the activity, or the entity carrying out the activity which makes the difference?), these two examples show that it is both artificial and counter-productive to distinguish between commercial and non-commercial/research regimes at the very beginning of a TDM project.

The argument of fair compensation

Those whose data is mined could make a *subsequent* argument asserting an entitlement for them to share in the commercial proceeds of the TDM activity. Data owners do not have the right to block experimentation and enterprise, but they might have an arguable right to a share of the commercial benefits that are achieved through the mining, analysis and use of data which is to be found in works to which they control the copyright. Looking at such a right for fair compensation of the right-holder (i.e., modelled on the private copying levy), the justification comes from the fact that whilst the right-holder may not be able to prevent the use of the work, nevertheless he/she might be entitled to payment of fair compensation, which could encourage rights owners to invest in making their databases available in usable, minable formats. On the other hand, calculating what amounts to fair compensation in specific cases could prove very difficult. In cases where right-holders have already received payment in some other form, for instance as part of a licence fee, it could be that no specific or separate payment is due. Moreover, the collection and distribution of fair compensation payments would necessarily occur through some scheme of collective rights management, with the drawbacks already mentioned above. In order to be sustainable and to avoid the need for future legislative updates, the provision should be drafted in neutral terms, sufficient to withstand the passage of time and likely changes in the associated technology.

[16] http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf#view=fit&pagemode=none

Recommendation

Data owners should be protected from practices that negatively affect their revenue, as opposed to practices that do not affect that revenue. Even this statement needs qualification: data owners should be protected against practices that negatively affect revenue in so far as this would reduce overall social welfare.

Europe needs a regime which enables any researcher, citizen, company or other entity to engage in TDM activities, using material to which they have lawful access, wherever they feel there is a good idea. The exact commercial rewards can be managed at subsequent stages, depending on the implementation of the mining outcome. The TDM exception should not be tailored using an approach which in any way blocks the innovative progress of science, regardless of the actors who contribute to this progress. Instead, the exception and its drafting should take into account that the outcome of TDM activities can be used in different ways, and with different associated economic benefits or rewards. For this purpose, an interpretative instrument, a 'fair use' approach (like the one used in the US) or some kind of "open norm" provision[17] might be the best way forward to tailor any TDM exception. This approach has the merit of allowing for flexibility in how the different elements, including commerciality, are factored in (bearing in mind that commerciality should not be the decisive factor, even if it plays a certain role). The transformative effect of creating entirely new knowledge from something that otherwise would have not been exploited should play a more important role than any commercial effect, because miners are creating welfare for the entire economy and society. If commercial text and data mining is allowed without further constraints or authorisation requirements, this can be expected in turn to result in more competition, broader services, and more creation and dissemination of knowledge.

The argument in support of limiting any TDM exception to activities conducted solely for research purposes has to do with the benefits for society as a whole of the research findings and associated solutions. However, confining the exception solely to non-commercial research activities may slow down the pace of innovation, for it is far from the case that only non-commercial research generates socially and economically valuable outcomes. Access to TDM can be expected to increase the productivity of research, because it increases research output with unchanged labour inputs. Both commercial and non-commercial research can be welfare-enhancing for society, and should therefore be stimulated by the IPR regime. Consequently, there is no valid economic argument to support a distinction between privately and publicly-financed TDM. A well-designed copyright regime should provide appropriate stimulus for all types of research, together with an appropriate level of protection for all rights owners. Once this balance has been reached, there is no reason to distinguish between commercial and non-commercial research. Data owners should be protected from practices that negatively affect their revenue, as opposed to practices that do not affect that revenue. Even this statement

[17] See (e.g.) the suggestion here: https://juliareda.eu/copyright-evaluation-report-explained/#open-norm

needs qualification: data owners should be protected against practices that negatively affect revenue in so far as this would reduce overall social welfare. In some cases, negative revenue effects may be more than compensated for by welfare benefits. Only a small contingent of stakeholders is reluctant to permit TDM to take place without prior authorisation, either because TDM is seen as a potential source of extra income or as a risk factor for their competitive interests. The challenge is to convince these stakeholders that the public interest in allowing TDM for research purposes prevails over individual royalty seeking / revenue-generating behaviour and, that the perceived risks are more theoretical than actual.

Framing the right approach for TDM in a legislative reform might amount to a comprehensive generalised exception for TDM, because this is a liberalised approach that allows everyone to decide what to do with their content. However, there should be some safeguards for organisations whose data is subject to subsequent commercial exploitation following mining. That could be a matter of negotiation between the parties, as it is the case in any business deal - within a framework of law which needs to be both as clear, whilst also as realistic, as possible. The framework also needs to represent or include a mechanism for arbitration when negotiation does not result in agreement. The primary right to monetise should belong to the person who had the mining idea, did the work - and paid for the work. But there should be a responsibility (supported by, or reflected in, the law) to consider the value of the text and data being mined in any consideration of the allocation of rewards from its subsequent reuse.

Any discussion of money and rights should take place much further down the chain.

Any discussion of money and rights should take place much further down the chain. The protection could be considered at the point at which some clearly commercially beneficial project, product, service, business or company has emerged. This is not unusual in the world of patents and could be replicated. In this way, scientists, publishers and the public at large could focus on solving the underlying scientific questions, instead of wasting or misallocating precious resources on administrative burdens before even starting to work on what could prove to be revolutionary ideas.

In order to be sustainable and to avoid the need for future legislative updates, the provision should be drafted in neutral terms, sufficient to withstand the passage of time and likely changes in the associated technology.

MINERS' OR PUBLISHERS' FEARS INFLUENCING THE TECHNICAL ARRANGEMENTS TO ENABLE TDM, AND THE IMPACT OF CHOICE OF DATA HANDOVER POINT

Many times, researchers are not IT experts. They need readily available tools, software and platforms, which can easily convert the input corpus in machine readable format. These miners are fine with using publishers' APIs. There are also other miners, who want to use their own APIs for different reasons, including for fear of revealing valuable insights into their research projects if they do their mining using publishers' APIs. These need and request publishers to make available content to be run through their own API.

Legally however, it might prove difficult to compel a database owner to provide the possibility to download externally. If a publisher offers an API and allows TDM to happen with this API on the source corpus, it might be legally difficult to impose an obligation requiring the database owner to allow downloading to take place.

Most publishers force miners to use their own APIs; this could in some cases be based on publishers' fears that (at least some) researchers would download the articles and re-sell them for entirely different purposes, thus starting to compete with the publisher. In practice, this would mean that the miner would be downloading copyrighted content for purposes other than TDM, and instead selling or redistributing the content, either in original form or for example in translated form.

And then there are alternative third party platforms (such as CrossRef), which provides a platform where miners can access numerous articles that have been made available by publishers (depending on the terms of agreement provided for each miner) and which exploit a common metadata format. Although CrossRef does not solve the licensing challenges described above, it does however provide an easier way to mine in the current jungle of potential TDM data sources. CrossRef represents members ranging from open access publishers to commercial publishers. Although it does not solve the licensing problems associated with TDM, the solution provided by CrossRef has been very positively received.

Recommendation

Although sometimes difficult, it would be useful to understand the intention of choosing the data handover location. If the goal or intention of requiring the data to be mined on the provider's or publisher's servers is based on an intention to retain a certain degree of control over what the miner is doing, this approach is too onerous and should not be encouraged. On the other hand, if miners use TDM as a bogus pretext for downloading and commercially exploting copyrighted material, this would represent a clear copyright infringement that would be actionable in court. Publishers argue that going to court is too costly and lengthy and prefer

"prevention rather than cure", even if this means imposing additional technical obstacles that impede miners from downloading copyrighted content onto their own servers for TDM purposes.

The recommendation coming out of the interview discussions was to develop a code of good practice, to which miners would sign up in return for being able to make their own copy of the publisher's dataset(s). Account also needs to be taken of the fact that sometimes the mining request may come from a mere IP or network address (e.g., one belonging to a university), making it difficult to establish the true identity of the individual miner. If miners are required to provide too many personal details, this could well also be received with resistance, e.g., for fear of too much invasion in the research project information being revealed.

If current trends continue to develop as they are today, everyone will eventually develop their own unique API, and we will end up with a proliferation of thousands of APIs, which will do nothing to simplify the complex existing situation.

Another point to take into account is that if current trends continue to develop as they are today, everyone will eventually develop their own unique API, and we will end up with a proliferation of thousands of APIs, which will do nothing to simplify the complex existing situation; however, we also note that various current projects are also on foot today, which aim to standardise the creation of these APIs. Even if TDM is to be allowed through a generalised exception, APIs will still be needed to do the actual mining. Trusted third party platforms which make APIs available should be encouraged; indeed, having a *trusted* third party in the mining process could provide a middle ground where publishers feel more confident that their content is not about to be misappropiated, and where miners feel they can engage in TDM without their project being put at risk of plagiarism or other sharp practice.

Any legislative solution should avoid fear of copyright infringements leading to technical measures being deployed so as to limit the practice of TDM. Provided that infringements remain actionable in court, putting in place technical obstacles to avoid copyright infringements appears to constitute the unwarranted blocking of innovation.

Bringing all stakeholders around a table would appear to be the most advisable solution, not least because there remains a degree of mistrust between some publishers and some researchers. Sometimes the presence of diverging interests can motivate such tension, but in other cases there can indeed be factors or aspects to which one category of stakeholder rightfully points, but which are not always foreseeable or even obvious for other categories of stakeholder.

The "Licences for Europe" exercise was a good approach - the Commission could try to replicate that approach here, taking into account the reasons for its failure.

It is encouraging that those participating appear (perhaps to varying extents) to share a genuine intention to find the best solution for scientific development, the advancement of society, the promotion of Europe's innovation and competitiveness, and so we should be able to avoid starting from the desired end result and building towards that.

CONCLUSIONS

The Commission needs to provide coherence and harmonisation for TDM across Europe, through a regulatory intervention proportional to the benefits of TDM and the costs of non-intervention. While doing so, the Commission should aim to achieve coherence in the legal provisions which it seeks to apply to TDM, with no consideration of 'commercial' versus 'non-commercial' purposes. Europe needs a regime which enables any researcher, citizen, company or other entity to engage in TDM activities, using material to which they have lawful access, wherever they feel there is a good idea. The exact commercial rewards can be managed at subsequent stages, depending on the implementation of the mining outcome. The protection could be considered at the point at which some clearly commercially beneficial project, product, service, business or company has emerged.

Confining the exception solely to non-commercial research activities may slow down the pace of innovation, for it is far from the case that only non-commercial research generates socially and economically valuable outcomes. Although an agreement for providing specific products and services could be concluded, and the owner can put in place specific mechanisms (e.g., a pay-wall) or require specific conditions for access that limit the use depending on the purpose, this should remain an opt-in choice and the Commission should not introduce a legal system which directly imposes such an artificial limitation between commercial and non-commercial purposes.

A generalised exception for TDM represents the needed liberalised approach, allowing everyone to decide what to do with their content. Many voices echo the fact that licences are not an alternative to a mandatory exception. Being aware that publishers do not always push for a licensing approach purely out of a desire to maximise the royalties that they receive, but also fears of overloading (or even blocking) the publisher's website or other servers as a consequence of having allowed too much traffic, a balance should be found between the measures imposed by publishers to avoid website/server overload or piracy and the actual needs of the miners. Using technical means to avoid piracy imposes a disproportionate burden for researchers, and it also creates a serious obstacle for research, whereas arguments based on the need of publishers to avoid the risk that their websites or other servers might be overwhelmed is yet to be better evidenced and requires proper substantiation.

Even if TDM is to be allowed through a generalised exception, APIs will still be needed to do the actual mining. Trusted third party platforms which make APIs available should be encouraged. Having a trusted third party in the mining process could provide a middle ground where publishers feel more confident that their content is not about to be misappropiated, and where miners feel they can engage in TDM without their project being put at risk of plagiarism or other sharp practice.

Bringing all stakeholders around a table would appear to be the most advisable solution, not least because there remains a degree of mistrust between some publishers and some researchers. Sometimes the presence of diverging interests can motivate such tension, but in other cases there can indeed be factors or aspects to which one category of stakeholder rightfully points, but which are not always foreseeable or even obvious for other categories of stakeholder.

In order to be sustainable and to avoid the need for future legislative updates, the provision should be drafted in neutral terms, sufficient to withstand the passage of time and likely evolution of the associated technology.

AUTHORS

Diana Cocoru is the Head of Policy and Research Development at OFE. She holds a degree in Law, one in International Economic Relations and a Master in European Business Law. She gained experience in the decision-making process while working for a Member of the JURI committee (European Parliament). Diana promoted the benefits of open source software while working for an NGO with impact in developing countries. Then she joined OpenForum Europe, where she deals with copyright among other policy dossiers, closely monitoring policy developments around text and data mining.

Mirko Boehm is the CEO of Endocode, where he specialises in consulting to and mentoring small to large businesses on complex software development endeavours, the use of open source products and methods in organizations, and software-related issues of business strategy and intellectual property. The Open Invention Network protects the open source ecosystem by acquiring patents and licensing them royalty-free to entities. Mirko is responsible for the Linux System Definition, which defines the technical scope of the patent non-aggression agreements. Mirko is an OFA Fellow and the co-chair of the OFE IP Task Force.

For any question please contact Diana Cocoru, at *diana@openforumeurope.org*.

OpenForum Europe (OFE) *is a not-for-profit, independent European based organisation which focuses on openness within the IT sector. We draw our support not only from some of the most influential global industry players, including Google, IBM, Oracle and Red Hat, but most importantly from across the European SME and consumer organisations and the open community. OFE also hosts a think tank, focussed around our global network of OpenForum Academy Fellows, each contributing significant innovative thought leadership on core topics. Views expressed by OFE do not necessarily reflect those held by all its supporters.*

OFE Limited, a private company with
liability limited by guarantee
Registered in England and Wales
with number 05493935
Registered office: 1 Blunt Road,
South Croydon, Surrey CR2 7PA, UK