

Computer Ethics - Philosophical Enquiry (CEPE) Proceedings

Volume 2019 *CEPE 2019: Risk & Cybersecurity*

Article 11

5-28-2019

Legal and Technical Issues for Text and Data Mining in Greece

Maria Kanellopoulou - Botti

Department of Archives, Library Science and Museology, Ionian University, Greece

Marinos Papadopoulos

Ionian University, Greece

Christos Zampakolas

Ionian University, Greece

Paraskevi Ganatsiou

Ionian University, Greece

Follow this and additional works at: https://digitalcommons.odu.edu/cepe_proceedings



Part of the [Applied Ethics Commons](#), [Archival Science Commons](#), [Digital Humanities Commons](#), [Information Security Commons](#), [Internet Law Commons](#), [Legal Ethics and Professional Responsibility Commons](#), [Risk Analysis Commons](#), and the [Science and Technology Law Commons](#)

Custom Citation

Kanellopoulou - Botti, M., Papadopoulos, M., Zampakolas, C., & Ganatsiou, P. (2019). Legal and technical issues for text and data mining in Greece. In D. Wittkower (Ed.), *2019 Computer Ethics - Philosophical Enquiry (CEPE) Proceedings*, (19 pp.). doi: 10.25884/yp3n-dq78 Retrieved from https://digitalcommons.odu.edu/cepe_proceedings/vol2019/iss1/11

This Paper is brought to you for free and open access by ODU Digital Commons. It has been accepted for inclusion in Computer Ethics - Philosophical Enquiry (CEPE) Proceedings by an authorized editor of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.

Legal and Technical Issues for Text and Data Mining in Greece

Maria Bottis
Ionian University

Marinos Papadopoulos
Ionian University

Christos Zabakolas
Ionian University

Paraskevi Ganatsiou
Ionian University

Abstract

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. The proposed participation in the 2019 CEPE Conference aims at presenting certain issues related to the existing legal framework as well as technical/librarianship issues that apply to Web harvesting and archiving. The aforesaid proposed conference participation will elaborate upon the applicable legal framework with the aim to shed light upon what is legally sound and what is not in relation to web harvesting techniques and processes. It will also elaborate upon technicalities of TDM leveraged for the implementation of TDM. Currently, the EU Commission aims at promoting the efficient use of text and data mining (TDM) for scientific research purposes. Regarding TDM, the EU Commission opts for making Member States to provide for an exception to the rights provided for in article 2 of Directive 2001/29/EC, articles 5(a) and 7(1) of Directive 96/9/EC and article 11(1) of the proposed Directive for reproductions and extractions made by research organizations in order to carry out text and data mining of works or other subject-matter to which they have lawful access for the purposes of scientific research. Thus, regarding TDM in the EU legal environment, the new Directive considers text and data mining to be an exception to the reproduction right of Copyright aimed solely for research. The exception for scientific research can, in certain circumstances, cover the acts of reproduction performed in the course of data analysis activities even in the existing "acquis communautaire" through the provision of article 5(3) of Directive 2001/29/EC. Web harvesting and archiving in Greek academic libraries is at its embryonic stage, currently, at least in consideration of article 4§4(b) of Law 4452/2017 which rules that the National Library of Greece is empowered with the right to deploy TDM in Greece and to oversee the deployment of TDM through other libraries. For academic libraries in Greece, the research upon the methods and applications for web harvesting as well as

upon the policies related to said subject matter is of special interest. Legal stumbling blocks exist, both with respect to the data collection in the Web harvesting phase as well as to data sharing in the archiving and making available to the public of the Web harvested output.

Keywords: *web harvesting, web archiving, TDM, text mining, libraries and archives*

TDM in the proposed new EU Directive on Copyright in the DSM

This paper elaborates upon the 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). As of the writing of this paper, the text of the new Directive remains to be published in the Official Journal of the European Union; on March 26, 2019, Axel Voss, a German politician and lawyer who serves as the assigned Rapporteur in drafting the aforesaid 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. On April 15, 2019, the text of the new EU Directive on Copyright in the DSM was adopted by the European Council after the EU Parliament's vote; on April 17, 2019, the final Act of said EU Directive on Copyright in the DSM was signed, thus the end of the procedure in the EU Parliament and the awaiting of the publication of the new EU Directive on Copyright in the DSM in the Official Journal.

The need for statutory exception from copyright for the sake of TDM and not the licensing¹ has long been requested in consideration of the fact that existing legal framework in the EU does cater for such an exception as well as of the fact that the "*Acquis Communautaire*" as it has been implemented in EU Member States does not cover TDM activity and is to blame for legal uncertainty in the EU regarding TDM, scientific research, and copyright protection.²

¹ See, for example, *IFLA Statement on Text and Data Mining*, (2013), available at URL: <https://www.ifla.org/publications/node/8225> (last check, May 25, 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.

² 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

Thus, the text of the new Directive on Copyright in the DSM has been approved through the European Parliament legislative resolution of 26 March 2019 on the proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market (COM(2016) 0593 – C8-0383/2016 – 2016/0280(COD)), and the position of the European Parliament adopted at first reading on 26 March 2019 with a view to the adoption of Directive (EU) 2019/... of the European Parliament and of the Council on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.

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 rightholders, 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 so called “*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.

For the EU legislator, TDM is just a means to achieve the goal of Digital Single Market. The goal for an EU DSM is a goal for 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 (CNECT - Communications Networks, 2016)³ 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

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.

³ See **COM(2015) 192 final**, 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, available at URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2015%3A192%3AFIN> [last check, May 26, 2019].

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 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. The suggestion 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 is titled “*Text and data mining for the purpose of scientific research*”; article 4 is titled “*Exception or limitation for text and data mining*.”⁴ 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) be implemented across all EU Member States in order to ensure effective harmonization of the law;
- b) do not be subject to contractual overrides regarding TDM implemented for scientific purpose; and
- c) do 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. 59) For this reason, the

⁴ Considers text of the new Directive on Copyright in the DSM as in PE-CONS 51/19, April 2, 2019.

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*” ...

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 aforesaid provision of art.7(1) does not refer 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 there’s no possibility for contractual override of TDM in the case of TDM implemented for scientific purpose.

There were many suggestions on how to encourage TDM for research purposes without fear of infringing IP rights. The goal for such an encouragement through legislative action could be achieved in a number of ways: (Hargreaves, et al., 2014, p. 52) through an adjustment of licensing practices; through a revised, normative interpretation of the reproduction right in copyright; through the introduction of a new mandatory exception in copyright and database laws, or 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.

In consideration of proposal for a new Directive on Copyright in the DSM, i.e. COM(2016)593 final 2016/0280(COD), and the voted text of the new Directive by the EU Parliament on March 26, 2019, there’s no doubt that 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, a.k.a. the encouragement of TDM through licensing was deemed to be inefficient and not adequate for creating legal certainty among Member States regarding TDM for scientific research (CNECT - Communications Networks, 2016, part 2/3 pp.51-52)⁵. 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 (CNECT - Communications Networks,

⁵ 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).

2016, part 1/3 pp.104-105)⁶ 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. (European Commission, 2016, p. 106).

As said, 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 said 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—should, therefore, be amended.(CNECT - Communications Networks, 2016, 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 said provision in their national law, there’re significant differences in the texts and accorded protection of national laws.

The new Directive on Copyright in the DSM includes art.3 and art.4 which address the issue of TDM. Article 3 is titled “*Text and data mining for the purpose of scientific research*”; article 4 is titled “*Exception or limitation for text and data mining.*”

⁶ 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. See, more at European Commission, (2016), [last check, May 26, 2019] (European Commission, 2016)

Article 3

Text and data mining for the purposes of scientific research

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.

Article 4

Exception or limitation for text and data mining

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 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.:

- d) They are designed to be implemented across all EU Member States in order to ensure effective harmonization of the law;
- e) They are not subject to contractual overrides regarding TDM implemented for scientific purpose; and
- f) They are not subject to lock-up behind technological protection measures.

According to Recital 5 of the new Directive on Copyright in the DSM, 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. Therefore, 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—should be amended.

Because of the fact that most exceptions or limitations to copyright in the EU legal framework are non-mandatory, they are not implemented the same in EU Members' legal systems, and they are not fully adapted to the use of technologies such as TDM technologies used in scientific research. Therefore, there has been in Europe legal uncertainty concerning 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 exceptions and limitations to copyright is a cause of failure, actually in the process of harmonization of copyright rules applicable in all Member States of the EU. 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, caused significant difficulties in leveraging on the existing legal framework for Copyright for covering the TDM activity.

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, 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.(Europese Commissie, 2016, 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 (Europese Commissie, 2016, Recital 11).

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 correlations⁷. 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.(Europese Commissie, 2016, Recital 8)⁸ 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 program

⁷ See definition of TDM in the art.2(2) of the voted text of the Directive on Copyright in the DSM.

⁸ 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.

mining 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. (European Commission, 2016, 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. (European Commission, 2016, 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 understood 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. Thus, use of works in the TDM process equals to “*non-expressive use*” (Sag, 2009, v.103)—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). TDM 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).

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.

To undertake TDM a researcher must access, or arguably make a copy of the articles/data in order to apply the necessary algorithms. In the European legal system such access and the making of copy is bound by the exclusivity power of copyright

holder irrespectively of arguments claiming that it is the facts dispersed throughout the content and relationship between the facts which are of interest to scientific researchers, neither of which are in themselves protected by copyright.

TDM, embedding algorithmic applications and algorithms for scientific research

TDM is indeed a technological solution that leverages on the development of new algorithms and the use of new information technology applications. TDM is of such pivotal importance to research and of such high economic value that it needs to be readily available not only to academic researchers, but also to scientific research conducted in the commercial arena. The text of the new Database Directive on Copyright in the DSM seems to acknowledge that. Economic arguments suggest that the welfare gains from commercial TDM would greatly exceed those available from non-commercial TDM. This argument also holds that making a distinction in law between ‘*commercial*’ and ‘*non-commercial*’ research would be difficult if not impossible, especially in a time when academics are encouraged, increasingly, to collaborate and ‘co-create’ with business. (Hargreaves et al., 2014, p.14)

To undertake TDM a researcher must access, or arguably make a copy of the articles/data in order to apply the necessary algorithms. As seen above hereto, for the US legal framework such access and the making of copy of the copyrighted work in order to apply the necessary algorithms is covered by the ‘*fair-use*’ doctrine of US Copyright Act. In the European legal system, though, such access and the making of copy is bound by the exclusivity power of copyright holder irrespectively of arguments claiming that it is the facts dispersed throughout the content and relationship between the facts which are of interest to scientific researchers, neither of which are in themselves protected by copyright. (Hargreaves et al., 2014, p.20). Such arguments have had a stronger heard in Australia; the Australian Industry Information Association (AIIA) suggested that the introduction of a specific exception to permit TDM “*would not negatively impact on the original data provider’s rights and commercial interests because the technology is not intended to reprint the original data, but to provide a synthesized result. These outcomes do not interfere with the economic value of the copyright material nor compete with it*” (Hargreaves et al., 2014, p.21). The AIIA has been a fervent supporter for the adoption in Australian Copyright law of a ‘fair use’ exception in place of the ‘fair dealing’ in the sense of ‘far use’ as in the US Copyright law, (Carter, 2018).⁹ In the same vein, the Australian Libraries Copyright Committee

⁹ See AIIA, (2018), **Copyright Modernisation Consultation AIIA response**, available at URL: https://aiia.com.au/_data/assets/pdf_file/0004/88915/AIIA-response-Copyright-modernisation-V-Final.pdf [last check, July 1, 2019], according to which the AIIA strongly urges government to adopt a ‘fair use’ exception in place of the ‘fair dealing’ exceptions. This is notwithstanding the additional exceptions proposed. Specifically, AIIA supports the US model of the four statutory fairness factors and six illustrative purposes – harmonisation with common international practices will mean less barriers for Australian businesses to exploit their copyright and use copyrighted material. While some of the additional fair dealing exceptions proposed (i.e. the incidental or technical use exception and the text and data mining exception) will go some way in reducing the barriers, the fundamental problem with

welcomes the recommendation for the adoption of a flexible fair use exception in Australian copyright law (Australian Libraries Copyright Committee, 2016)¹⁰ and considers TDM to be within the transformative non expressive uses of copyrighted works, thus to be fit for an open-ended exception from copyright such as ‘fair use’ (Neylon, 2012).¹¹

TDM depends on and involves the application of algorithms which are essentially the backbone of computational methods applied to solve problems/improve performance based on experience. Algorithms are behind computational data driven tasks and the use of statistics, probability theory and optimization to learn from them. The application of algorithms for text and document classification is typical in libraries and archives wherein documents available online are harvested and archived. As the number of documents available online and the size of each document constantly increase, properly classifying them becomes more difficult but also more imperative. Text classification is the process of identifying the category in which a document belongs (selected from a specified set of categories). Text classification is a key task in Natural Language Processing (NLP).

NLP refers to any task of automated processing of natural language (written and spoken) such as machine translation, automatic summarization, paraphrasing and many others. NLP techniques can, for instance, be used for practical applications such as opinion mining and trend detection based on information available of the web. The entire World Wide Web can be thought of as a large collection of linguistic information that can be search, processed and classified. This approach is referred to as the “*Web as Corpus*” (Tsolakidou, 2018, p.28).

Very often, text classification is seen as a supervised learning task in which labeled documents are given as input to the classifier in order for it to accurately identify the categories of new documents (Tsolakidou, 2018, pp.29-30). Text classification leverages on algorithms that process text data at scale in applications such as:

- ✓ *Document Organization and Retrieval*: This refers to the creation and management of large digital libraries of documents, web collections, scientific literature, or even social feeds. TDM may be used for the harvesting and archiving from the Web of works that fit the thematic interest of a library’s special or general collection of works. In that case, feature selection becomes an important issue for text classification. This refers to determining the features which are most relevant to the classification

fair dealing still remains. For AIIA, fair dealing is more complex, more costly and less fixable than the alternative fair use option.

¹⁰ See ALCC, (2016), ***Australia’s libraries and archives support the fair use***, available at URL: <http://libcopyright.org.au/news/australias-libraries-and-archives-support-fair-use> [last check, July 1, 2019], according to which the ALCC welcomes recommendations to allow room in Australian copyright system for all uses that are fair, rather than privileging some uses and users above others.

¹¹ For ALCC, uses which may have been characterized as transformative, such as text and data mining, may be better seen as ‘non-expressive’ or ‘orthogonal’ uses. Fair use in the US Copyright law provides the flexibility for new technologies to develop which may straddle the two definitions, and similarly providing courts with the tools to deem when such uses will unreasonably harm the copyright owner. See more at Australian Law Reform Commission, Australian Government, ***Non-consumptive Use – Text and Data Mining***, available at URL: https://www.alrc.gov.au/publications/8-non-consumptive-use/text-and-data-mining#_ftnref89 [last check, July 1, 2019].

process which is very important because some of the words are much more likely to be correlated to the class distribution than others.

In the for-profit organizations' market algorithms are commonly used for text and data classification in applications such as:

- ✓ *News filtering*: online news services deal with a large volume of articles created daily. The sheer volume makes manual organization very hard. Therefore, automated classification methods can be very useful for news categorization in web portals.
- ✓ *Opinion Mining/Sentiment Analysis*: Customer reviews or opinions are often short text documents which can be mined to determine useful information such as whether the reviewer is positively or negatively inclined and even his emotional state.

The above text and data classification algorithmic applications, these that are common for for-profit organizations, may be used in the context of a non-profit library that caters for its public image and how it is perceived by its stakeholders. Algorithmic applications that can be used in the library environment may also include:

- ✓ *Native Language Identification (NLI)*. NLI is the task of identifying the native language of authors of texts written in a (potentially) foreign language. NLI is modeled as a text classification task with labels corresponding to native languages. The basis of NLI is the assumption that one's mother tongue influences the way they acquire and produce second languages (Second Language Acquisition – SLA) and that traits easily identifiable in speech production should be identifiable in written texts as well (Tsolakidou, 2018, p.31).
- ✓ *Word-sense disambiguation* which is the process of identifying the particular meaning of a word based on the way it is used in a sentence and its context. A more advanced task is *Named Entity Recognition (NER)* which involves identifying and tagging among others, people's names, organizations and geographical locations within the text.
- ✓ The *bag-of-words (BoW)* model of text. As the name suggests, this model treats documents like bags of words, i.e. as containers where the order of items does not matter. Bags are essentially sets that are allowed to have more than one instances of the same item, meaning that a word may be found in the bag (document) multiple times. This is referred to as multiplicity and it is maintained in this model. The idea behind BoW is that documents are similar if they have similar content and that we can learn something about the meaning of the document from its content (Tsolakidou, 2018, p.35).
- ✓ The *Term Frequency-Inverse Document Frequency (TF-IDF)* which is a numerical statistic used very often in information retrieval applications to estimate the importance of a term in a document, a collection or an entire corpus. The idea behind TF-IDF is simple and straightforward and relies on the two factors included in its name. The combination of these two factors tends to correspond to the way human minds tend to evaluate search relevance. TF-IDF measures the relative concentration of a term in a given set of documents/articles. If a word is common in

a given item but relatively rare elsewhere then the score should and will be high, i.e. the document is very relevant to the search term. Inversely, if a word occurs few times in one document and many times in other documents the TF-IDF score will be relatively low (Tsolakidou, 2018, pp.40-42).¹²

- ✓ The *BM 25*, with BM standing for Best Matching is a variation/improvement on TF-IDF which focuses on assessing the relevance of a document to a query and is widely used for results ranking in search engines. It is often referred to as *Okapi BM 25* as it was developed in the context of the Okapi information retrieval system at London's City University in the 1980s and 1990s. BM 25 combined previous variants *BM 11* and *BM 15* into a single weighting function. In this function, the IDF component is preserved while the TF component is redefined and based on two new parameters (k1 and b) (Tsolakidou, 2018, pp.43-44).

Word and document embedding algorithms that can be used in the library environment may include:

- ✓ *Word2vec* is a tool for computing continuous distributed representations of words that was created by a team of researchers led by Tomas Mikolov at Google in 2013 and is distributed as open source software with an Apache License. Word2vec is essentially a group of related models that provides an efficient implementation of the continuous bag-of-words and skip-gram model architectures to compute distributed vector representations of words from very large data sets (Tsolakidou, 2018, pp.47-54). The tool first constructs a vocabulary from the training text data and then learns vector representation of words. Following this transformation, the vector representations can be fed into many natural language processing applications such as text classification or machine translation.
- ✓ *Doc2vec* is a tool for computing continuous distributed representations of phrases or sentences; it's a document embedding algorithmic tool that can provide vector embeddings for entire documents (Tsolakidou, 2018, pp.61-64).

Tinkering with TDM in Greece

A recent development in Greece's legal framework on the National Library of Greece (NLG) stipulates for activities that are within the TDM operation. Specifically, law 4452/2017 which is titled "*Regulation on State Language Certificate subject matter, on the National Library of Greece and on other provisions*" includes in its text the provision of art.4(4)(b) according to which the National Library of Greece operates as the official National Depository and Archive of digital publications, data and metadata produced in the country or related to Greek culture. This operation includes the monitoring and archiving of the Internet (web archiving) or other technology environment. To this end, the National Library of Greece shall undertake, allocate and coordinate the actions concerned at national level.

¹² see, also, the same for applications and variations of TF-IDF on pp.42-43.

This provision of art.4(4)(b) of law 4452/2017 is the first provision in the Greek legal system that caters for TDM activities, actually. Art.4(4)(b) of law 4452/2017 preceded any EU regulation upon TDM in Greece and the Europe's "*acquis communautaire*".

There is a paradox in the ruling of law 4452/2017 though: the Greek legislator rules upon the key TDM-player in the Greek market despite the fact that it has yet to rule upon the TDM-game!

Once the provision of art.4(4)(b) of law 4452/2017 became effective, NLG made its first attempts with TDM. On February 2017 NLG deployed TDM for the first time targeting Greek websites at national level. This first attempt was a broad crawling of the Web for websites under the .gr domain or websites under the .edu or .com domains which were composed in Greek.

These first attempts of web harvesting and archiving in Greece were implemented in five different stages extending in two major harvests. They are described succinctly in the following table hereto.

Working Stages Of Web Archiving In Greece By NLG

Stage I	Economic and technical study on the needs and content of the Greek web harvest. Study of international experience	1 st web harvest: broad crawl – national level: text data only
Stage II	Definition of "Greek" sites to be mined	
Stage III	Data Analysis of 1st web harvest to create a National Web Archiving System	
Stage IV	Installing and checking the operation of tools for all phases of national web archiving: extraction, archiving / classification and finally, user search and access): Heritrix for harvesting, Solr for indexing and Open Wayback for web site reconstitution. Netarchive Suite using.	2 nd web harvest: broad - national level: text only) thematic (text and images)
Stage V	Developing a National Archiving System of Greek Web ("ΕΣΑΕΙ"): the Greek user/librarian interface	

The harvesting of works in the Greek Web was twofold:

1. Bulk harvesting &
2. Selective harvesting

In either bulk or selective the harvesting focused on the Greek Web, i.e. targeted works on the Internet composed in the Greek language, and also targeted works uploaded under *.gr* domains or under the *.edu* or *.com* domains. There exist more than 170.000 Web sites registered in the *.gr* top-level domain, or are hosted under the *.net*, *.com*, or *.org* top-level domains. In order to accumulate the first web pages, both endogenous and exogenous discovery (Masanés, 2006) was used. At first, two web portals¹³ were used, in order to compile about 45000 seed URLs, which were used for the exogenous discovery. URLs that followed Robots Exclusion Protocol, contained multimedia or required password were excluded from the harvesting.

The selective harvesting was implemented in consideration of certain criteria These were the:

1. Subject/Topic criterion: three subject harvestings were carried out, regarding the collections of “*Local Government*”, “*News*” and “*Education*”.
2. Creator/Provenance criterion:
 - a. The sites for the “*Local Government*” collection were harvested through the site of the Ministry of Interior Affairs, combining research where needed, in order to find the relevant site in case of broken links or non-updated URLs.
 - b. The News collection derived from a List of the Secretary General for Media and Communication.
 - c. The sites for the Education collection were harvested from the first two bulk crawls and the domain was the main criteria.
3. Type/Format criterion: only formats related to text and images have been harvested.

Web archive preservation is being implemented by keeping copies of the collection in different servers kept in different buildings of NLG. The user can access works harvested and archived leveraging on the NLG Curator tool. The National Archiving System of Greek Web («ΕΣΑΕΙ» National System)¹⁴ has a user interface in Greek and English languages and searching tools to archive formed from the Greek web archiving process and TDM procedure. The NLG Curator Tool was based on the *Netarchive Suite* solution. NLG emphasized on the use of open, widely used software, according to international best practices. Through *Netarchive Suite*, NLG can perform both bulk and subject crawls.

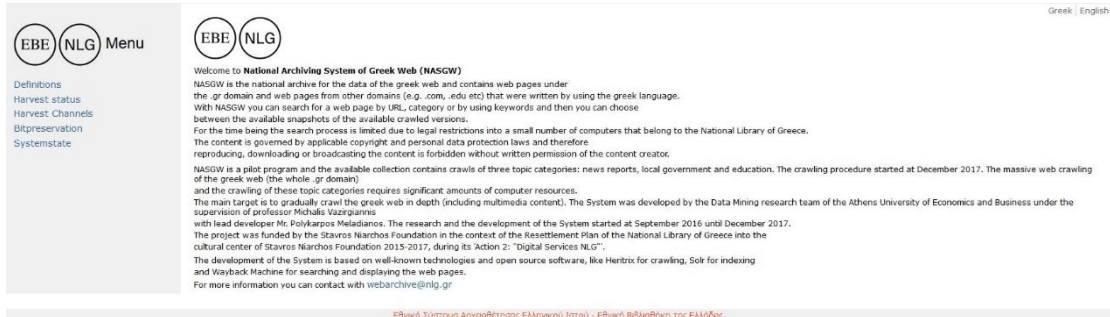
Web archive preservation is being implemented by keeping copies of the collection in different buildings of NLG. So far, no other preservation method has been used (for example emulation, migration etc.). Copies of the collected material are being kept at the Valiano building (which used to host NLG until 2018 and in the new establishment of NLG, in the Stavros Niarchos Foundation Cultural Center).

¹³ dmoz (www.dmoz.org/World/Greek) and greek sites (<http://www.greek-sites.gr/>)

¹⁴ Greek logo of the System (“ΕΣΑΕΙ”) connotes the ancient Greek language, and specifically the phrase (εσαεί < ἔς ἀεί) which means “forever”.

NLG Curator tool user has the ability to search through the subject crawls with subject, URL or full text. The user can also limit the results according to the time that the web pages were harvested.

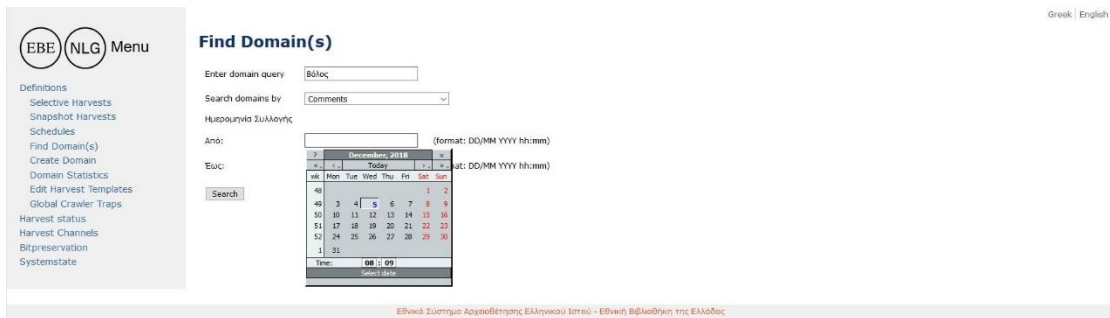
The following captions give an impression of NLG’s Curator tool, currently:



Caption 1: Index of NLG’s Archiving System



Caption 2: Search domain tool of NLG’s Archiving System



Caption 3: Search domain tool of NLG’s Archiving System

References

Australian Libraries Copyright Committee. (2016). Australia’s libraries and archives support fair use | Australian Libraries Copyright Committee. Retrieved July 1,

2019, from <http://libcopyright.org.au/news/australias-libraries-and-archives-support-fair-use>

Carter, J. (2018). Copyright Modernisation Consultation Paper, (July), 2–6. Retrieved from <http://apraamcos.com.au/about-us/government-submissions/>

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*.

CNECT - Communications Networks, D. D. (2016). *COMMISSION STAFF WORKING DOCUMENT IMPACT ASSESSMENT on the modernisation of EU copyright rules Accompanying the document 2/3*. Brussels. Retrieved from <https://ec.europa.eu/transparency/regdoc/rep/10102/2016/EN/SWD-2016-301-F1-EN-MAIN-PART-2.PDF>

Europese Commissie. (2016). EXPLANATORY MEMORANDUM Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on copyright in the Digital Single Market COM(2016)593/F1, 0280, 1–33. Retrieved from <https://ec.europa.eu/transparency/regdoc/rep/1/2016/EN/1-2016-593-EN-F1-1.PDF>

Hargreaves, I., Guibault, L., Handke, C., Valcke, P., 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*. Retrieved from http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf

Masanés, J. (2006). *Web Archiving*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-46332-0_1

Neylon, C. (2012). Text and data mining. In *ALPSP International Conference*. Retrieved from https://www.alrc.gov.au/publications/8-non-consumptive-use/text-and-data-mining#_ftnref89

Sag, M. (2009). *COPYRIGHT AND COPY-RELIANT TECHNOLOGY 103* (No. Vol.103). Retrieved from <https://poseidon01.ssrn.com/delivery.php?ID=0650040990940721010800260241201260770080340680210650360671011111070170761250021241180170160590470501200970801180071020770660701230470290510780210831000731120230800230400521261160050010851180240750991020311100741>

Sag, M. (2019). The New Legal Landscape for Text Mining and Machine Learning. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3331606>

Tsolakidou, E. (2018). *Word and document embeddings: An application in the Greek language*. Aristotle University of Thessaloniki.