Scientific Research Publishing

Open Journal of Philosophy, 2019, 9, 369-395
http://www.scirp.org/journal/ojpp
ISSN Online: 2163-9442
ISSN Print: 2163-9434

Text and Data Mining in Directive 2019/790/EU Enhancing Web-Harvesting and Web-Archiving in Libraries and Archives*

Maria Bottis[1], Marinos Papadopoulos[2], Christos Zampakolas[3], Paraskevi Ganatsiou[4]

[1]Associate Professor, Department of Archives, Library Science and Museology, Ionian University, Corfu, Greece
[2]Attorney-at-Law, Independent Researcher, PhD, MSc, JD, Athens, Greece
[3]Archivist/Librarian, Independent Researcher, PhD, MA, BA, Ioannina, Greece
[4]Educator, MA, BA, Prefecture of Ionian Islands, Corfu, Greece
Email: botti@otenet.gr, marinos@marinos.com.gr, christoszampakolas@gmail.com, pganatsiou@gmail.com

How to cite this paper: Bottis, M., Papadopoulos, M., Zampakolas, C., & Ganatsiou, P. (2019). Text and Data Mining in Directive 2019/790/EU Enhancing Web-Harvesting and Web-Archiving in Libraries and Archives. *Open Journal of Philosophy, 9,* 369-395.
https://doi.org/10.4236/ojpp.2019.93024

Received: June 20, 2019
Accepted: August 25, 2019
Published: August 28, 2019

Abstract

Text and Data Mining (hereinafter, TDM) issue for the purpose of scientific research or for any other purpose which is included in the provisions of the new EU Directive on Copyright in the Digital Single Market (hereinafter, DSM). TDM is a term that includes Web harvesting and Web Archiving activities. Web harvesting and archiving pertains to the processes of collecting from the web and archiving of works that reside on the Web. Web harvesting and archiving is one of the most attractive applications for libraries which plan ahead for their future operation. When works retrieved from the Web are turned into archived and documented material to be found in a library, the amount of works that can be found in said library can be far greater than the number of works harvested from the Web. This paper aims at presenting certain issues related to the existing legal framework as well as technical/librarianship issues that apply to TDM which includes Web harvesting and archiving activities. This paper elaborates upon the applicable new provisions of Directive 2019/790/EU on Copyright in the DSM with the aim to shed light upon issues such as the notion of "*lawful access*", the beneficiary of the mandatory exception for TDM, the purpose-specific TDM described in art.3 of the new Directive on Copyright in the DSM, and the application of the "*three-step test*" in TDM.

Keywords

Libraries, Archives, Web Harvesting, Web Archiving, Data Analysis, Text & Data Mining, TDM, Text Mining

*Work co-funded by Greece and the European Social Fund (ESF) through the Operational Program "Human Resources Development, Education and Lifelong Learning" for the implementation of the ESF & the Youth Employment Initiative in Greece.

1. Text and Data Mining in Directive 2019/790/EU on Copyright in the Digital Single Market

Text and Data Mining (TDM) for scientific research or for any other purpose is included in the provisions of the new EU Directive on Copyright in the Digital Single Market (DSM)[1]. TDM is a term that includes Web Harvesting and Web Archiving activities which have been within the interests of public libraries—national libraries in most cases—and archives of countries which leverage on new Information Technology tools and applications aiming to enrich their national repositories with works that reside mostly on the Internet. Web Harvesting and Web Archiving were included in the drafted provisions of the EU Proposal for an Amendment of EU Copyright Law from the outset of the effort to amend EU Copyright law. The term that prevailed regarding the description in law of Web Harvesting and Web Archiving was "Text and Data Mining" or TDM.

The need for a statutory exception from copyright for the sake of TDM and not licensing has long been requested by many organizations such as the International Federation of Libraries (IFLA) (IFLA Statement on Text and Data Mining, 2013)[2]. The existing legal framework in the EU does not contain this exception and the "*acquis Communautaire*" does not cover TDM and is to blame for legal uncertainty in the EU regarding TDM, scientific research, and copyright protection (IFLA Statement on Text and Data Mining, 2013)[3].

[1]The text of the new Directive was published in the Official Journal of the European Union on May 17, 2019; it is the new Directive 2019/790/EU on Copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. See Directive 2019/790/EU on copyright and related rights in the Digital Single Market and amending Directive 96/9/EC and Directive 2001/29/EC available at URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2019.130.01.0092.01.ENG&toc=OJ:L:2019:130:TOC [last check, July 1, 2019]. On April 17, 2019, the final Act of the Directive on Copyright in the Digital Single Market (DSM) was signed after the voting process of April 15, 2019 before the EU Parliament in which the text of the new EU Directive on Copyright in the DSM was adopted. Before that vote and on March 26, 2019, Axel Voss, a German politician and lawyer who serves as the assigned Rapporteur in drafting the new Directive on Copyright, i.e. Proposal COM(2016)593 final 2016/0280(COD), presented the amended proposal before the European Parliament which voted on it and adopted the compromise amendment No.271 to the proposal for a new Directive on Copyright in the DSM. See Proposal for a Directive of the European Parliament and of the Council on Copyright in the Digital Single Market COM/2016/0593 final-2016/0280(COD) available at URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2016:0593:FIN [last check, July 1, 2019]. The objective of the new Directive on Copyright in the DSM is to contribute to the functioning of the internal EU market, provide for a high level of protection for right holders, facilitate the clearance of rights, and create a framework in which the exploitation of works and other protected subject matter can take place. That harmonized European legal framework contributes to the proper functioning of the internal market in the EU, and stimulates innovation, creativity, investment and production of new content, also in the digital environment, in order to avoid the fragmentation of the internal market in the EU. The protection provided by that European legal framework—the "*Acquis Communautaire*"—also contributes to the Union's objective of respecting and promoting cultural diversity while at the same time bringing European common cultural heritage to the fore.

[2]*IFLA Statement on Text and Data Mining*, (2013), available at URL: https://www.ifla.org/publications/node/8225 (last check, July 1, 2019), according to which IFLA does not support licensing as an appropriate solution for TDM. If a researcher or research institution, or another user accessing information through their library, has lawfully acquired digital content, including databases, the right to read this content should encompass the right to mine. Further, the sheer volume and diversity of information that can be utilised for text and data mining, which extends far beyond already licensed research data bases, and which are not viewed in silos, makes a licence-driven solution close to impossible.

[3]*IFLA Statement on Text and Data Mining*, (2013), ibid., according to which IFLA maintains that legal certainty for text and data mining (TDM) can only be achieved by (statutory) exceptions… Copyright and database laws can affect the ability of libraries to fulfil their mandates and deliver information services for the benefit of their patrons, and can impede the use of materials by library users in ways that would benefit communities—for scholarship, research, improvements in health and science, creativity and social inclusion. The text, documents or databases that are mined may well be subject to copyright, related rights and/or database rights. The extraction and copying of content one already has legal access to, and its transformation into a machine-readable format, can touch on the rights holder's exclusive reproduction right. In addition, technical protection measures attached to databases that prevent reproduction are subject to legal protection. The technical act of copying involved in the process of TDM falls by accident, not intention, within the complexity of copyright laws… Researchers must be able to share the results of text and data mining, as long as these results are not substitutable for the original copyright work—irrespective of copyright law, database law or contractual terms to the contrary. Without this right, legal uncertainty may prevent important research and data driven innovation putting researchers, institutions and innovators at risk.

For the EU legislator, TDM is a means to achieve the goal of Digital Single Market: the free movement of goods, persons, services and capital, where individuals and businesses can seamlessly access and exercise online activities under conditions of fair competition, and a high level of consumer and personal data protection, irrespective of their nationality or place of residence.

The EU DSM Strategy (European Commission, COM (2015) 192 final, 2015) considers three pillars in its foundation:

1) Better access for consumers and businesses to online goods and services across Europe. This requires the rapid removal of key differences between the online and offline worlds to break down barriers to cross-border online activity.

2) Creating the right conditions for digital networks and services to flourish. This requires high-speed, secure and trustworthy infrastructures and content services, supported by the right regulatory conditions for innovation, investment, fair competition and a level playing field.

3) Maximizing the growth potential of the European Digital Economy. This requires investment in ICT infrastructures and technologies such as Cloud computing and Big Data, and research and innovation to boost industrial competitiveness as well as better public services, inclusiveness and skills.

Regarding the achievement of the first pillar, i.e. better access for consumers and businesses to online goods and services across Europe, there's a requirement for a more harmonized copyright regime which provides incentives to create and invest while allowing transmission and consumption of content across borders, building on Europe's rich cultural diversity. To this end, the European Commission has been working on proposed solutions that include:

a) Portability of legally acquired content,

b) Cross-border access to legally purchased online services while respecting the value of rights in the audiovisual sector,

c) Greater legal certainty for the cross-border use of content for specific purposes (e.g. research, education, text and data mining, etc.) through harmonized exceptions,

d) Clarification of the rules on the activities of intermediaries in relation to copyright-protected content and

e) Modernization of enforcement of intellectual property rights, focusing on commercial-scale infringements (the "follow the money" approach) as well as its cross-border applicability.

The TDM issue, as well as other copyright provisions included in the new Directive 2019/790/EU on Copyright in the DSM, pertains to the harmonization of exceptions and limitations in copyright law of EU Member States, the creation of legal certainty for cross-border use of content for the purpose of scientific research.

The EU legislator has considered-at least for the time being-recommendations made by various scholars upon the TDM and how it should be regulated in the proposed Directive on Copyright in the DSM, also in the realm of libraries. The

suggestion was that it is best to have a mandatory exception for TDM which would be inspired from, and contain partly the same conditions as the scientific research exception, but which would have its own characteristics prevailed. Article 3 in the text of the new Directive 2019/790/EU is titled "*Text and data mining for the purpose of scientific research*";[4] article 4 is titled "*Exception or limitation for text and data mining*."[5] The mandatory character of the provisions of art.3 and art.4 in the text of the new Directive on Copyright in the DSM can normally be decomposed into three elements, i.e.: (Hargreaves et al., 2014: p. 57).

a) They have to be implemented across all EU Member States in order to ensure effective harmonization of the law;

b) They must not be subject to contractual overrides regarding TDM implemented for scientific purpose; and

c) They must not be subject to lock-up behind technological protection measures.

Even when the owner (or holder) of the data cannot exercise copyright or database rights, contractual restrictions or technical protection measures may render TDM more burdensome or even impossible (Hargreaves et al., 2014: p. 57). For this reason, the wording of the TDM exception in the new Directive on Copyright in the DSM as it was voted on March 26, 2019 by the EU Parliament rules that:

a) Art.3 (1) and art.4 (1): Member States "*shall provide for an exception*"…

[4]Directive 2019/790/EU. The provision of article 3 in the text of Directive 2019/790/EU on copyright in the DSM has as follows:

1) *Member States shall provide for an exception to the rights provided for in Article 5*(*a*) *and Article 7*(*1*) *of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, and Article 15*(*1*) *of this Directive for reproductions and extractions made by research organizations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access.*

2) *Copies of works or other subject matter made in compliance with paragraph 1 shall be stored with an appropriate level of security and may be retained for the purposes of scientific research, including for the verification of research results.*

3) *Rightholders shall be allowed to apply measures to ensure the security and integrity of the networks and databases where the works or other subject matter are hosted. Such measures shall not go beyond what is necessary to achieve that objective.*

4) *Member States shall encourage rightholders, research organizations and cultural heritage institutions to define commonly agreed best practices concerning the application of the obligation and of the measures referred to in paragraphs 2 and 3 respectively.*

[5]Directive 2019/790/EU. The provision of article 4 in the text of Directive 2019/790/EU on copyright in the DSM has as follows:

1) *Member States shall provide for an exception or limitation to the rights provided for in Article 5*(*a*) *and Article 7*(*1*) *of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, Article 4*(*1*)(*a*) *and* (*b*) *of Directive 2009/24/EC and Article 15*(*1*) *of this Directive for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining.*

2) *Reproductions and extractions made pursuant to paragraph 1 may be retained for as long as is necessary for the purposes of text and data mining.*

3) *The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.*

4) *This Article shall not affect the application of Article 3 of this Directive.*

The wording is not "*may provide*" but "*shall provide*" which indicates the mandatory character of the proposed provision.

b) Art.3 (3): "*Rightholders shall be allowed to apply measures to ensure the security and integrity of the networks and databases where the works or other subject-matter are hosted. Such measures shall not go beyond what is necessary to achieve that objective*."

These measures include technological protection measures such as DRM. Thus, technical protection measures may not render TDM burdensome or even impossible.

c) Art.3 (4): "*Member States shall encourage rightholders, research organizations and cultural heritage institutions to define commonly agreed best practices concerning the application of the obligation and of the measures referred to in paragraphs* 2 *and* 3 *respectively*."

The TDM exception is set as an obligation, and EU Member States must encourage rightholders, research organizations and cultural heritage institutions to define best practices concerning the application of the obligation as well as of the measures referred to:

- in Art.3 paragraph 2, i.e. the storage of copies of works or other subject matter which have been harvested from the web with an appropriate level of security and the retain of such stored works or other subject matter for the purposes of scientific research including the verification of research results, and
- in Art.3 paragraph 3, i.e. the application of Technical Protection Measures (TPMs) to ensure the security and integrity of networks and databases where works are hosted, but without going beyond what is necessary to achieve the objective of the mandatory TDM.

d) Art.4 (3): "*The exception or limitation provided for in paragraph* 1 *shall not apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online*."

Essentially, Art.4 (3) sets an opt-out option from the mandatory exception of TDM for any other purpose except for scientific research through the application of means such as machine-readable means in the case of content made available publicly online or contractual agreements. Unless there's an explicit expression of the rightholders of works or database that they do not allow TDM for any other purpose except for scientific research, the TDM exception or limitation to copyright is applicable. To this end, art.7 (1) of the new Directive on Copyright in the DSM rules that "*Any contractual provision contrary to the exceptions provided for in Articles* 3, 5 *and* 6 *shall be unenforceable*." The provision of art.7 (1) does not make reference to art.4 of this Directive, thus art.4 (3), free from the restriction of art.7, allows for an opt-out from the mandatory nature of the exception of TDM.

e) Art.4 (4): "*This article shall not affect the application of Article* 3 *of this Directive*."

This means that no contract may override TDM in the case of TDM implemented for scientific purposes.

There were many suggestions on how to encourage TDM for research purposes without fear of infringing intellectual property rights. The goal for such an encouragement through legislative action could be achieved in a number of ways, namely: (Hargreaves et al., 2014: p. 57)

1) Through an adjustment of licensing practices

2) Through a revised, normative interpretation of the reproduction right in copyright

3) Through the introduction of a new mandatory exception in copyright and database laws, or

4) Through the adoption of an "*open norm*" designed to guide the courts to take a more flexible view of what users are permitted to do.

The EU legislator opted for the choice of introducing a mandatory exception for TDM covering uses pursuing scientific research purposes and limited to certain beneficiaries, i.e. research organizations and cultural heritage institutions (art.3), but also allowing uses for other purposes either non-commercial or commercial and not limited to certain beneficiaries (art.4), and also of ensuring that TDM regulation cannot be over-ridden through the enforcement of restrictive contractual clauses—in the case of TDM implemented for scientific purposes (art.3)—or technological protection measures.

The point of contention between the introduction of a new mandatory exception and the facilitation of TDM in consideration of the existing exception for scientific research has found its solution in the introduction of a new mandatory exception. The license option and the encouragement of TDM through licensing was deemed to be inefficient and inadequate to create legal certainty among Member States regarding TDM for scientific research (EC, SWD (2016) 301 final PART 2/3, 2016: pp. 51-52)[6]. The extent to which TDM in Europe is facilitated by any existing exceptions to either EU copyright or database law appeared unclear. The application of a copyright and database exception relating to teaching or scientific research is optional and has not been implemented at all in some Member States. This has contributed to uncertainty in the European scientific research community (EC, SWD (2016) 301 final PART 1/3, 2016: pp. 104-105)[7].

[6]Researchers have generally considered that licenses-based solutions would not be able to fully solve the problems of legal uncertainty they face as regards the use of TDM techniques. This was also confirmed in these stakeholders' replies to a 2013-2014 public consultation (institutional users such as libraries and universities generally considered licenses an inadequate source of transaction costs for TDM and indicated that a legislative change is needed to introduce a mandatory exception for text and data mining in EU copyright law).

[7]Researchers are generally convinced of the potential of TDM but they put forward legal uncertainty, caused by the current copyright rules, as one of the reasons for the slow development of TDM in the EU (in addition to aspects unrelated to copyright, such as lack of awareness and skills, infrastructural challenges, etc.). A considerable level of legal uncertainty exists among researchers regarding TDM and copyright law. Research organizations and researchers do not always know whether TDM is copyright-relevant at all, whether it may be covered by an exception or whether a specific rightholders' authorization is required.

Moreover, it was considered that unless a TDM mandatory exception applicable horizontally for all Member States were passed, the possibility of enacting different TDM legislations in Member States is possible, and as a consequence the fragmentation of the Single Market is more than likely to increase over time as a result of Member States adopting TDM exceptions at national level which could be based on different conditions, which is likely to happen in the absence of intervention at EU level (EC, SWD (2016) 301 final PART 1/3, 2016: p. 106).

The introduction of the exception or limitation regarding TDM in the text of the new Directive on Copyright in the DSM is mandatory. According to Recital 5 of the Directive, the existing exceptions and limitations in European Union law should continue to apply, including to text and data mining, education, and preservation activities, as long as they do not limit the scope of the mandatory exceptions or limitations provided for in the proposed new Directive on Copyright in the DSM, which need to be implemented by Member States in their national law. Directive 96/9/EC—the Database Directive—and Directive 2001/29/EC—the so called InfoSoc Directive or Directive on Copyright in the Information Society—are, therefore, amended by Directive 2019/790/EU (Directive 2019/790/EU) (Directive 2019/790/EU, Recital 5).

TDM is treated as a means for research and innovation which allows uses of copyrighted works as well as of non-copyrighted material which are not clearly covered by the existing "*Acquis Communautaire*" on exceptions and limitations to copyright. Through this reference on research and innovation—the text of Recital 5 includes education, and preservation of cultural heritage, too—the EU legislator makes a nuanced reference to art.5 (3) (a) of Directive 2001/29/EC which caters for non-mandatory exceptions or limitations to the reproduction right of art.2 of the InfoSoc Directive as well as to the right of communication to the public of works and the right of making available to the public of other copyrighted subject-matter of art.3 of the InfoSoc Directive. According to art.5 (3) (a) of the InfoSoc Directive Member States may provide for exceptions or limitations to the rights provided for in Articles 2 and 3 in—among other cases—case of use for the sole purpose of illustration for teaching or scientific research, as long as the source, including the author's name, is indicated, unless this turns out to be impossible and to the extent justified by the non-commercial purpose to be achieved. Not all EU Members have adopted the provision of art.5 (3) (a) of the InfoSoc Directive, and among those EU Members which have implemented this provision in their national law, there are significant differences in the texts and accorded protection of national laws.

2. Removing Legal Uncertainty for TDM

Most exceptions or limitations to copyright are non-mandatory. They are, therefore, not implemented in the same way in the EU Members' legal systems, and are not fully adapted to the use of technologies such as TDM technologies used in scientific research. Legal uncertainty follows, on TDM as well as other

exceptions or limitations too, which the new Directive on Copyright in the DSM aims to alleviate. The non-mandatory nature of most of InfoSoc Directive's list of exceptions and limitations to copyright is a cause of failure in the process of harmonization of copyright rules applicable in all Member States of the EU (Geiger et al., 2018: pp. 14-15)[8]. The non-harmonized EU legal framework for exceptions and limitations, especially those pertaining to scientific research and teaching, which have not been implemented nationally by EU Member States in the same way due to their non-mandatory nature, cause significant difficulties in leveraging on the existing legal framework for Copyright for covering the TDM activity.

In Greece, for example, there is no provision in Copyright law 2121/1993 for an exception or limitation of copyright for scientific research; article 21 of law 2121/1993—the Greek Copyright Law—does not address an exception or limitation for scientific research, but rather it addresses the teaching limitation solely (Law of Greece 2121/1993)[9]. Art.81 of law 3057/2002 through which law 2121/1993 was amended in consideration of the provisions of the InfoSoc Directive did not include the exception or limitation of copyright for scientific research. Libraries, therefore, like other entities, cannot benefit from this exception.

But even in cases of EU Members' legal systems which cater for the scientific research exception or limitation, which seems to remain at the founding causes for setting a new mandatory exception of the copyright for TDM, it is quite common that the provision in national law benefiting scientific research is not fully adapted to the use of technologies such as TDM for scientific purposes, and certainly is not adapted to the use of technologies for TDM for any other purpose aside from scientific research. Moreover, where researchers, via a library or not, have lawful access to content, for example through subscriptions to publications or open access licenses, the terms of the licenses could exclude text and data mining. As research is increasingly carried out with the assistance of digital technology, there was a risk that the European Union's competitive position as a research area would suffer, unless steps were taken to address the legal uncertainty concerning text and data mining (Directive 2019/790/EU, Recital 10). These steps were taken in the form of the new Directive on Copyright in the DSM, which among other issues, rules upon TDM in its provisions of art.3 and art.4.

[8]Geiger et al. (2018), pp. 14-15 references in footnotes 65, 68, 70. A unified and mandatory approach is especially crucial in the digital environment as the internet involves uses that, most of the time, affect several copyright legislations, leading to a major insecurity regarding what is allowed. See also Stamatoudi, 2016: pp. 251-283.

[9]According to art. 21 of law 2121/1993, which is titled "Reproduction for Teaching Purposes" *It shall be permissible, without the consent of the author and without payment, to reproduce articles lawfully published in a newspaper or periodical, short extracts of a work or parts of a short work or a lawfully published work of fine art work exclusively for teaching or examination purposes at an educational establishment, in such measure as is compatible with the aforementioned purpose, provided that the reproduction is effected in accordance with fair practice and does not conflict with the normal exploitation. The reproduction must be accompanied by an indication of the source and of the names of the author and the publisher, provided that these names appear on the source.*

Though the existing exceptions in EU Copyright law on research, teaching and education aim at achieving public policy objectives, objectives among EU Member States remain different. As new types of uses have emerged, it remains uncertain whether these exceptions are still adapted to achieve a fair balance between the rights and interests of authors and other rightholders on the one hand, and of users on the other. Besides, these exceptions remain national and legal certainty around cross-border uses is not guaranteed. As a consequence, cross-border collaborations of researchers are hindered by the lack of sameness in understanding and applying the research exception or limitation to copyright which could cover under certain conditions the TDM, too; this affects directly TDM activities since researchers are unaware—or face high transaction costs for clearance—of whether TDM would be lawful across all EU jurisdictions involved in the research collaboration (Geiger et al., 2018: pp. 12-13). The situation of legal uncertainty is further affected by combinations of contractual and technical measures which are frequently used to create insurmountable hurdles for researchers engaging in TDM projects. Actually, contractual and technological barriers are also frequently used to prevent TDM activities on materials not protected by copyright or on public domain subject matter (Geiger et al., 2018: p. 13), and the Court of Justice has ruled that the use of contractual and technological means on non-protected by copyright or the *sui generis* right databases is not illegal (Case C-30/14, 2015)[10].

The new Directive on Copyright in the DSM aims to tame the legal uncertainty concerning text and data mining by providing for a mandatory exception for universities and other research organizations, as well as for cultural heritage institutions, and libraries, to the exclusive right of reproduction and to the right to prevent extraction from a database. In line with the existing European Union research policy, which encourages universities and research institutions to collaborate with the private sector, the EU legislator aims to encourage research organizations throughout the EU to benefit from the TDM mandatory exception or limitation from copyright in the provisions of art.3 and art.4 of the new Directive on Copyright in the DSM when their research activities are carried out in the framework of public-private partnerships and/or in cross-border collaborations (Directive 2019/790/EU, Recital 11). The intention of TDM provisions in the new Directive is to alleviate legal uncertainty upon applicable copyright law and to enable research organizations and cultural heritage institutions to continue to be the beneficiaries of the TDM exception, and rely on their private partners for carrying out TDM, including by using their technological tools (Directive 2019/790/EU, Recital 11).

[10]Case C-30/14, (2015), Ryanair Ltd v PR Aviation BV; in this case the CJ ruled that Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases must be interpreted as meaning that it is not applicable to a database which is not protected either by copyright or by the sui generis right under that directive, so that Articles 6(1), 8 and 15 of that directive do not preclude the author of such a database from laying down contractual limitations on its use by third parties, without prejudice to the applicable national law.

3. What TDM Is

In the text that was adopted on March 26, 2019, TDM is understood as the automated analytical technique aimed at analyzing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations (Directive 2019/790/EU, Recital 11)[11]. In addition to texts, the term "*text*" is broad enough to include fixed images, sound recordings, and audio-visual works. TDM is meant to be the automated computational analysis of information in digital form, such as text, sounds, images or data that is enabled through the use of new computational technologies (Directive 2019/790/EU, Recital 8)[12]. In a broad sense, TDM is called any activity where computer technology is used to index, analyze, evaluate and interpret mass quantities of content and data (Caspers et al., 2016: p. 9). TDM makes the processing of large amounts of information with a view to gaining new knowledge and discovering new trends possible. TDM is also an inherent part of Artificial Intelligence and Machine Learning research. Machine Learning refers to a cluster of statistical and programming techniques that give computers the ability to learn from exposure to data without being explicitly programmed (Sag, 2019: p. 7). TDM technologies are prevalent across the digital economy, however there is widespread acknowledgment that TDM can in particular benefit the research community and, in so doing, support innovation (Directive 2019/790/EU, Recital 8).

Such technologies benefit universities and other research organizations, as well as cultural heritage institutions since they could also carry out research in the context of their main activities. However, in the European Union, such organizations and institutions are confronted with legal uncertainty as to the extent to which they can perform TDM of content. In most instances, TDM can involve acts protected by copyright, by the sui generis database right or by both, in particular, the reproduction of works or other subject matter, the extraction of contents from a database or both which occur for example when the data is normalized in the process of TDM. Where no exception or limitation applies, an authorization to undertake such acts is required from rightholders (Directive 2019/790/EU, Recital 8). TDM requires making a copy of the materials used for text and data mining to be read by a computer.

However, computer reading is not the same as human reading. Computers read by applying mathematical functions to the text or other subject matter inserted to them for TDM in the process of generating abstract statistics. In the TDM process the computer used does not comprehend or enjoy the copyrighted works inserted to it in the way humans do by reading. The computer simply processes the materials to produce metadata. This use of the copyrighted works does not threaten the rights of authors as they have been traditionally unders-

[11]Definition of TDM in the art.2(2) of the voted text of the Directive on Copyright in the DSM.

[12]See, also, European Commission, (2016), ibid., according to which *Text and Data Mining* (*TDM*) *is a term commonly used to describe the automated processing* ("*machine reading*") *of large volumes of text and data to uncover new knowledge or insights.*

tood despite the fact that there is reproduction of works in the process of TDM (Sag, 2019: pp. 8-9). The statistical information or the knowledge produced as an output of the TDM process is not the outcome of human appreciation of the expressive qualities of copyrighted works which are reproduced so as to become legible by the computer used in the TDM process.

This reasoning drives to the reasonable conclusion that use of works in the TDM process equals to "*non-expressive use*" (Sag, 2009)—sometimes referred to as "*non-consumptive use*", i.e. is an act of reproduction of copyrighted work that is not intended to enable human enjoyment, appreciation, or comprehension of the copyrighted expression as such (Sag, 2019: p. 9). DM does not communicate the expression of the copyrighted works submitted to it, but rather it generates valuable information about the works submitted to it that is different from what is expressed by the works submitted to it (Sag, 2019: pp. 21-22). TDM and other non-expressive uses do not communicate original expression to the public (i.e., to any human reading audience for the purpose of being read, understood, or appreciated). As such, even though these uses involve technical acts of copying, they do not conflict with the copyright owner's exclusive rights (Sag, 2019: p. 10).

As a consequence of the above reasoning in the US law, in 2011 the Hargreaves Review recommended that the UK implement an exception for "*non-consumptive use*". This use was defined as use of a work enabled by technology which does not trade on the underlying creative and expressive purpose of the work. The idea is to encompass the uses of copyright works, where copying is really only carried out as part of the way technology works. For instance, in data mining or search engine indexing, copies need to be created for the computer to analyze; the technology provides a substitute for reading all the documents. These new uses happen to fall within the scope of copyright regulation is essentially a side effect of how copyright has been defined, rather than being directly relevant to what copyright is supposed to protect (Hargreaves, 2011: p. 47).

In the US, courts have found that reproducing copyrighted works as one step in the process of knowledge discovery through text data mining is transformative, and thus ultimately a fair use of those works that fits in the first fair use factor of the US Copyright Act (Authors Guild v. Hathi Trust, 755 F.3d 87, 97-98, 2d Cir., 2014)[13]. Since the US Supreme Court's 1994 decision in *Campbell v. Acuff-Rose* (Campbell v. Acuff-Rose Music (92-1292), 510 U.S. 569, 1994), the central focus of the first factor in the fair-use doctrine of the US Copyright law, Section 107 (US Copyright Act, 1976)[14] has been on whether the defendant's use

[13]Authors Guild v. HathiTrust, 755 F.3d 87, 97-98 (2d Cir.2014); ibid., at 101 (concluding library digitization for the purpose of permitting full-text searches fair use).

[14]Section 107 of the US Copyright Act, a.k.a. the fair-use doctrine includes the following four factors: 1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; 2) the nature of the copyrighted work; 3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and 4) the effect of the use upon the potential market for or value of the copyrighted work.

is "*transformative*" (Sag, 2019: p. 20). In the evaluation of "*the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes*" the "*transformative*" use is apt to the real question whether the defendant's use of the copyrighted work "*adds something new, with a further purpose or different character, altering the first with new expression, meaning, or message*" (Sag, 2019: p. 20)[15]. Classic transformative uses are generally fair uses because, in spite of communicating some of the author's original expression, they do not substitute for it. Parody, commentary, criticism, illustration, and explanation all may include large portions of the author's original expression, but these transformative uses do not usually pose any risk of expressive substitution of the author's copyrighted work (Sag, 2019: p. 25).

In 2014 in *Authors Guild Inc., v HathiTrust*, (Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 97, 2d Cir., 2014) the US Court of Appeals for the Second Circuit concluded that "*the creation of a full-text searchable database is a quintessentially transformative use*", and concluded that "… *the result of a word search is different in purpose, character, expression, meaning, and message from the page* (*and the book*) *from which it is drawn. Indeed, we can discern little or no resemblance between the original text and the results of the* [*HathiTrust Digital Library*] *full-text search*" (Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 97, 2d Cir., 2014; Sag, 2019: p. 23). In 2015, in the *Authors Guild Inc., v. Google Inc.*, case the US Court of Appeals for the Second Circuit found that Google's copying of the entire text of library books to create a search index was "*highly transformative*" (Authors Guild v. Google, Inc., 804 F.3d 202, 2d Cir., 2015: pp. 216-217). The court explained that "*as with HathiTrust* (*and iParadigms*), *the purpose of Google's copying of the original copyrighted books is to make available significant information about those books, permitting a searcher to identify those that contain a word or term of interest, as well as those that do not include reference to it. In addition, through the ngrams tool, Google allows readers to learn the frequency of usage of selected words in the aggregate corpus of published books in different historical periods. We have no doubt that the purpose of this copying is the sort of transformative purpose described in Campbell*" (Authors Guild v. Google, Inc., 804 F.3d 202, 2d Cir., 2015: p. 24).

Therefore, TDM has been found by US Courts to be "*transformativeuse*" of copyrighted works that fits in the first factor of the US fair-use doctrine. The concept of "*transformative use*" fits in the concept of "*non-expressive use*" the latter being considered as a subset of the former (Sag, 2019: pp. 21, 24 et seq.). TDM is "*transformative*" in the sense that it does not merely supersede the objects of the copyrighted works but instead it adds something new, with a further purpose or different character, it has a "*transformative*" outcome (Sag, 2019: p. 21)[16]. Also, TDM is "*non-expressive use*" in the sense that it does involve the use

[15]Sag, M., (2019), ibid., citation in *Campbell v. Acuff-Rose Music* (92-1292), 510 U.S. 569 (1994) therein.

[16]Sag, M., (2019), ibid., citation in *Campbell v. Acuff-Rose Music* (92-1292), 510 U.S. 569 (1994) therein.

of copyrighted work for a purpose that does not ultimately (or substantially) convey the original expression encoded within the copyrighted work and thus it does not infringe copyright in consideration of the US Copyright legislation (Sag, 2019: pp. 24-25). "*Non-expressive use*" of copyrighted work generates information about a work, that information may be useful, it may be valuable, it may even affect the demand for the underlying work, but metadata about a work does not in any way fulfill the public's demand for the author's original expression (Sag, 2019: p. 25). By definition, a non-expressive use of a copyrighted work does not usurp the copyright owner's communication of her original expression to the public because the expression is not communicated (Sag, 2019: p. 25). Thus, in the US legal theory and jurisprudence it has been firmly associated with the fair-use doctrine that certain "*non-consumptive uses*" or "*non-expressive uses*" ought not to be protected by copyright, and this is based on the fundamental distinction in copyright law: that between ideas and expression (Sag, 2019: p. 11 et seq.)[17].

In the US legal theory and practice, allowing text mining and other similar non-expressive uses of copyrighted works without authorization in consideration of fair-use (US Copyright Act, Section 107 (1)) is entirely consistent with the fundamental structure of copyright law because, at its heart, copyright law is concerned with the communication of an author's original expression to the public. The fair-use doctrine permits copying (or distribution, display, or performance) without permission in certain circumstances, depending on the purpose, proportionality, and effect of that copying. Copying that amounts to fair use is not merely excused, it is not infringement and thus requires no further license or excuse. The fair-use doctrine as is provisioned in US Copyright Act, Section 107 describes four statutory factors which are interrelated and must be "*explored, and the results weighed together, in light of the purposes of copyright*" (Sag, 2019: pp. 19-20)[18].

4. Transformative/Non-Expressive Use in the EU and TDM

Thus, for the US legal framework access and the making of copy of a copyrighted work in the TDM process is covered by the "*fair-use*" doctrine of US Copyright Act (US Copyright Act, Section 107 (1)). The "*fair-use*" doctrine

[17]Sag mentions a couple of seminal per the idea-expression dichotomy in Copyright law cases before the US Supreme Court, such as the *Harper & Row* case (*Harper & Row, Publishers, Inc. v. Nation Enterprises*, 471 U.S. 539, 556 (1985) available at URL: https://supreme.justia.com/cases/federal/us/471/539/ [last check, July 1, 2019]) which strikes a definitional balance between the First Amendment and the Copyright Act by permitting free communication of facts while still protecting an author's expression. Also, the *Eldred v. Ashcroft* case (*Eldred v Ashcroft* (01-618) 537 U.S. 186 (2003) 239 F.3d 372, available at URL: https://www.law.cornell.edu/supct/html/01-618.ZO.html [last check, July 1, 2019]) in which the Court found that the idea-expression distinction is one of copyright's "*built-in First Amendment accommodations*" and that as a result of the idea-expression distinction, "*every idea, theory, and fact in a copyrighted work becomes instantly available for public exploitation at the moment of publication*".

[18]See, also, *Campbell v. Acuff-Rose Music* (92-1292), 510 U.S. 569 (1994) case to which Sag, M. (2019) cites.

permits copying (or distribution, display, or performance) without permission in certain circumstances, depending on the purpose, proportionality, and effect of that copying. Copying that amounts to fair use is not merely excused, it is not infringement and thus requires no further license or excuse. The "*fair-use*" doctrine as is provisioned in US Copyright Act, Section 107 describes four statutory factors which are interrelated and must be "*explored*, *and the results weighed together*, *in light of the purposes of copyright*."

Arguments for the introduction of an open norm similar to the "*fair-use*" doctrine of the US Copyright Act have been voiced in Europe, both before and during the drafting of the new Directive on Copyright in the DSM. For example, the 2011 Hargreaves Review regarding the amendment of the UK Copyright law recommended that the UK should implement an exception for "*non-consumptive use*", which was defined as use of a work enabled by technology which does not trade on the underlying creative and expressive purpose of the work (Hargreaves, 2011: p. 44 et seq.)[19]. According to the 2011 Hargreaves Review the idea is to encompass the uses of copyrighted works where copying is really only carried out as part of the way technology works. For instance, in data mining or search engine indexing, copies need to be created for the computer to analyze; the technology provides a substitute for reading all the documents. The fact that these new uses of works through the application of technology happen to fall within the scope of copyright regulation is essentially a side effect of how copyright has been defined.

There have been similar arguments posed by other scholars too claiming that it is the facts dispersed throughout the content and the relationship between the facts which are of interest to scientific researchers, neither of which are in themselves protected by copyright, and therefore the introduction of an open norm in the EU Copyright system would have become a solid legal basis for covering TDM as well as other not-yet discovered technology-enhanced and technology-neutral uses of works/data that though involve a reproduction of the protected material, they are "*non-consumptive uses*" or "*non-expressive uses*" which ought not to be protected by copyright. All of them agree that an inclusion of flexibility—which the open norm entails—should not be achieved by legalizing questionable interpretative approaches developed by national courts, but rather it should find expression through a new flexible norm that should be included in the EU legal framework for Copyright and stand in systematic contrast to the

[19]The 2011 Hargreaves Review considered whether the more comprehensive American approach to copyright exceptions, based upon the so-called "*Fair Use*" defense, would be beneficial in the UK. In order to make progress at the necessary rate, the 2011 Hargreaves Review concluded that the UK needs to adopt a twin track approach: pursuing urgently specific exceptions where these are feasible within the current EU framework, and, at the same time, exploring with our EU partners a new mechanism in copyright law to create a built-in adaptability to future technologies which, by definition, cannot be foreseen in precise detail by today's policy makers. This latter change will need to be made at EU level, as it does not fall within the current exceptions permitted under EU law. The 2011 Hargreaves Review recommend that the UK Government should press at EU level for the introduction of an exception allowing uses of a work enabled by technology which do not directly trade on the underlying creative and expressive purpose of the work (this has been referred to as "*non-consumptive*" use).

closed list of limitations and exceptions to Copyright described in article 5 of the InfoSoc Directive.

Flexible interpretations of the listed exceptions and limitations of article 5 of the InfoSoc Directive either by the Court of Justice of the EU or national courts could not suffice in replacement of an open norm in the EU Copyright legal system. This option of flexible interpretation by the court neither can promise flexibility in the application of EU Copyright law for the future nor does it provide positive legal certainty for uses of copyrighted works that are based on new technologies.

Therefore, the argument that the solution for flexibility in EU Copyright law could be satisfied through an open norm in the existing catalog of the exceptions and limitations of the InfoSoc Directive has been voiced in Europe, thus the need to open the existing closed list of exceptions and limitations to copyright so as to accommodate within it a European open norm doctrine similar to the American "*fair use*" doctrine. This solution stresses the need for reforming article 5 of the InfoSoc Directive especially with regard to limitations and exceptions in a way that respects the imperative of legal certainty and contributes to the "Digital Agenda" and the "Digital Single Market" strategic orientations of EU Copyright law.

However, this argument for the introduction in the EU Copyright law of an open norm doctrine similar to the American "*fair use*" doctrine did not prevail in the drafting of the new Copyright Directive in the DSM. The US legal reasoning upon the meaning of reproduction in the case of copying that is "*non-expressive use*" or "*non-consumptive use*" and "*transformative use*" was not accepted as such by the drafters of Directive 2019/790/EU. The EU law supports the idea/expression dichotomy; international treaties and conventions on Copyright in the EU such as the TRIPs Agreement (TRIPs Agreement, 1995)[20] and the WIPO Copyright Treaty (WIPO Copyright Treaty, 1996)[21] sustain that the European Copyright law considers the reproduction right to be at the core of author's copyright and vests it with a broad meaning that includes technical rather than functional scope.

In the EU Copyright law, the notion of reproduction is accepted at its broadest meaning as is clearly stated in art.2 of the InfoSoc Directive, which includes "…*direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part*". Recital 21 of the InfoSoc Directive provides a broad definition of the acts of reproduction is needed to ensure legal certainty within the internal market in the EU. And it has been confirmed as such—i.e. broad definition of the notion of "*reproduction*"—by the Court of Justice in the *Infopaq* case (Case C-5/08 Infopaq International A/S v Danske DagbladesForening [2009] Eu ZW 655, 2009). In the EU Copyright law, the meaning of re-

[20]Art.9(2) provides that "*Copyright protection shall extend to expressions and not to ideas, procedures, methods of operation or mathematical concepts as such*".

[21]Art.2 provides that "*Copyright protection extends to expressions and not to ideas, procedures, methods of operation or mathematical concepts as such*".

production is to determined technically rather than functionally (Walter & Lewinski, 2010: p. 968). Thus, copying of works in the framework of the TDM process in the EU Copyright law falls within the legal meaning of reproduction which is an exclusive right of the author of a work.

Under Recital 9 of the new Directive on Copyright in the DSM, TDM can also be carried out in relation to mere facts or data not protected by copyright, and in such instances no authorization is required under copyright law. There can also be instances of TDM that do not involve acts of reproduction or where the reproductions made fall under the mandatory exception for temporary acts of reproduction provided for in art.5 (1) of Directive 2001/29/EC (Directive 2001/29/EC, 2001)[22], which should continue to apply to TDM techniques that do not involve the making of copies beyond the scope of that exception.

The statutory exception of TDM pertains but is not limited to activities which are confined to acts of "*automated processing of large amounts of structured digital textual content*, *for purposes of information retrieval*, *extraction*, *interpretation*, *and analysis*" which are undertaken for scientific research purposes. In her benchmark 2011 report, Eefke Smit refers to TDM as "*automated tools*, *techniques or technology to process large volumes of digital content that is often not well structured—to identify and select relevant information; to extract information from the content*, *to identify relationships within/between/across documents and incidents or events for meta-analysis*" (Smit & Van der Graaf, 2011). Aside from the term "*text and data mining*" which is usually referred with the TDM initials, the notions of text mining, text data mining, content mining, and computational text analysis are often used interchangeably with the "*text and data analysis*" or the "*text and data mining*" with the aim to describe a TDM inquiry (Bergman et al., 2013) or an analytical TDM approach (Reilly, 2012: pp. 75-76).

TDM works in the following manner (Geiger et al, 2018: pp. 5-6):

1) It identifies input materials to be analyzed, such as works, or data individually collected or organized in a pre-existing database;

2) It copies substantial quantities of materials—which encompasses

a) pre-processing materials by turning them into a machine-readable format compatible with the technology to be deployed for the TDM so that structured data can be extracted. Preprocessing typically encompasses the following tasks (Tsolakidou, 2018: pp. 33-34):

- *Tokenization*: this is typically the first step in a natural language processing solution and it refers to splitting the text into meaningful character sequences/self-contained semantic units, e.g. words or sentences. A naïve toke-

[22]According to art.5(1) of Directive 2001/29/EC, temporary acts of reproduction referred to in Article 2 [of this Directive], which are transient or incidental [and] an integral and essential part of a technological process and whose sole purpose is to enable: (a) a transmission in a network between third parties by an intermediary, or (b) a lawful use of a work or other subject-matter to be made, and which have no independent economic significance, shall be exempted from the reproduction right provided for in Article 2 [of this Directive].

nization solution involves removing punctuation and splitting the text by blank spaces.

- *Normalization*: this involves removing morphological variations from words such as capitalization, plural number or tenses, in order to grasp similarities between them (e.g., the same word in singular and plural), obviously with a loss of information. Two types of techniques are used, *stemming* and *lemmatization*. In the former, language specific patterns are recognized, using for example the rules for converting words from singular to plural or verb tenses. This technique is simple, fast and applicable for large volumes of text. Lemmatization involves using a dictionary (such as Word Net that is both a dictionary and a thesaurus) to extract the roots of common words. This approach can be more accurate compared to stemming, but it is more resource intensive and dictionaries may be incomplete for certain languages. The two methods can complement each other and they are often used in conjunction.
- *Parsing*: this involves a group of functions that are used after term isolation and document cleanup, i.e., after normalization and parsing, which facilitate working in higher abstraction layers. Typically, parsing includes morphological and syntactical analysis of tokens in order to identify their role within sentences (e.g. noun, verb, adjective or object-verb-subject), which is referred to as *Part-of-Speech* (*POS*) *tagging*.

b) possibly, but not necessarily, uploading the pre-processed materials on a platform, depending on the TDM technique to be deployed;

3) It extracts the data; and

4) It recombines data to identify patterns into the final output.

Once access to content is available or granted, TDM generally implies the reproduction of the text or the data, either temporarily, e.g. by caching the content or permanently, e.g. by creating a database of key elements for facilitating searches (index). There are also TDM technologies which allow for analyzing content without making any copies of the analyzed content, e.g. by website crawling or screen-scrapping. TDM tools involving minimal copying of few words or crawling through data and processing each item separately could be operated without running into potential liability for copyright infringement. This follows from the fact that copyright law does not protect data but only original expressions within copyright protected subject matters. In this respect, the proposal for a new Directive on copyright in the DSM clarifies that "*Text and data mining may also be carried out in relation to mere facts or data which are not protected by copyright and in such instances no authorization would be required*" (Directive 2019/790/EU, Recital 9). The amendment no.271 of May 26, 2019 of the text of this proposed Directive added in Recital 9 the specific mention that works and other subject matter not protected by copyright or the sui generis right can be freely mined. This may happen in relation to mere facts or data that are not protected by copyright, and in such instances no authorization is required under copyright law; and it may also happen in instances of TDM that do not involve acts of reproduction or where the reproductions made may fall under the mandatory ex-

ception for temporary acts of reproduction provided for in art.5 (1) of Directive 2001/29/EC, which should continue to apply to TDM techniques that do not involve the making of copies beyond the scope of that exception (Geiger et al., 2018: p. 6).

Content that is text and data mined may come in different formats, such as machine-readable formats (e.g. XML) or PDFs, which may be more or less easily mined. The data retrieved often needs to be normalized, annotated and aggregated into a corpus to allow for an efficient use of mining software. The normalization and annotation can be done either by the publishers, including as part of a commercial offer (e.g. data in an XML format, provided in a structured way) or by the researchers themselves, which is more the case for researchers in the public interest research organizations, who tend to prefer using their own tools (relying also more on PDFs than commercial users). The normalization and annotation phase of TDM activity involves the preprocessing to standardize materials into machine-readable formats; activity in this phase might trigger infringement of the right of reproduction of works found online (Geiger et al., 2018: p. 6). Likewise, the uploading of the pre-processed material on a platform—which might occur or not depending on whether the TDM technique adopted makes use of a TDM software crawling data to be analyzed directly from the source—might also violate the right of reproduction. The process of analyzing the texts or data is to be distinguished from its result. The output of TDM might consist for example of a summary of the analyzed text and data, visualizations such as graphics or charts, but also of new knowledge, patterns, and combinations of data that may lead to new discoveries and research results. However, the analysis and extraction of the TDM process, i.e. the phase where data is finally extracted—can also infringe upon the right of reproduction depending on the mining software deployed and the character of the extraction (EC, SWD (2016) 301 final PART 1/3, 2016: p. 158).

Regarding TDM activity on databases, TDM might involve the reproduction, translation, adaptation, arrangement, and any other alteration of a database protected by copyright, which means the original selection and arrangement of the database's content (Geiger et al., 2018: p. 6). TDM activity might, also, infringe sui generis database right, in particular the extraction—and to a minor extent the re-utilization—of substantial parts of a database or the repeated extraction of insubstantial parts of a database. In this context, even if extraction does occur without reproduction of the original materials, extraction itself would infringe upon the exclusive sui generis right provided to the database owner (Geiger et al, 2018: p. 7; C-203/02, The British Horseracing Board Ltd. and Others v. William Hill Organization Ltd., 2004). According to the CJ (C-203/02, The British Horseracing Board Ltd and Others v. William Hill Organization Ltd., 2004), the infringement occurs by unauthorized actions for the purpose of reconstituting, through the cumulative effect of acts of extraction, the whole or a substantial part of the contents of a database protected by the sui generis right

and/or of making available to the public, through the cumulative effect of acts of re-utilization, the whole or a substantial part of the contents of such a database, which thus seriously prejudice the investment made by the maker of the database. Article 7 (5) of the Database Directive refers to unauthorized acts of extraction or re-utilization the cumulative effect of which is to reconstitute and/or make available to the public, without the authorization of the maker of the database, the whole or a substantial part of the contents of that database and thereby seriously prejudice the investment by the maker.

5. The Notion of "Lawful Access"

Both the provisions of art.3 and art.4 of the new Directive on Copyright in the DSM require "*lawful access*" to works which may be submitted to TDM (EC, SWD (2016) 302 final, 2016)[23]. The requirement of "*lawful access*" to works for TDM is not new. It is met in the provisions of Section 29A of the UK law which rules for TDM as of 2014 (Law of UK:Copyright, Designs and Patents Act, 1988)[24].

Among the very few EU Members—save for the UK's provision of art.29A as is reported above hereto (as of the composition of this text the UK remains an EU Member on the verge of its BRexit from the EU)—which have already amended their copyright legislation catering for TDM none of them include a "*lawful access*" requirement (Law of UK: Copyright, Designs and Patents Act, 1988)[25].

TDM provision in law has been implemented in France, Germany and Estonia, so far. Regarding French law (Law of France: Law No.2016-1231 for a Digital Republic, 2016)[26] and its provision on TDM, there's no requirement for "*lawful access*".

In the Estonian Copyright Act (1992)[27] (Law of Estonia: Estonian Copyright Act, 1992) while the exception for TDM does not contain "*lawful access*" or sim-

[23]European Commission & SWD 302 Final (2016), on the modernization of EU copyright rules, available at URL: http://ec.europa.eu/transparency/regdoc/rep/10102/2016/EN/SWD-2016-302-F1-EN-MAIN-PART-1.PDF [last check, July 1, 2019] according to which *For TDM, the preferred option is a mandatory exception applicable to research organizations acting in the public interest such as universities or research institutes. The exception would allow them to carry out TDM on content they have lawful access to, for the purposes of scientific research.* The Executive Summary of the Impact Assessment is available at URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52016SC0302&from=EN [last check, July 1, 2019].

[24]The UK legislator amended its Copyright law by S.I. 1992/3233, regulation 7, S.I. 1997/3032, regulation 8 and S.I. 2003/2498, regulation 9. Section 29A that was added to the Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014 came into force on June 1st, 2014.

[25]See Section 29A par.1 for the requirement of "lawful access"; section 29A was added to UK's the Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014 which came into force on June 1st, 2014.

[26]In France, the legislator of Law No.2016-1231 for a Digital Republic (Loi pour uneRépublique-numérique), introduced TDM exceptions both applying to works (art.L.122-5, 10 of the CPI) and databases (art.L.342-3, 5 of the CPI).

[27]The Estonian legislator amended the country's Copyright Act of 1992 and as of 01/01/2017 introduced TMD in paragraph 3 of art.19 titled "*Free use of works for scientific, educational, informational and judicial purposes.*"

ilar requirement, it demands mention of the name of the author of the work, if it appears thereon, the name of the work and the source publication.

Regarding the text in Section 60d of the German Copyright law (Law of Germany: German Act on Copyright and Related Rights, 1965)[28] it does not impose a "*lawful access*" requirement. Also, it does not limit the source materials that can be mined to text and data. With regard to databases, their reproduction is being qualified as constituting "*normal use*" in accordance with section 55a, first sentence.

The notion of "*lawful access*" which is a requirement for TDM in art.3 and art.4 of the new Directive on Copyright in the DSM has been criticized since it could hamper TDM in the sense of de facto subject TDM research to private ordering. According to the European Copyright Society, "the exception can effectively be denied to certain users by a right holder who refuses to grant "*lawful access*" to works or who grants such access on a conditional basis only" (European Copyright Society, 2017: p. 4; Max Planck Institute for innovation and Competition, 2016; Geiger et al, 2018: p. 22.). The deployment of TDM can effectively be denied to certain users by a right holder who refuses to grant "*lawful access*" to works or who grants such access on a conditional basis only. Moreover, this deference to private ordering allows publishers to price TDM into their subscription fees, thus subjecting TDM to lawful access will make TDM research projects harder to run by raising related costs (Geiger et al., 2018: p. 22).

The notion of "*lawful access*" is somehow delineated in Recital 10 of the new Directive by giving the examples of access to a work through a subscription or access to a work through an open access license. Further, Recital 14 points out that "*lawful access*" should be understood as covering access to content based on an open access policy or through contractual arrangements between rightholders and research organisations or cultural heritage institutions, such as subscriptions, or through other lawful means. For instance, in cases of subscriptions taken by research organisations or cultural heritage institutions, the persons attached thereto and covered by those subscriptions should be deemed to have lawful access. Lawful access should also cover access to content that is freely available online.

In the case of "*lawful access*" to a database, the notion of "*lawful access*" of a database should be interpreted as meaning users who respect the terms of use and the conditions of access to a database provided by the publisher (or the library); in this case "*lawful access*" is similar to the notion of "*normal use*" of the database (Geiger et al., 2018: p. 25) or "*lawful use*" which is described in Recital 33 of the InfoSoc Directive as a use which is authorized by the rightholder or which is not restricted by law (Directive 2001/29/EC, Recital 33)[29]. It should also be interpreted as meaning access to a database allowed by an existing exception

[28]In 01/09/2017 Germany amended its Copyright law and the amendment has come into force as of 01/03/2018 introducing TDM in Section 60d titled "*Text and data mining.*"

[29]Directive 2001/29/EC, Recital 33 according to which *A use should be considered lawful where it is authorised by the rightholder or not restricted by law.*

or limitation to copyright. In the first case, "*lawful access*" may mean that these conditions of access—normal use allowed through a contractual agreement or a license—provided by the publisher (or the library) can easily prohibit data analysis, and in that case, the TDM exception is completely circumvented and becomes useless in practice.

By no means "*lawful access*" to database allows the circumvention of Technical Protection Means (TPMs) which are set by the rightholder of the database with the aim to protect it. Thus, "*lawful access*" to database does not confer the user with the right to circumvent TPMs set for database's protection. On the other hand, the application of TPMs by the rightholder of the database does not allow him/her to go beyond the proportionality principle. To this end, Recital 16 of the new Directive delineates that in case of TPMs for the protection of security and integrity of a database those TPMs should remain proportionate to the risks involved and should not exceed what is necessary to pursue the objective of ensuring the security and integrity of the database. The rightholder who sets the TPMs must not go beyond the proportionality principle and undermine the effective application of the TDM exception set by the new Directive (Directive 2019/790/EU, Recital 16).

6. The Beneficiary of TDM Exception

In the case of art.3 of the new Directive, the beneficiary of the TDM exception is a research organization and a cultural heritage institution.

The "*research organization*" is defined in art.2 (1) of this Directive, according to which it means a university, including its libraries, a research institute or any other entity, the primary goal of which is to conduct "*scientific research*" or to carry out educational activities involving also the conduct of scientific research: (a) on a not-for-profit basis or by reinvesting all the profits in its scientific research; or (b) pursuant to a public interest mission recognised by a Member State, and in such a way that the access to the results generated by such scientific research cannot be enjoyed on a preferential basis by an undertaking that exercises a decisive influence upon such organisation.

According to Recital 12 of this Directive, "*scientific research*" should be understood to cover both the natural sciences and the human sciences. Proper for running "*scientific research*" is any entity which meets the criteria of art.2 (1) which may not be necessarily a university or a higher education institution and their libraries, but it can be a hospital or a research institution other than a teaching one (Directive 2019/790/EU, Recital 12).

The public-interest mission of the scientific research organization to which art.2 (1) (b) refers could be the outcome of public funding that this organization is granted or could be the result of specific provision in the national law of Member State to which this organization is headquartered or it could be recognized in contractual agreement involving the organization and the State. For the avoidance of any doubt, Recital 12 of this Directive delineates through a negative

description the notion of "*research organization*" by stressing that organisations upon which commercial undertakings have a decisive influence allowing such undertakings to exercise control because of structural situations, such as through their quality of shareholder or member, which could result in preferential access to the results of the research, should not be considered research organisations for the purposes of this Directive (Directive 2019/790/EU, Recital 12).

The notion of "*cultural heritage institution*" is defined in art.2 (3) of this Directive as a publicly accessible library or museum, an archive or a film or audio heritage institution. Recital 13 of this Directive delineates the notion of "*cultural heritage institution*" as any library or museum that is publicly accessible regardless of the type of works or other subject matter that they hold in their permanent collections, as well as archives, film or audio heritage institutions. They should also be understood to include, inter alia, national libraries and national archives, and, as far as their archives and publicly accessible libraries are concerned, educational establishments, research organisations and public sector broadcasting organisations (Directive 2019/790/EU, Recital 13).

In the case of art.4 of the new Directive, the beneficiary of the TDM exception is not restricted to a research organization or a cultural heritage institution. This means that the beneficiary of the exception or limitation for TDM provided in art.4 can be any non-profit or for-profit entity or individual. This fact is confirmed by Recital 18 of Directive 2019/790/EU according to which TDM techniques are widely used both by private and public entities to analyse large amounts of data in different areas of life and for various purposes, including for government services, complex business decisions and the development of new applications or technologies (Directive 2019/790/EU, Recital 18).

7. Purpose-Specific TDM

Art.3 of the Directive 2019/790/EU is purpose-specific: the exception of TDM is provided for the purposes of scientific research. The term "*scientific research*" is understood as in the definition of "*research*" put forward by the OECD; according to it research is understood as "*creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications*" (De Wolf & Partners, 2014: p. 55; Manual, 2002). Scientific research lies in the ambit of that definition. In any case of questionable research activity—in any case of doubts upon the "*scientific*" nature of the research—the burden would lie on the shoulders of the user to prove that his/her research is "*scientific*" and/or that it is implemented within a scientific framework and that the TDM activity undertaken within this framework was carried out for scientific research purposes, too.

Licensing works or other subject matter for uses falling outside the scope of the TDM mandatory exception of the new Directive on Copyright in the DSM is

possible, in consideration of Recital 18 of this Directive which posits that rightholders should remain able to license the uses of their works or other subject matter falling outside the scope of the mandatory exception provided for in this Directive for text and data mining for the purposes of scientific research and of the existing exceptions and limitations provided for in Directive 2001/29/EC (Directive 2019/790/EU, Recital 18).

8. TDM and the Application of the "Three-Step Test"

Art.7 of the Directive 2019/790/EU on Copyright in the DSM rules on common provisions applicable to the exceptions provided through this new Directive on Copyright. Art.7 (2) of said Directive refers to art.5 (5) of the InfoSoc Directive which sets the rule of the so called "*three-step test*". According to art.7 (2) of the new Directive on Copyright in the DSM:

Article 5 (5) *of Directive* 2001/29/*EC shall apply to the exceptions and limitations provided for under this Title. The first, third and fifth subparagraphs of Article* 6 (4) *of Directive* 2001/29/*EC shall apply to Articles* 3 *to* 6 *of this Directive*.

The first sentence of paragraph 2 of article 7, i.e. "*Article* 5 (5) *of Directive* 2001/29/*EC shall apply to the exceptions and limitations provided for under this Title*" focuses on the application of the three-step test on all other exceptions or limitations provisioned in the Title II of the Directive on Copyright in the DSM. Thus, the tree-step test is definitely applicable on TDM; the application of rule of art.5 (5) of the InfoSoc Directive on TDM is clearly addressed through the provision of art.7 (2) of the Directive on Copyright in the DSM.

The three-step test is not new in Copyright law. It is embodied not only in the Berne Convention (art.9 (2) but also in the TRIPs Agreement (art.13) and the WIPO Internet Treaties (art.10 of WIPO Copyright Treaty (WIPO Copyright Treaty, 1996) & art.16 of WIPO Performances and Phonograms Treaty (WIPO Performances and Phonograms Treaty, 1996). The substantial part of its wording in these international treaties has remained unchanged. The three-step test as is also provisioned in art.5 (5) of the InfoSoc Directive has become the cornerstone for almost all exceptions to all intellectual property rights at the international—European—level. The three-step test provision of article 5 (5) of Directive 2001/29/EC form a uniform element of European Copyright law which always consists of the same building blocks:

1) It is enunciated that Copyright limitations must be certain special cases.

2) It is clarified that these limitations may not conflict with the normal exploitation of the work, and

3) It must be ensured that the limitations do not unreasonably prejudice the legitimate interests of the author.

The three-step test sets forth three abstract criteria:

- Criterion 1: As a general rule, limitations to Copyright law are allowed in certain special cases

This rule is delineated by two subsequent criteria determining that:

- Criterion 2: there may neither be a conflict with the normal exploitation of the work;
- Criterion 3: nor an unreasonable prejudice to the legitimate interests of the author (and/or right-holder and/or user—depending on the interpretation).

In addition to article 5 (5) of Directive 2001/29/EC the wording of article 9 (2) of the Berne Convention, of article 13 of the TRIPs Agreement and of articles 10 and 16 of the WIPO Copyright and Performances & Phonograms Treaties give evidence of this structural legal edifice of the "*three-step test*" in Copyright. The three criteria have always been understood to be cumulative (WTO Document WT/DS160/R, 2000 in case USA—Section 110 (5) of the US Copyright Act presented on June 15, 2000; Marinos, 2003: p. 93 et seq.; Kallinikou, 2008: pp. 277-278)[30]. This means that limitations on Copyright have to meet all three criteria to be considered permissible, that is, all of them apply jointly to limitations so that if a limitation fails to comply with any one of the steps, it does not pass the test (Knights, 2000: p. 3). And if it does not pass the test, the limitation to Copyright law must be abolished. All three criteria of the three-step test must be deemed relevant tests deserving the same interpretative effort.

The three-step test is located at the interface between the author's exclusive rights and privileged uses. Its three steps make it possible to approach the core of Copyright's balance in stages. The first step is the furthest from the core of Copyright's nature and correspondingly is of a general nature. Its sets forth the basic rule of criterion 1, i.e. limitations of Copyright must be restricted to certain special cases. Copyright limitations which are incapable of fulfilling this criterion are inevitably doomed to fail. The second step delineates the basic rule of criterion 1 by setting up criterion 2, i.e. limitations to Copyright must not lead to conflict with the normal exploitation of the work; the conflict with the normal exploitation of the work is not permissible. At this stage no additional instruments, like the payment of the equitable remuneration, for the reconciliation of the interests of authors and users are necessary. Limitations that fail to meet this condition cannot be countenanced at all. The third step sets up criterion 3 which is the closest among all three of them to the core of Copyright's nature. In order to pass, limitation on Copyright must not be an unreasonable prejudice to the legitimate interests of the author and/or right-holder (and for some legal theorists, it must not be an unreasonable prejudice to the legitimate interests of the author and/or right-holder and/or user of the copyrighted work).

[30]WTO Decision Panel has expressed the opinion that the three conditions of the Three-Step test must apply on a cumulative basis, each being a separate and independent requirement that must be satisfied. See case *USA—Section 110(5) of the US Copyright Act* presented on June 15, 2000, WTO Document WT/DS160/R, available at URL: http://www.wto.org/english/news_e/news00_e/1234da.pdf [last check, July 1, 2019], 6.97; and that each of the three steps must be presumed to mean something different from the other two, or else there would be redundancy. See case *Canada—Patent Protection of Pharmaceutical Products* presented on March 17, 2000, WTO Document WT/DS114/R, available at URL: http://www.wto.org/english/tratop_e/dispu_e/7428d.pdf [last check, July 1, 2019], 7.21.

9. Conclusion

It follows from the above that TDM is expected to benefit libraries and archives in connection to web-harvesting and web-archiving, greatly reinforcing the adventure of scientific research in Europe. We need to see, however, the way the various EU Member-States will incorporate the exception of text and data mining and also, the way in which libraries and other cultural organizations will act upon this incorporation. Text and data mining in libraries do not operate within a vacuum; libraries and archives will need to devote to this endeavor perhaps significant resources of personnel and infrastructure; a matter of political decision, also, the use of TDM remains in the hands of library administrators. The profound benefits of web-harvesting and web-archiving will hopefully aid at this kind of important, for scientific research, decision-making in the European libraries and archives of the 21st century.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Bergman, C., Hunter, L., & Rzhetsky, A. (2013). *Announcing the PLOS Text Mining Collection*. https://blogs.plos.org/everyone/2013/04/17/announcing-the-plos-text-mining-collection/

Case C-203/02 (2004). *The British Horseracing Board Ltd and Others v. William Hill Organization Ltd.* http://curia.europa.eu/juris/liste.jsf?num=C-203/02

Case C-30/14 (2015). *Ryanair Ltd v. PR Aviation BV.* http://curia.europa.eu/juris/document/document.jsf?docid=161388&doclang=EN

Case C-5/08 (2009). *Infopaq International A/S v Danske Dagblades Forening. Eu ZW 655.* http://curia.europa.eu/juris/liste.jsf?num=C-5/08

Caspers, M., Guibault, L., McNeice, K., Piperidis, S., Pouli, K., Eskevich, M., & Gavriilidou, M. (2016). *Reducing Barriers and Increasing Uptake of Text and Data Mining for Research Environments Using a Collaborative Knowledge and Open Information Approach*. Baseline Report of Policies and Barriers of TDM in Europe. https://cordis.europa.eu/project/rcn/197301_en.html

De Wolf & Partners (2014). *Study on the Legal Framework of Text and Data Mining* (*TDM*). European Union. https://publications.europa.eu/en/publication-detail/-/publication/074ddf78-01e9-4a1d-9895-65290705e2a5/language-en

European Commission & COM 192 Final (2015). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee, and the Committee of the Regions, a Digital Single Market Strategy for Europe.* Document: 52015DC0192, Brussels. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2015%3A192%3AFIN

European Commission & SWD 301 Final Part 1/3 (2016). *Commission Staff Working Document, Impact Assessment on the Modernization of EU Copyright Rules.* https://ec.europa.eu/digital-single-market/en/news/impact-assessment-modernisation-eu-copyright-rules

European Commission & SWD 301 Final Part 2/3 (2016). *Commission Staff Working Document, Impact Assessment on the Modernization of EU Copyright Rules.* Brussels. https://ec.europa.eu/transparency/regdoc/rep/10102/2016/EN/SWD-2016-301-F1-EN-MAIN-PART-2.PDF

European Commission & SWD 302 Final (2016). *Commission Staff Working Document, Executive Summary of the Impact Assessment, on the Modernization of EU Copyright Rules.* Document: 52016SC0302, Brussels. http://ec.europa.eu/transparency/regdoc/rep/10102/2016/EN/SWD-2016-302-F1-EN-MAIN-PART-1.PDF

European Copyright Society (2017). *General Opinion on the EU Copyright Reform Package.* https://www.ivir.nl/publicaties/download/ECS_opinion_on_EU_copyright_reform.pdf

Frascati Manual (2002). *Proposed Standard Practice for Surveys on Research and Experimental Development*. OECD. https://www.oecd-ilibrary.org/docserver/9789264199040-en.pdf?expires=1542611355&id=id&accname=guest&checksum=39B756986E0ECF728154E3785B2AA363

Geiger, C., Frosio, G., & Bulayenko, O. (2018). *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market-Legal Aspects.* Centre for International Intellectual Property Studies (CEIPI) Research Paper. https://doi.org/10.2139/ssrn.3160586 http://www.europarl.europa.eu/RegData/etudes/IDAN/2018/604941/IPOL_IDA(2018)604941_EN.pdf

Hargreaves, I. (2011). *Digital Opportunity: A Review of Intellectual Property and Growth.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/32563/ipreview-finalreport.pdf

Hargreaves, I., Guibault, L., Handke, C., Martens, B., Lynch, R., & Filippov, S., (2014). *Standardisation in the Area of Innovation and Technological Development, Notably in the Field of Text and Data Mining—Report from the Expert Group*. European Union.

IFLA (2013). *IFLA Statement on Text and Data Mining.* https://www.ifla.org/publications/node/8225

Kallinikou, D. (2008). *Πνευματική Ιδιοκτησία και Συγγενικά Δικαιώματα* (*Copyright and Related Rights*). Athens: P.N. Sakkoulas.

Knights, R. (2000). *Limitations and Exceptions under the "Three-Step Test" and in National Legislation—Differences between the Analog and Digital Environments.* WIPO/DA/MVD/00/4, 1-18. https://www.wipo.int/edocs/mdocs/copyright/en/wipo_da_mvd_00/wipo_da_mvd_00_4.pdf

Marinos, M. T. (2003). *Περιορισμοί και εξαιρέσεις από το δικαίωμα πνευματικής ιδιοκτησίας—Ερμηνεία και Έλεγχος των τριών βημάτων* (*Limitations and Exceptions to Copyright—Analysis and Application of the Three-Step-Test*). In M. T. Marinos (Ed.). *Κοινωνία Πληροφοριών και Πνευματική Ιδιοκτησία—Η ελληνικήρύθμιση* (*Information Society and Copyright Law—The Hellenic Rule*). Athens, Thessaloniki: Sakkoulas Publications.

Max Planck Institute for Innovation and Competition (2016). *Position Statement on the Proposed Modernisation of European Copyright Rules.* Research Paper No. 17-12. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3036787

Reilly, B.F. (2012). *When Machines Do Research, Part 2: Text-Mining and Libraries*.

Sag, M. (2009). Copyright and Copy-Reliant Technology. *Northwestern University Law Review*, *103*, 1607-1682. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1257086

Sag, M. (2019). The New Legal Landscape for Text Mining and Machine Learning. *Journal of the Copyright Society of the USA*, *66*. https://doi.org/10.2139/ssrn.3331606 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3331606

Smit, E., & Van der Graaf, M. (2011). *Journal Article Mining 2011: A Research Study into Practices*, *Policies*, *Plans.....and Promises.* Publishing Research Consortium, Amsterdam. http://publishingresearchconsortium.com/index.php/128-prc-projects/research-reports/journal-article-mining-research-report/160-journal-article-mining

Stamatoudi, I. (2016). Text and Data Mining. In I. Stamatoudi (Ed.) *New Developments in EU and International Copyright Law,* Kluwer Law International (pp. 251-282).

Supreme Court of the United States (1994). *Case Campbell v. Acuff-Rose Music* (*92-1292*), *510 U.S. 569.* https://www.law.cornell.edu/supct/html/92-1292.ZO.html

Tsolakidou, E. (2018). *Word and Document Embeddings: An Application in the Greek Language.* Master Thesis, Thessaloniki: Aristotle University of Thessaloniki. http://geolib.geo.auth.gr/index.php/grelit/article/view/12320

United States Court of Appeals, Second Circuit (2014). *Case Authors Guild v. HathiTrust*, *755 F.3d 87*, *97-98* (*2d Cir. 2014*). https://casetext.com/case/authors-guild-inc-v-hathitrust-1

United States Court of Appeals, Second Circuit (2015). *Case Authors Guild v. Google*, *Inc.*, *804 F.3d 202.* https://scholar.google.gr/scholar_case?case=2220742578695593916&q=Authors+Guild+v.+Google,+Inc.,+804+F.+3d+202&hl=en&as_sdt=2006&as_vis=1

Walter, M., & Lewinski, S. (2010). *European Copyright Law—A Commentary.* Oxford: Oxford University Press.

World Intellectual Property Organization (1996). *WIPO Copyright Treaty.* Geneva. http://www.wipo.int/treaties/en/ip/wct/trtdocs_wo033.html

World Intellectual Property Organization (1996). *WIPO Performances and Phonograms Treaty.* Geneva. http://www.wipo.int/treaties/en/ip/wppt/trtdocs_wo034.html