"A right to 'read' for machines: assessing a black-box analysis exception for data mining"

Marco Caspers
University of Amsterdam
Institute for Information Law
Vendelstraat 7, 1012 XX
Amsterdam, The Netherlands
M.Caspers@uva.nl

Lucie Guibault
University of Amsterdam
Institute for Information Law
Vendelstraat 7, 1012 XX
Amsterdam, The Netherlands
L.Guibault@uva.nl

ABSTRACT

This panel looks into the impact of the current copyright framework in the European Union on text and data mining (TDM) and discusses the impact of introducing a TDM exception in EU copyright law. A design of this exception is proposed for the panel, and is partially based on findings in the Horizon 2020 FutureTDM project. This project aims to improve uptake of text and data mining (TDM) in the EU and, in that regard, has studied the legal barriers to TDM and will be developing and recommending a policy framework in the future. Part of this policy framework will consist of possible actions to be undertaken by the - European and national - legislators.

A TDM exception is considered to include in the recommendations and we therefore broach the topic to discuss the possible legal, economic and practical impact of such an exception with experts from the field. The TDM exception, as proposed for this panel, is inspired by the "black-box analysis" exception from the Software Directive, which allows lawful users of a program to perform any of the acts of loading, displaying, running, transmitting or storing the program to "determine the ideas and principles" underlying it. The authors of the panel believe that this underlines the general principle of copyright law: namely, that ideas and facts are not protected. Therefore, proposition to be discussed is that a similar exception should be introduced for copyright law in general, that would allow reproductions to be made of works for the sole purpose of extracting facts and ideas underlying them. This would allow TDM activities, where machines 'read' lawfully accessed works just as the human reading of works does not require further authorization from the copyright holder.

Keywords

EU copyright law, copyright exception, TDM (text and data mining), software protection, access to ideas.

ASIST 2016, October 14-18, 2016, Copenhagen, Denmark.

INTRODUCTION

This panel is organized by partners in the FutureTDM project consortium.[1] FutureTDM aims to improve uptake of text and data mining (TDM) in the EU by actively engaging with stakeholders such as researchers, developers, publishers and SMEs. The partners in the FutureTDM consortium share the ambition behind the EC's call to develop policy and legal frameworks to reduce the barriers of TDM uptake and with it, promote the awareness of TDM opportunities across Europe. As a result, the consortium offers a concept that not only focuses on the required identification, assessment and analysis of current TDM obstacles, but also creates a practitioner-driven emphasis through engagement of workshops and discussions. An outcome of the FutureTDM project will include, guidelines that offer informed recommendations to practitioners from various disciplines, and propose solutions to overcome legal and policy barriers impeding TDM opportunities. In the context of overcoming legal barriers, we have thought of proposing a TDM exception that we would like to discuss with experts and people from the fields of academia, industry, non-profit organizations and governments.

PANEL PROPOSITION: AN EXCEPTION FOR TDM

Introduction

As a basic principle of copyright law, ideas are not protected. It is rather the original expression of ideas that copyright law seeks to protect. This principle prevents ideas being monopolized by authors of works, thereby ensuring the free flow of ideas in society. Other individuals may reuse those ideas and express them in their own – original – way. With the advent of computer technology, the ideas expressed in digital works are not easily ascertainable without the use of a device, which itself requires the making of a reproduction. As that reproduction falls within the scope of the copyright owner's exclusive right, it becomes literally impossible to gain access to the ideas at the root of a work without infringing the owner's copyright,

[1] See for further information on this Horizon 2020 project: http://www.futuretdm.eu/news/about-futuretdm/.

unless the act of reproduction is authorized by law or by the owner. This problem also arises in the context of text and data mining (TDM) activities.

Problems under current copyright framework

Directive 2001/29/EC on Copyright in the Information Society ('Copyright Directive') obliges Member States to grant authors of works the exclusive right to make reproductions, communicate (or make available) and distribute their works to the public. Although Member States have some leeway to grant other exclusive rights to authors under national law, the mere act of accessing a work does generally not fall under the ambit of any exclusive right. As a consequence, someone who has lawfully acquired a book cannot be prevented from reading the book. And, since ideas are not protected under copyright law, the reader can re-use the ideas behind the book to express and discuss them in his own original way. This is consistent with the idea-expression dichotomy and this would enable scientists to discuss, verify and build on existing (published) research, knowledge and ideas – to which they have lawful access to – without making any copyright infringement.

The amount of published knowledge has increased exponentially over the last decades. With the help of TDM techniques, this has opened up opportunities to mine large collections of works to find certain patterns and it has enabled academics to keep up with an overload of publications in some fields through machine assisted literature review. In these cases, machines rather than human beings have access to the contents that are to be mined.

This is problematic from a copyright perspective. As opposed to access and reading by human beings, a computer cannot access and 'read' a work without making any reproductions, whether they are temporarily made in its random-access memory (RAM) or (more) permanently in a long-term storage such as a hard-disk. The making of such copies undoubtedly falls within the scope of the reproduction right, which must be interpreted rather broadly. This is explicitly stated in the preamble of the Copyright Directive and, according to the CJEU in its Infopaq decision, follows from the language of Article 2 of the Copyright Directive, stating that authors have an exclusive right to "authorise or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part" of their works. Therefore, even though the person controlling the computer may have lawful access to those works and merely makes the copies for the purpose of extracting ideas and facts from the works, these acts of reproduction require the copyright holder's authorization; the person has a 'right' to read the work, while the computer has not.[2]

We do not consider such an approach to be consistent with the basic principle of copyright law, which excludes the protection of ideas. Although TDM may be exempted from copyright law by certain exceptions, this will only solve the problem under certain circumstances. For example, reproductions made for the purpose of scientific research may only be permitted for non-commercial purposes. This may become problematic within a changing academic field where the border between what is academic and what is commercial is blurring. Moreover, the implementation of copyright exceptions in European's national copyright laws is fragmentary, leading to legal uncertainty and jurisdictional issues in cross-border collaborations. But above all, these exceptions do not solve the underlying problem: that the extraction of ideas and facts is not necessarily allowed under the current copyright system.

Solution: copyright rules for software

Fortunately, the European legislator has already addressed this issue earlier in relation to the copyright protection of computer programs. Where computer programs are protected under copyright law, Article 5(3) of the Software Directive (2009/24/EC) allows for the so-called black box analysis of that software. More precisely, it entitles the person who has the right to use the program to "observe, study or test the functioning of the program in order to determine the ideas and principles which underlie any element of the program if he does so while performing any of the acts of loading, displaying, running, transmitting or storing the program which he is entitled to do".

Likewise, we propose that lawful access to works *in general* should imply that reproductions are allowed to be made for the sole purpose of determining the ideas behind a work. This could be achieved in the form of a copyright exception, but it should rather be formulated as a general limitation to the scope of the reproduction right: to the extent reproductions are made for the sole purpose of extracting unprotected ideas from a work, it should not fall within the scope of the author's exclusive right. This would not only permit many acts carried out in TDM activities, it would also be a future proof and technology neutral instrument that allows any process of knowledge extraction that involves reproductions that should not be relevant from a copyright perspective. Such an exception will of course have its boundaries, for which the criteria for a lawful decompilation of a computer under the Software Directive provides useful insights. For example, reproductions may not be used for other goals than solely extracting ideas and they should also be necessary for that purpose.

---

[2] The latter expression is derived from the proposition "The right to read is the right to mine", as coined by Murray-Rust (2012).

The exceptions in the Software Directive are mandatory for Member States to implement. A limitation on the reproduction right for the making of copies for the sole purpose of extracting ideas should also be made mandatory, preferably in a way that right holders may not prevent such reproductions in any way. Currently, of the twenty-one exceptions provided by the Copyright Directive, twenty are optional. This has resulted in a fragmentary copyright landscape in the EU. For example, according to Caspers & Guibault (2016), the exception of Article 5(3)(a), which enables Member States to allow reproductions made solely for scientific research purposes without the right holder's consent, is implemented very narrowly – if at all – in many Member States; only in some jurisdictions would it possibly exempt reproductions made in TDM activities. This is very likely to affect TDM researchers who work or collaborate on a cross-border scale. Therefore, our proposed limitation would have to be mandatory to prevent a fragmentary copyright landscape in Europe, which would especially be problematic where the fundamental principles of copyright are affected.

ISSUES TO BE DISCUSSED

The panel will discuss the proposed exception, that is designed according to the principles of Article 5(3) of the Software Directive, from three angles:

1. Legal perspective: legal feasibility
2. Economic perspective: economic consequences
3. Practical perspective: impact on future TDM practice

Legal perspective

From a legal perspective, the proposed exception will be discussed as regards its legal feasibility: how does it fit within the European and international copyright framework? For example, how does this exception relate to the functions of copyright works and the authors' interests in content appropriation.[3] Moreover, it is discussed whether the scope of such an exception is sufficient to cover the aimed uses: the making of reproductions by computer hardware and the extraction of unprotected knowledge, facts and ideas from the works.

Economic perspective

From an economic perspective, the consequences on the market of copyright works will be discussed, with a particular focus on the market position of several stakeholders and possible changes in the value chain. For example, panelists will discuss the effect on exploitation opportunities of copyright works and the economic gains that may be achieved by such an exception.

Practical perspective

From a stakeholder perspective, the practical feasibility of the proposed TDM exception will be discussed by the panellists. The focus will be on whether such an exception will be the right solution. From a miner's perspective, it will be assessed whether this exception will provide him or her the right legal instrument and certainty to mine lawfully accessed works, without fearing any legal consequences from right holders.

DESIGN OF THE PANEL

Schedule

The panel will start with two presentations by partners in the FutureTDM consortium to introduce the issues and the proposed solution in the form of a copyright exception. Each presentation will last 5 to 10 minutes. This will be followed by a presentation from each external expert of 10 to 15 minutes per presentation. This will allow a discussion and Q&A with the audience for 25 to 40 minutes.

Contributions by FTDM consortium partners

Freyja van den Boom

Freyja van den Boom is researcher on law, ethics and technology. She is currently affiliated with Open Knowledge and, in that capacity, involved as a partner within the FutureTDM consortium.

Freyja van den Boom will share her findings from the Knowledge Cafés that she organized in the course of the project, as well as from the thirty stakeholder interviews she conducted, with a focus on how barriers in copyright law are regarded by different stakeholders. Some findings are already published in deliverables 2.3 and 4.3, which are available at http://www.futuretdm.eu/knowledge-library/.

Marco Caspers (moderator)

Marco Caspers (LLM) is a researcher in copyright law at the Institute for Information Law (University of Amsterdam), with a main focus on the intersection of copyright law and technology. He also works for the Technology Transfer Office of the University of Amsterdam. As partner in the FutureTDM consortium, he maps the legal barriers to TDM, as well as the possible solutions to overcome such barriers.

Marco Caspers will present the findings from the FutureTDM project with regard to the way copyright law serves as a barrier to TDM. The results follow from a comparative analysis, carried out by Caspers & Guibault (2016), between the implementation of copyright exceptions in different Member States, as well as an analysis of the European copyright framework. This will be followed by a proposition for a TDM exception that the panel will discuss. Marco will moderate the discussion.

[3] Cf. Borghi & Karapapa (2011).

Contributions by invited experts

Aleksei Kelli

Aleksei Kelli is Professor of Intellectual Property Law (the University of Tartu, Estonia). He has acted as the Head of an Expert Group on the Codification of the Intellectual Property Law, the Ministry of Justice of Estonia (2012-2014). Aleksei has been the main intellectual property expert in research and innovation policy monitoring program coordinated by the Estonian Ministry of Education and Research. Dr Kelli is responsible for the management of IPRs of digital language resources at the University of Tartu and the Institute of the Estonian Language and he is an appointed intellectual property expert to CLARIN (Common Language Resources and Technology Infrastructure) Committee for Legal Issues.

As a legal expert, Dr Kelli will share his views on the legal feasibility and possible implementation of the proposed TDM exception and will contribute on the general discussion. He can use his experience from the copyright reform debate in Estonia, which included the possibilities of an exception for TDM (Kelli et al. 2012).

Matěj Myška

JUDr. Matěj Myška, Ph.D. is a Senior Assistant Professor at the Institute of Law and Technology, School of Law, Masaryk University and editor-in-chief of the Review of Law and Technology ("Revue pro právo a technologie"). Since 2013, he has been cooperating with the Technology Transfer Office of the Masaryk University as a lawyer and Creative Commons Czech Republic as the Legal Lead. His professional focus lies in private ICT law, especially digital copyright.

As a legal expert with a background in copyright and digital developments, Dr Myška will share his views on the legal aspects of the proposed TDM exception and will contribute on the general discussion.

Christian Handke

Christian Handke is Assistant Professor (tenured) of Cultural Economics at Erasmus University Rotterdam. He is program coordinator of the Master in Cultural Economics and Entrepreneurship. Since 2012, Dr Handke also works as Senior Researcher at University of Amsterdam, where he participates in the research project on Copyright in an Age of Access.[4] His research focuses on cultural economics and the economics of copyright, innovation and technological change, as well as the record industry.

As an economist, Dr Handke will focus on the empirical evidence as regards the incentive argument in favor of copyright protection and in particular for the argument that a TDM exception would promote academic research of this type.

Penny Labropoulou

Penny Labropoulou is a Senior Researcher at the Institute for Language and Speech Processing, Research Centre "Athena", with a specialisation in Language Technology and Computational Linguistics. Since 2010, she has been actively involved in infrastructures (META-SHARE, CLARIN and OpenMinTeD)[5] empowering the sharing of knowledge and language resources and language technologies, the interaction of data resources and web services/workflows and their exploitation in the advancement of research. Through her involvement in metadata modeling activities for language resources, she has taken a particular interest to legal metadata and, in general, legal processes and instruments related to their deployment.

Drawing from her experience in the fields of NLP and TDM, Penny Labropoulou will present the legal complexities and challenges TDM experts and end users face, especially when combining datasets and software from various sources, and discuss how the proposed TDM exception can alleviate the "legal chaos" burden from such activities.

Discussion

After the presentation from the FutureTDM project partners and the three invited experts, there will be approximately 25 to 40 minutes left for a discussion among the panelists and the audience. Having a panel with various approaches, the pros and cons of a TDM exceptions can be weighed from different angles. Given the mixed audience that attends the Annual Meeting of ASIS&T, the discussions will provide a lot of input on the proposed copyright exception, as well as further issues that may need to be further addressed by the project.

ACKNOWLEDGMENTS

We want to like to thank all the people, and in particular the partners in the FutureTDM consortium, who have provided their input and reflections on the design and organization of this panel.

REFERENCES

Borghi, M. & Karapapa, S. 'Non-Display Uses of Copyright Works: Google Books and beyond.' *Queen Mary Journal of Intellectual Property* 1(1), 21-52.

Caspers, M. & Guibault, L. (2016). *Baseline report of policies and barriers of TDM in Europe* (FutureTDM Deliverable 3.3). Retrieved from

[4] See for more information on the project: http://www.ivir.nl/onderzoek/acs.

[5] See for more information on these projects: http://www.meta-share.eu, https://www.clarin.eu and http://openminted.eu.

http://www.futuretdm.eu/knowledge-library/?b5-file=2374&b5-folder=2227.

Kelli, A., Tavast, A. & Pisuke, H. (2012). 'Copyright and Constitutional Aspects of Digital Language Resources: The Estonian Approach.' *Juridica International* (XIX), 40-48.

Murray-Rust, P. (2012). The right to read is the right to mine. Retrieved July 21, 2016 from https://blogs.ch.cam.ac.uk/pmr/2012/05/31/the-right-to-read-is-the-right-to-mine/.