

## Extended Abstract:

# Why We Need a Text and Data Mining Exception (But it is Not Enough)

Thomas Margoni, Giulia Dore

School of Law

University of Stirling, UK

E-mail: thomas.margoni@stir.ac.uk, giulia.dore@stir.ac.uk

### Abstract

Text and Data Mining (TDM) has become a key instrument in the development of scientific research. Its ability to derive new informational value from existing text and data makes this analytical tool a necessary element in the current scientific environment. TDM crucial importance is particularly evident in a historical moment when the extremely high amounts of information produced (scholarly publications, databases and datasets, social networks, etc.), make it unlikely, if not impossible, for humans to read them all. Nevertheless, TDM, at least in the EU, is often a copyright infringement. This situation illustrates how certain legal provisions stifle scientific development, instead of fostering it, with significant damage for EU based researchers and research institutions and for the European socio-economic competitiveness more in general. Other countries leading the scientific and technological development have already implemented legislative or judicial solution permitting TDM, also for commercial purposes.

This extended abstract suggests, as it has been already advocated in literature and in policy documents, that a mandatory TDM exception, not limited to non-commercial research, is needed to bring the EU on the same level playing field as other jurisdictions, such as the US and Japan.

However, this extended abstract further argues that, while in the short-term a TDM mandatory exception can and should be implemented by the EU legislator, by way of a harmonising Directive(s), for the long-term sustainability of the EU copyright framework, a broader, general and technology-neutral exception should instead be considered. The latter should take the form of a fair use like standard and indeed be part of a more structured intervention in the field of copyright, by means of a Regulation that would provide uniformity to the whole EU copyright framework.

**Keywords:** text and data mining, copyright exceptions and limitations, fair use, EU law.

## 1. Introduction

The increasing role played by Text and Data Mining in today's research sector is demonstrated by the attention that institutions, case law, policy documents and scholarly literature is dedicating to this topic (Brook et al, 2014; De Wolf, 2014). Overall, TDM potentialities have been widely illustrated by recent studies that established how mining existing content appears to be a crucial tool that serves both scientific and economic progress (JISC, 2012). One of TDM's most powerful features resides in the possibility for researchers to derive new information from the exterminated amount of existing knowledge.

Nevertheless, especially in the EU, TDM often represents an act of copyright infringement, or better a Sui Generis Database Right (SGDR) infringement. In fact, the current EU legal framework requires that all acts of reproduction, even if temporary, partial and indirect, be authorised by the right holder (see Art. 2 Directive 2001/29/EC and Art. 7 Directive 96/9/EC). Accordingly, to the extent that it is necessary to make such a temporary and transient copy for TDM purposes, TDM constitutes a copyright (or most likely SGDR) infringement.

As it is known and well documented in the literature, the section of the EU legal framework that should balance the broad protection afforded to copyright holders (mainly Art. 5 Directive 2001/29/EC, but also Articles 6 and 9 of the Database Directive 96/9/EC) have been drafted following a different paradigm: 21 exceptions listed exhaustively (i.e. Member States cannot create additional ones), but not mandatory, except for one (i.e. of the remaining 20 Member States can decide which

ones to implement). It is clear how this provision not only fails to harmonise EU copyright law in the field of exceptions and limitation, but also creates a strong unbalance in the relationship between the protection of the legitimate interest of right holders on the one hand, and the protection of other fundamental rights such as freedom of expression, which includes the freedom of artistic expression and scientific inquiry, property and the freedom to conduct a business, on the other (Hugenholtz, 2000; Guibault, 2010).

The resulting situation impacts directly on the legitimacy of TDM (De Wolf, 2014) because, on the one hand it does not allow MS to create new exceptions to address scientific development, while on the other fails to achieve the objective of a harmonised internal market in the field of copyright.

At this regard, the paper will argue that a TDM exception, not limited to non-commercial purposes, as suggested by the Hargreaves report (Hargreaves, 2011) should be implemented as soon as possible. Nevertheless, this type of exception will not probably stand the test of technological development. In two, three or five year time, when the new scientific breakthrough in the field of data analysis will be ready, the EU will have to go through this same, inefficient process once again, losing again in terms of competitiveness in favour of other more flexible legal systems.

TDM is but another example that what the EU really needs is a broad, flexible and technology-neutral standard to address the complex relationship between and among right holders, citizens/consumers and technological progress. A European fair use standard as part of a systematic intervention to uniformise EU copyright law.

## 2. The EU legal framework

### 2.1. Copyright

Art. 2 of Directive 2001/29/EC requires that all acts of reproduction, even if temporary, partial and indirect, need to be authorised by the right holder. The Directive clarifies that a broad definition of reproduction “is needed to ensure legal certainty within the internal market”, however does not offer any evidence of why a broad definition will enhance certainty more than, for instance, a balanced definition (see Recital 21).

Contrast this, with the fact that all the copyright limitations listed in the InfoSoc Directive (Directive 2001/29/EC), with the exception of Art. 5.1 (acts of temporary reproduction which are transient or incidental and an integral and essential part of a technological process) are not mandatory, but left to the discretion of Member States. The consequence is a fragmented and uncertain legal framework for TDM in the EU in clear contradiction with a harmonised internal market. Clearly, this situation represents a hurdle for the wide adoption of TDM in the EU.

### 2.2. SGDR

The SGDR is a peculiar EU form of protection for databases which are protected regardless of any originality. What is protected here is the “substantial investment” in quantitative or qualitative terms that the maker of the database puts in it. This substantial investment can take the form of time, money, labour or any other resources spent in the making of a DB. Importantly, when talking about “making” the database, the substantial investment has to be in the obtaining, verification and presentation of the data and not in their creation. So for example, a football league cannot benefit from SGDR protection in the fixture lists of the teams playing in the league as these data are considered to be created. The extent to which scientific databases can be said to be constituted by created or obtained data is not clearly settled in case law. In particular, the dichotomy between creating and obtaining data is not necessarily solved at the epistemological level.

The maker of a database qualifying for SGDR protection enjoys two main exclusive rights: the right to prevent extraction, that is to say the permanent or temporary transfer, of a substantial part of the database; the right of re-utilisation of the database, namely making them available to others.

Exceptions and limitation to SGDR are even narrower than those accorded to copyright, yet they are listed following the same exhaustive but not mandatory technique. MS have the faculty to exempt uses for private purposes (only for non electronic databases); illustration for teaching or scientific research (to the extent justified by the non commercial purpose to be achieved); and for public security or administrative or judicial procedure (Art. 9 Database Directive).

## 3. National examples

In the United States, courts have established that acts of web and text and data mining are transformative and therefore are covered by the fair use defence, regardless of whether they are conducted for commercial purposes

(*Authors Guild, Inc. v. Google, Inc.*, 954 F. Supp. 2d 282, 291 (S.D.N.Y.2013); Aff'd 2015 2d Circuit; *Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014); see in general the study of the US Association of Research Libraries (ARL, 2015)).

Other countries, such as Japan, have drafted specific TDM exceptions not limited to commercial purposes (Japan Copyright Act, Article 47septies).

Within the EU, the UK has recently implemented a TDM exception for lawfully accessed works or other subject matter. While on the one side the exception cannot be limited by contractual agreements to the contrary, it only operates for non commercial purposes, a limit dictated by the reported EU legal framework (Hargreaves, 2011).

## 4. Conclusions

The EU has only one option if it intends to enjoy the benefits of scientific, technological and economic development in the field of data: the creation of a mandatory exception that clearly and unambiguously allows activities such as TDM. Realistically, this will have to be done in two stages: in the short term a dedicated exception for TDM activities, not limited to non commercial purposes mandatory for all EU MS, by way of an amending directive(s). In the long term, a more systematic intervention to create a uniform internal market for copyright purposes, which should implement a broad, flexible and technology neutral counter balance to exclusive rights: a European fair use.

## 5. Acknowledgements

The authors want to thank all members of the OpenMinTeD consortium and those who have provided their valuable insights and external contribution to the project.

This work has received funding from the European Union's Horizon 2020 research and innovation programme (H2020-EINFRA-2014-2) under grant agreement No. 654021. It reflects only the author's views and the EU is not liable for any use that may be made of the information contained therein.

## 6. Bibliographical References

- Borghi, M. and Karapapa, S. (2012). *Copyright and Mass Digitization*. Oxford, UK: Oxford University Press.
- Brook, M., Murray-Rust, P. and Oppenheim, C. (2014). The Social, Political and Legal Aspects of Text and Data Mining (TDM). *D-Lib Magazine*, 20 (11/12).
- Cox, K. L. (2015). *Text and Data Mining and Fair Use in the United States*, ARL (issue brief).
- Ghafele, R. and Gibert, B. (2012). *The economic value of fair use in copyright law: counterfactual impact analysis of fair use policy on private copying technology and copyright markets in Singapore*.
- Guibault L (2010). Why Cherry-Picking Never Leads to Harmonisation: The Case of the Limitations on Copyright under Directive 2001/29/EC. *Jipitec*, 1: 55.
- Guibault, L. and Wiebe, A. (2013). *Safe to be open. Study on the protection of research data and recommendations for access and usage*. Göttingen: Universitätsverlag Göttingen.
- Guibault, L. and Margoni, T. (2015). Legal Aspects of

Open Access to Publicly Funded Research, in *Enquiries into Intellectual Property's Economic Impact*, pp. 373-414, OECD (report).

Hargreaves, I. (2011). *Digital Opportunity. A Review of Intellectual Property and Growth* (report).

Hugenoltz, B. (2000). Why the Copyright Directive is Unimportant, and possibly invalid, *European Intellectual Property Review*, 2000-11, pp. 499-505.

Margoni, T., Caso, R., Ducato, R., Guarda P. and Moscon, V. (2016). Open Access, Open Science, Open Society, *ElPub2016*, accepted paper.

McDonald, D., Kelly, U. (2012). *The Value and Benefits of Text Mining*, JISC (report).

Triaille, J. P., de Meeûs d'Argenteuil, J. and de Franquien A. (2014). *Study on the legal framework of text and data mining*, DE WOLF (study).

Berne Convention for the Protection of Literary and Artistic Works (Sept. 9, 1886), as revised.

Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, OJ L 77, 27.3.1996, 20–28.

Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001, on the Harmonization of certain aspects of copyright and related rights in the information society, OJ L 167, 22.6.2001, 10–19. Copyright law of Japan, Law No. 48 (May 6, 1970), as amended.

UK Copyright, Designs and Patents Act 1988 (Nov 15, 1988), as revised.

The U.S. Copyright Act, 17 U.S.C. (Oct 19, 1976), as revised.

WIPO Copyright Treaty WCT (Dec 20, 1996)

Trade-Related Aspects of Intellectual Property Rights, TRIPS (Apr 15, 1994).