European Copyright Law and the Text and Data Mining Exceptions and Limitations

With a focus on the DSM Directive, is the EU Approach a Hindrance or Facilitator to Innovation in the Region?

Author: Charlotte Gerrish
Supervisor: Kacper Szkalej
Table of Contents

Abbreviations ................................................................................................................... 4
Abstract .............................................................................................................................. 5
1. Introduction .................................................................................................................. 5
  1.1 Overview.................................................................................................................. 5
  1.2 Aim and Research Questions .................................................................................. 6
  1.3 Method & Material .................................................................................................. 6
  1.4 Delimitations ........................................................................................................... 7
  1.5 Outline ..................................................................................................................... 7
2. Background .................................................................................................................. 8
  2.1 A new era ............................................................................................................... 8
  2.2 Europe’s Digital Single Market – the Aim to Create a Thriving Environment for Big Data and TDM .... 8
  2.3 Big Data, TDM and Society .................................................................................... 9
    2.3.1 The Concept of “Big Data” - Volume, Velocity, Variety - and Value?................. 10
    2.3.2 Big Data Analytics, AI and Machine Learning .................................................. 11
    2.3.3 TDM as a Method of Managing Big Data .......................................................... 12
  2.4 Proposal for a Directive on Copyright in the Digital Single Market ..................... 13
3. TDM and its Relationship with Copyright .................................................................. 16
  3.1 IPR Applicable to Big Data .................................................................................... 16
    3.1.1 Sources of Copyright Law within the European Union ..................................... 16
    3.1.2 What is Copyright? ......................................................................................... 17
    3.1.3 How Can Copyright Protection Extend to Big Data? ....................................... 17
    3.1.4 The Copyright Monopoly ................................................................................. 19
  3.2 TDM Techniques- Tensions with Copyright ....................................................... 20
    3.2.1 TDM - Acts of Copyright Infringement? ............................................................ 20
    3.2.2 When is the Threshold for Copyright Infringement Met? ................................ 22
      3.2.2.1 TDM Carried Out on Single Words ......................................................... 23
      3.2.2.2 TDM Carried Out on Strings of Words or Phrases ................................. 24
      3.2.2.3 Workarounds ......................................................................................... 25
  4. Pre-DSM Directive Approach to TDM Exceptions and Limitations across the EU ...... 28
    4.1 General Limitations and Exceptions to Copyright Infringement: International Focus ...... 28
    4.2 Limitations and Exceptions Applicable to TDM: EU Focus ................................ 31
      4.2.1 Temporary Acts of Reproduction ................................................................. 32
      4.2.2 Teaching and Scientific Research ................................................................. 34
      4.2.3 Incidental Inclusion ...................................................................................... 35
    4.3 Limitations and Exceptions Applicable to TDM: Member State Focus ................. 36
      4.3.1 England and Wales ....................................................................................... 36
5. The DSM Directive – the Solution to Risk-Free TDM in Europe? 

5.1 The Journey of the TDM Exception in the DSM Directive 

5.1.1 The Proposal 

5.1.2 The DSM Directive – the Final Word 


5.2.1 Harmonisation 

5.2.2 Express Recognition of TDM and Expanded Scope 

5.2.3 Rule Against Contractual Override 

5.2.4 Certainty for Copyright Holders 

5.3 Negative Aspects of the DSM Directive’s TDM Provisions 

5.3.1 The Harmonization is Still Open to Fragmentation 

5.3.2 Undesirable Difference in Treatment Between Research and Other TDM Activities 

5.3.2.1 The Broader TDM Exception is Devoid of Function Due to the Possibility for Contractual Override 

5.3.2.2 The Scope of the Broader TDM Exception Remains Unclear 

5.3.3 Undesirable Restrictions for Research Organizations and Commercial Entities Alike 

5.3.3.1 Qualification Restrictions for the Research Exception 

5.3.3.2 The Issue of “Lawful Access” 

5.3.3.3 The Licensing Burden 

5.3.4 Unresolved Issues Related to Coexistence with TPMs 

5.3.5 Incompatibility with the International Community 

5.4 The Overall Impact of DSM Directive’s Approach to TDM 

6. Another Way: Alternatives and the Future 

6.1 Fair Use – The American Example 

6.1.1 Authors Guild v. Google, Inc – the Google Books Saga 

6.1.2 Fair Use – Not Without Problems 

6.2 A Broad-Ranging Exception – The Japanese Example 

6.3 Making the Best of Things 

6.3.1 Practical Suggestions to Overcome Existing Barriers to TDM in Europe 

6.3.1.1 Education and Raising Awareness 

6.3.1.2 Lawful Access Solutions - Closing the Value Gap 

6.3.1.3 Managing Barriers Created by the Application of TPMs 

7. Conclusion 

8. Bibliography 

9. Appendix 1
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>Artificial Intelligence</td>
</tr>
<tr>
<td>Berne Convention</td>
<td>Berne Convention for the Protection of Literary and Artistic Works</td>
</tr>
<tr>
<td>CDPA</td>
<td>Copyright, Designs and Patents Act 1998, as amended (UK)</td>
</tr>
<tr>
<td>CJEU</td>
<td>Court of Justice of the European Union</td>
</tr>
<tr>
<td>Commission</td>
<td>European Commission</td>
</tr>
<tr>
<td>CPI</td>
<td>Intellectual Property Code (Code de la propriété intellectuelle) (France)</td>
</tr>
<tr>
<td>DRM</td>
<td>Digital Rights Management</td>
</tr>
<tr>
<td>DSM</td>
<td>Digital Single Market</td>
</tr>
<tr>
<td>EU</td>
<td>European Union</td>
</tr>
<tr>
<td>EUR</td>
<td>Euro</td>
</tr>
<tr>
<td>ICO</td>
<td>Information Commissioner's Office (UK)</td>
</tr>
<tr>
<td>ICT</td>
<td>Information and Communication Technology</td>
</tr>
<tr>
<td>IoT</td>
<td>Internet of Things</td>
</tr>
<tr>
<td>IP</td>
<td>Intellectual Property</td>
</tr>
<tr>
<td>Member State(s)</td>
<td>Member countries of the EU - Austria, Belgium, Bulgaria, Croatia, Republic of Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden and the UK</td>
</tr>
<tr>
<td>IPR</td>
<td>Intellectual Property Right(s)</td>
</tr>
<tr>
<td>OCR</td>
<td>Optical Character Recognition</td>
</tr>
<tr>
<td>OECD</td>
<td>The Organisation for Economic Co-operation and Development</td>
</tr>
<tr>
<td>OJ</td>
<td>Official Journal of the European Union</td>
</tr>
<tr>
<td>R&amp;D</td>
<td>Research and development</td>
</tr>
<tr>
<td>SME</td>
<td>Small and Medium Enterprise</td>
</tr>
<tr>
<td>TDM</td>
<td>Text and data mining by automated means, including scanning, extracting, analysing data and compiling results for gaining new insights or to assist machine learning development, without human intervention</td>
</tr>
<tr>
<td>TPMs</td>
<td>Technical Protection Measures</td>
</tr>
<tr>
<td>TRIPS Agreement</td>
<td>The Agreement on Trade-Related Aspects of Intellectual Property Rights</td>
</tr>
<tr>
<td>UK CofA</td>
<td>UK Court of Appeal</td>
</tr>
<tr>
<td>US</td>
<td>United States of America</td>
</tr>
<tr>
<td>USD</td>
<td>United States Dollar</td>
</tr>
<tr>
<td>WCT</td>
<td>World Intellectual Property Organization Copyright Treaty</td>
</tr>
<tr>
<td>WIPO</td>
<td>World Intellectual Property Organisation</td>
</tr>
<tr>
<td>WTO</td>
<td>World Trade Organisation</td>
</tr>
</tbody>
</table>
Abstract

We are in a digital age with Big Data at the heart of our global online environment. Exploiting Big Data by manual means is virtually impossible. We therefore need to rely on innovative methods such as Machine Learning and AI to allow us to fully harness the value of Big Data available in our digital society. One of the key processes allowing us to innovate using new technologies such as Machine Learning and AI is by the use of TDM which is carried out on large volumes of Big Data. Whilst there is no single definition of TDM, it is universally acknowledged that TDM involves the automated analytical processing of raw and unstructured data sets through sophisticated ICT tools in order to obtain valuable insights for society or to enable efficient Machine Learning and AI development. Some of the source text and data on which TDM is performed is likely to be protected by copyright, which creates difficulties regarding the balance between the exclusive rights of copyright holders, and the interests of innovators developing TDM technologies and performing TDM, for both research and commercial purposes, who need as much unfettered access to source material in order to create the most performant AI solutions. As technology has grown so rapidly over the last few decades, the copyright law framework must adapt to avoid becoming redundant. This paper looks at the European approach to copyright law in the era of Big Data, and specifically its approach to TDM exceptions in light of the recent DSM Directive, and whether this approach has been, or is, a furtherance or hindrance to innovation in the EU.

1. Introduction

1.1 Overview

This paper will first focus on the European approach to copyright law in respect of Big Data and TDM, firstly discussing the concepts of Big Data, Machine Learning, AI and TDM. Next, we consider whether Big Data is capable being protected by copyright in light of the Infopaq decision of the CJEU. The exceptions which exist in international law and then which existed at an EU level prior to the adoption of the DSM Directive that are capable of applying to TDM (notably Article 5 of the InfoSoc) are then discussed, along with an appraisal of national law examples from the UK (Section 29A of the CDPA) and from France (Article L.122-5-10 of the CPI). We then analyse the TDM provisions in the recent DSM Directive in light of the proposal and the final text. Next, we assess the legal approach to TDM in the EU when compared with the US and Japan. Finally, and we look at potential solutions to Europe’s approach to TDM and how actors can reconcile innovative activities in light of the DSM Directive. We then provide
a final conclusion as to whether Europe’s approach to TDM is indeed a hindrance or furtherance for innovation in the region and consider whether further research must be conducted in other legal areas to allow us to fully appraise this question.

1.2 Aim and Research Questions
The aim of this paper is to ascertain whether the EU-approach to copyright law and TDM is ultimately a furtherance or hindrance to innovation such as for the exploitation of Big Data Analytics and the development of Machine Learning and AI in Europe. The following questions shall be examined:

- Is Big Data capable of copyright protection?
- How can TDM processes amount to copyright infringement of a copyright holder’s exclusive rights (with a focus on the right to control reproduction and communication to the public)?
- Which limitations and exceptions to TDM activities in Europe existed prior to the advent of the DSM Directive, and are they sufficient?
- Are the provisions related to TDM in the DSM Directive sufficient to further risk-free innovation in the region from both a research and commercial perspective?
- Could Europe have sought inspiration for the DSM Directive from the US and Japan to foster an innovative environment?
- How can European TDM actors, copyright holders and courts best implement the DSM Directive TDM provisions to further innovation in the EU?
- Is the EU approach to copyright law and TDM overall (particularly in light of the DSM Directive) a furtherance or hindrance to European innovation?

1.3 Method & Material
This paper combines several methods of research, given that Big Data, TDM and the related copyrighted issues are situated across commercial, technological and legal disciplines. The paper relies on doctrinal research with a focus on legislation and jurisprudence. Since Big Data and TDM is a global concern, it has also been appropriate to include a comparative element between EU and national laws (i.e., France and the UK). For the purposes of this paper, we have focused on the specific national examples from France and from England and Wales, as these are two jurisdictions with which we are familiar in both an academic and practical sense. It is also appropriate to analyse the American and Japanese legislative and case-law approach to the topic; these jurisdictions having been chosen as they are arguably Europe’s key competitors in the field of technology and innovation. Finally, as the subject matter of Big Data
and TDM is heavily relevant to innovation and business in Europe, we have also undertaken a socio-legal and empirical methodology by reviewing industry insights and obtaining interview data from TDM-innovators and IP law practitioners.

1.4 Delimitations
The scope of this paper is limited to a discussion of the application of copyright law exceptions and limitations in the EU as they apply to Big Data and TDM activities, with a focus on a copyright holder’s economic rights only, notably related to the rights of reproduction and communication to the public. Any impact on the copyright holder’s moral rights is not dealt with in detail. Furthermore, other issues such as the so-called _sui generis_ database right (Directive No. 96/9/EC), TPMs and privacy law (i.e., Regulation No. 2016/679/EC) and their potential impact on Big Data, TDM and innovation in Europe are not discussed at length. The discussion of the DSM Directive is limited to the provisions related to TDM; other controversial provisions on “link tax” and “upload filters” are not discussed. Finally, this paper does not deal in depth with the potential clash with the international community and the risk of non-EU based TDM operators nonetheless being bound by the DSM Directive’s provisions for TDM activities (which may have extra-territorial application) when such activities are either carried out in Europe, on European source data or on works owned by EU-based copyright holders.

1.5 Outline
This paper begins with this introductory overview (1). Then, an analysis of the background to the research questions is conducted, including an understanding of Big Data and the environment in which TDM exists (2). Next, TDM and its relationship with copyright, how Big Data can be protected by copyright and how TDM processes can constitute copyright infringement of the exclusive rights of reproduction and communication to the public are discussed (3). Thereafter, we deal with the EU’s approach to copyright exceptions as applicable to TDM prior to the DSM Directive, in the context of international treaties, and later with examples from English and French law (4). A discussion of the DSM Directive’s provisions (from the proposal to the final text) related to TDM is then undertaken, with an emphasis on its positive and negative aspects (5). The position in other jurisdictions as an alternative to the EU approach, and the future of TDM and copyright law in Europe is then considered (6). We then conclude the question (7) and provide a bibliography of the resources used herein (8).
2. Background

2.1 A new era

We are in a new era – the fourth industrial revolution, an economy of internet, automation and AI. In the last 5 years, global internet traffic has grown 17-fold, resulting in the widespread use and development cloud computing, blockchain, Big Data and the IoT. To support this increase in widespread technology, the US spent nearly 500 billion USD on R&D and whilst modest in comparison, the EU nonetheless spent nearly EUR 300 million in 2015. As part of this new industrial revolution, we have moved away from the traditional analogue world to a connected environment and new techniques are being developed to further our information society. Such techniques require an adapted legal framework to protect the rights of the various actors, but which also allows for innovation to be exploited without risk for the benefits of society as a whole.

2.2 Europe’s Digital Single Market – the Aim to Create a Thriving Environment for Big Data and TDM

Against the backdrop of this fourth industrial revolution, it was inevitable that the EU committed to a political, policy-based and legal overhaul of existing frameworks to reap the benefits of innovation and technology. In 2014, the Commission iterated that the European digital economy had been “slow in embracing the data revolution compared to the US and lacks comparable industrial capability”. The Commission also noted research and innovation funding on data in the EU is “sub-critical” and the “corresponding activities are largely uncoordinated” whilst acknowledging “the complexity of the current legal environment [within the EU] together with the insufficient access to large datasets and enabling infrastructure create entry barriers to SMEs and stifle innovation”. It is within this context the EU committed to the Digital Agenda for Europe almost a decade ago.

---

6 Towards a thriving data-driven economy, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, COM (2014) 442 final, p. 2.
7 Ibid.
8 Ibid.
Against the backdrop of the 2008 global economic crisis, the Commission stated its intention to “deliver sustainable economic and social benefits from a digital single market” to create a Europe for the future. As recently as 2017, the Commission further reiterated:

“Digital innovation, driven by the combination of Big Data, cloud computing, mobile technologies and social media, is one of the most powerful drivers of change and the best opportunity for Europe to move back to a growth path.”

The Commission’s commitment to create a thriving environment for data and innovation is therefore key to ensure that Europe: (a) remains relevant in our technological society; and (b) maintains a firm place as key global player in ICT and the developer of new technologies. In order to achieve its goals, the Commission confirmed the steps to be taken to further the Digital Agenda initiative would include: (i) exploiting the single market to lay out joint technology roadmaps from research to commercialization for harnessing innovation to social need; (ii) by creating industry-led initiatives for open innovation in order to drive value creation and growth across the economy in areas such as the IoT and in key enabling technologies in ICT; and (iii) more generally, by revisiting the existing applicable legal frameworks, including copyright.

2.3 Big Data, TDM and Society

Before reviewing the copyright framework applicable to Big Data, we must understand the environment in which it applied. One of the key trends to arise out of the increased use of technology and the IoT, is the value attributed to Big Data and its analysis by automated TDM. “Organisations utilise Big Data and analytics solutions to navigate the convergence of their physical and digital worlds,” whether to enable strategy decisions, manage customer experience or to facilitate the delivery of digital services. It is anticipated that worldwide revenue forecasts for Big Data and business analytics solutions will reach USD 260 billion in 2022. Big Data is therefore big business.

9 A Digital Agenda for Europe, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, COM/2010/0245 f/2.
11 A Digital Agenda for Europe, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, COM/2010/0245 f/2 at para. 2.5.2.
12 Ibid at para. 2.5.3.
14 Ibid.
As the Commission confirmed, “leveraging data-driven innovation, enabled by Big Data technologies, is a powerful value generator and constitutes a real benefit for Europe’s economy as a whole”\(^\text{15}\). It is therefore unsurprising the EU wished to commit to this data-driven economy so as to ensure that Member States and the DSM are best-positioned to attain a large market-share of this highly relevant industry.

2.3.1 The Concept of “Big Data” - Volume, Velocity, Variety - and Value?

What is Big Data? “Big Data” could be determined by the “Volume” of information, text and data available – some experts have described the notion of Big Data as simply being “large pools of data”\(^\text{16}\) – literally, “big” data. However, as highlighted by the OECD, the problem with defining Big Data in terms of size, is that this method of quantification depends on the “\textit{evolving performance of available storage technologies}”\(^\text{17}\). “Volume” is therefore only one method for defining Big Data.

Indeed, “Volume, Velocity and Variety”, otherwise known as the three “Vs”, are often considered by industry experts as being the three main characteristics of Big Data\(^\text{18}\). When considering “Velocity”, we must look at how and when data is processed, including the accessibility of data and its availability in real-time. To illustrate, the OECD considers the “\textit{primary benefit of [Big Data] is its capacity to provide real-time statistics that are timelier than official statistics}”\(^\text{19}\), in other words, the speed in which information is available. When defining “Variety”, this is diversity of “\textit{unstructured data sets}”\(^\text{20}\). These three Vs therefore allow us to classify “Big Data”.

A fourth “V” also defines Big Data – the notion of “Value”, which is not only related to income and revenues gained from Big Data but also to the socio-economic value obtained from the use of Big Data\(^\text{21}\). The ability to search data has very important implications for EU citizens’ right

\(^{20}\) Ibid. p.12.
\(^{21}\) Ibid.
to information\textsuperscript{22}. For example, without Big Data and TDM, journalists would never have been able to reveal the “Panama Papers” scandal\textsuperscript{23}. Big Data and TDM may also be used to assist in other areas of society, such as in the field of criminology\textsuperscript{24}, in banking\textsuperscript{25}, in online marketing\textsuperscript{26}, in medical discoveries\textsuperscript{27} as well as in public health management\textsuperscript{28}.

2.3.2 Big Data Analytics, AI and Machine Learning

Aside from harnessing the value of Big Data, being able to freely use, process and manipulate Big Data through scanning, reproduction and analysis is critical to the development of AI and machine learning\textsuperscript{29}. Broadly speaking, AI is the ability of a computer to perform tasks commonly associated with humans. In particular, AI is predicated on the analysis of Big Data in its varying shapes, sizes and forms. Machine learning is a set of techniques that allows computers to ‘think’ by creating mathematical algorithms based on accumulated data. AI uses machine learning, and TDM relies heavily on both AI and machine learning to produce the most relevant results\textsuperscript{30}. The ICO considers that Big Data, AI and machine learning are inherently linked:

“Big Data can be thought of as an asset that is difficult to exploit. AI can be seen as a key to unlocking the value of Big Data; and machine learning is one of the technical mechanisms that underpins and facilitates AI. The combination of all three concepts can be called ‘Big Data analytics’.”\textsuperscript{31}

\textsuperscript{26} Content and Influence Marketing, Measure conversation volume and sentiment, find key influencers and identify topics that resonate with target customers, all in real time, Ubermetrics, available at: https://www.ubermetrics-technologies.com/influencers-content-marketing and accessed on 11 May 2019.
\textsuperscript{29} Big data and data protection (GDPR and DPA 2018), by Richard Kemp, Partner, Kemp IT Law and Practical Law Data Protection, Practical Law Company, Thomson Reuters, Resources ID w-017-1623, 2019.
\textsuperscript{30} Ibid.
There is an undeniable interrelation between innovation, and TDM, AI and Big Data: Big Data feeds AI, and AI algorithms (which can form the basis of TDM techniques) which all need a free and unfettered access to Big Data to be able to function in the most efficient way.

Maryam Mazraei, Founder at UK-based data analytics company provides insight into how TDM is essential to its business model analysing start-up failures, and also for start-up investors generally:

“We track and analyse company data to enhance decision making and provide insights to tech investors and the start-up ecosystem. It is essential for us to have access to information for our automated data pipeline which requires TDM to be able to scrape relevant information for better analysis and learning”32.

By way of further example, AI fuelled by Big Data is used by airline, SAS, to de-ice its aircrafts more efficiently to reduce passenger delays (a common issue during Scandinavian winters)33. Whilst seemingly banal as an example, the point here is that Big Data and the related opportunities that machine learning provides to data analytics tools, have a meaningful impact on society. It is therefore important that the legal framework covering TDM activities, as a method of analysing Big Data and machine learning, is properly adapted to allow society to benefit from this value.

Whilst the benefits are clearly apparent, Big Data alone is highly diverse (i.e., the three V’s) and requires the capability to search and link data sets from unstructured sources before we are able to gain anything useful in monetary or societal terms. Without specific methods to manage Big Data, the likelihood of harnessing its value is almost impossible, even if those methods may entail legal risks, such as widescale copyright infringement.

2.3.3 TDM as a Method of Managing Big Data

In the age of Big Data, the OECD points out that “information is highly context-dependent and may not be of value out of the right context”34. The value of data itself is therefore not to be found in the data or text in isolation, but in the aggregation and analysis of that data35.

Mechanisms must be created to be able to efficiently deal with the three V’s to ensure that its value is harnessed. Text and data must be extracted from its original source to enable the analysis of information-sets leading to the discovery of patterns and trends, and the creation of new data sets.

Historically, extracting value from unstructured data was “labour-intensive”36. The extraction, aggregation and analysis of Big Data is almost impossible to perform manually, however TDM techniques, which are developed through innovation, automation, AI and machine-learning are able to do so efficiently37 and cost-effectively. An example of this is the development of internet search engines, such as Yahoo!, which initially edited web directories using humans. As internet content increased, the only way to manage such massive unstructured data sets was to create automated scanning of the content via TDM mechanisms38. Whilst this adoption of automated TDM techniques by Yahoo! was successful, Google was already miles ahead in terms of market-share due to its much earlier adoption of TDM via the implementation of its PageRank algorithm. It is hard for manual TDM to compete with automated procedures39, although as processes become more automated, the governing legal framework (for example, relating to the copyright issues arising out of TDM) must adapt to new risks, and also to new opportunities.

Since TDM is clearly a key method for managing Big Data, even for a simple internet keyword search, it is vital, as shall be discussed below, that the legal framework surrounding access to Big Data and TDM activities (particularly insofar as it relates to copyright) is sufficiently flexible to boost innovation, whilst also protecting the interests of various stakeholders - the owners of the initial text and data sources, the creators of the ultimate aggregated data forms as well as the end-users of the TDM results.

2.4 Proposal for a Directive on Copyright in the Digital Single Market

EU-based researchers have indicated the “uncertainties concerning the treatment of TDM activities under European and national copyright laws”40 have traditionally hindered the

---

39 Ibid.
development of TDM techniques in Europe. Indeed, the Commission itself acknowledged that “fragmented implementation of copyright rules and lack of clarity over rights to use data obstruct the development of cross-border data use and new applications of technologies (e.g. text and data mining)”\(^{41}\). Clearly, this risk of legal uncertainty of copyright law, and its impact on investment and development within Europe is a potential barrier to the success of EU’s Digital Agenda\(^{42}\). A method of addressing such risk is by implementing new legislation, namely copyright legislation, which is adapted to the fourth industrial revolution and its digital and technical particularities, including for TDM. Following on from the Digital Agenda and DSM strategy\(^{43}\), the European Parliament and the Council therefore attempted to tackle the uncertainties in 2016 when it issued its Proposal\(^{44}\).

It is important to state that copyright law existing prior to the Proposal was not necessarily unsound, \textit{per se}, but was nonetheless outdated compared to the new realities of digital technologies\(^{45}\). The aim of the Proposal was therefore to ensure an appropriate copyright environment for the all actors and new business models of the digital age, harmonising national laws in respect of online copyright across the Member States to provide legal certainty and to avoid fragmentation of the internal market, whilst enhancing cross-border access to copyrighted content\(^{46}\).

In terms of updating copyright law in order to facilitate management of Big Data through TDM techniques, the Commission’s intention was to make legislative proposals to:

“[R]educe the differences between national copyright regimes [as between Member States] and allow for wider online access to works by users across the EU, including greater legal certainty for the cross-border use of content for specific purposes (e.g. research, education, text and data mining) through harmonised exceptions”\(^{47}\).

Therefore, exceptions on TDM were clearly at the forefront of the legislator’s mind when compiling what would become, the DSM Directive.

\(^{41}\) A Digital Single Market Strategy for Europe, A Digital Agenda for Europe, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, COM/2015/0192 final, 6 May 2015 at para 4.1.


\(^{43}\) Towards a modern, more European copyright framework, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, COM/2015/0626 final, 9 December 2015.


\(^{45}\) Ibid.

\(^{46}\) Ibid.

\(^{47}\) A Digital Single Market Strategy for Europe, A Digital Agenda for Europe, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, COM/2015/0192 final, 6 May 2015 at para 2.4.
Indeed, even in the text of the Proposal, the Commission confirmed that as new uses of copyrighted work have emerged, “it remains uncertain whether [existing] exceptions to [copyright infringement] are still adapted to achieve a fair balance between the rights and interests of authors and other rightsholders on the one hand, and of users on the other”\textsuperscript{48}. Furthermore, the Commission further stated that the objective of the new legislation is to “guarantee the legality of certain types of uses in these fields, including across borders”\textsuperscript{49}.

Despite these good intentions aimed at improving the copyright law framework for new technologies, and specifically for TDM, the Proposal attracted much industry criticism. Members of 23 organisations representing universities, technology companies, telecommunications providers, start-ups, libraries, scientific and research funding organisations, open access publishers, journalists and non-profits opined in an open letter to the Commission that:

“We fail to comprehend why, although the [European] Commission accepts the fact that AI and machine learning need vast amounts of data for training, it is pushing for a restrictive TDM exception within the [DSM Directive] that will remove the ability for start-ups, businesses, or public-private collaborators to use TDM to develop any of the innovations it seeks to foster”\textsuperscript{50}.

The Proposal was therefore subject to modification and various debates throughout its legislative history, due to the general controversies surrounding some of its provisions\textsuperscript{51}, with the final DSM Directive adopted by the European Parliament on 26 March 2019 and approval by the Council of the EU on 15 April 2019\textsuperscript{52}, which provides for a slightly wider TDM exception than initially envisaged in the Proposal. The DSM Directive entered into force on 07 June 2019 and Member States shall have two years to transpose its provisions into their respective national legal frameworks (i.e., by 07 June 2021).

\textsuperscript{49} Ibid.
\textsuperscript{51} For further information on the general criticisms, see EU Copyright Reform/Expansion, Julia Reda at: https://juliareda.eu/eu-copyright-reform/ and accessed on 27 April 2019 and Poland’s recently filed complaint - Poland files complaint with EU’s top court over copyright rule change, Agnieszka Barteczko, David Goodman, Reuters, 24 May 2019 available at: https://www.reuters.com/article/us-eu-copyright-poland/poland-files-complaint-with-eus-top-court-over-copyright-rule-change-idUSKCN1SU0T9 and accessed on 29 May 2019
However, it is within the context of a traditionally unfavourable environment for TDM in the EU, which is completed by the arguable lack of sufficiency of the TDM provisions contained in the DSM Directive, that this paper shall consider whether the EU approach to copyright in the field of TDM has been, and continues to be, a hindrance to European innovation.

3. TDM and its Relationship with Copyright

3.1 IPR Applicable to Big Data

The main IPR which are likely to exist in respect of Big Data are copyright and *sui generis* database right. Patents may apply to software and business processes which *process* Big Data, but do not protect the text and data *content* of Big Data. Trade marks can apply to products related to Big Data (such as AI tools and search engines), but again, it is very unlikely that trade marks would apply to Big Data itself. Furthermore, in terms of quasi-intellectual property rights, Big Data may be protected by confidentiality and contractual provisions or as a trade secret. For the purposes of this paper, only copyright, including copyright which may arise in a database (but exclusive of *sui generis* database right) shall be discussed.

3.1.1 Sources of Copyright Law within the European Union

In the EU, the sources of copyright law comprise international treaties and EU legislation, and national laws of Member States. Despite the efforts to converge Member State laws, copyright law across the EU remains, essentially, national law which has resulted in uncertainties for copyright holder rights across the board (and not just for TDM) since the regimes as between EU Member States are not fully harmonised. The uncertainties are heightened in our cross-border, internet-age, and are further exacerbated by the fact that copyright arises automatically upon creation of the copyrighted work and does not require any

---

53 Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure.
formal registration procedure. This means that keeping track of copyrighted work, attributing ownership and identifying infringers is not always a simple task, particularly in our online environment. As discussed further below, the uncertainties of European copyright law prior to the DSM Directive, and presently, as applicable to TDM, are susceptible to constitute a hindrance to innovation in the region.

3.1.2 What is Copyright?

WIPO considers that copyright is “a legal term used to describe the rights that creators have over their literary and artistic works”57. According to this definition, the scope of copyright therefore seems rather wide, but limited to the creative fields, and so arguably, the nature of the text and data that is used for TDM would be excluded from copyright protection. It is, however, inaccurate to consider that copyright only applies to artistic works, or literary works in the traditional sense.

Generally speaking, “copyright protection extends only to expressions, and not to ideas, procedures, methods of operation or mathematical concepts as such”58. Yet, as WIPO points out, applicable legislation worldwide does not usually contain an exhaustive list of what may constitute a copyrightable work, but examples of works which are often protected by copyright include extend further than the traditional artistic notions to include: literary works, computer programs, databases, films, music, artistic works such as paintings, drawings, photographs, and sculpture, architecture; advertisements, maps, and technical drawings59 and websites, meaning the source data of TDM is capable of copyright protection, as further discussed below.

3.1.3 How Can Copyright Protection Extend to Big Data?

Academics argue that one of the “basic and fundamental principles of copyright law is that data is as such not protected, as copyright only protects the creative form, not the information incorporated in the protected work”60.

58 Ibid.
59 Ibid.
Indeed, Geiger et. al consider that TDM should not be concerned by any IPRs, whether copyright or otherwise, as TDM activities fall outside the scope of any monopoly IPR, and as such “any restriction would amount to undermining the underlying rationales of copyright protection and result in an inadmissible restriction of freedom of expression and information”\textsuperscript{61}. Arguably, the risk of copyright infringement in respect of TDM processes carried out on data, is a non-issue: data in itself is simply not capable of copyright protection.

Geiger’s arguments are understandable, but in the context of Big Data, and given the three Vs, mere “data” must be distinguished from “Big Data”. As such, it is nonetheless likely that literary copyright subsists in documents, publications, research and analysis, as well as in any technical documents, software and IT architecture which constitute Big Data, and which is ultimately subject to TDM activities. Copyright within a database (so-called “database copyright”) may apply to Big Data in some instances.

As an aside, database copyright must be distinguished from software and literary copyright and from the \textit{sui generis} database right as introduced by Directive 96/9/EC, which is heavily investment based and for which “the intellectual effort and skill of creating that data are not relevant in order to assess the eligibility for database protection”\textsuperscript{62}. Whilst detailed discussion of \textit{sui generis} database right is outside the scope of this paper, database copyright may be available where a database is not covered by the \textit{sui generis} right (for example, due to a lack of expenditure or investment, of if the verification or presentation of the data is trivial).

This is confirmed by various provisions of EU legislation which state that works such as computer programs, or databases are protected by copyright only if they are original, i.e., they derive from the author’s own intellectual creation\textsuperscript{63}. To illustrate, these provisions were transposed into UK law under section 3(1)(a) of the CDPA which states that: “literary work means […] a table or compilation other than a database”.

This provision, as debated by the CJEU and the UK CofA in the Football Dataco saga, further confirms the EU’s position - literary work \textit{copyright protection} may exist in a database, provided that: (i) the selection or arrangement of the database’s contents are the result of the

\textsuperscript{61} Ibid.
\textsuperscript{62} Judgment in Case C-604/10 (Football Dataco Ltd & ors v. Yahoo! UK ltd & ors).
\textsuperscript{63} See Articles 1(3) Articles 1(3) of Directive 91/250, 3(1) of Directive 96/9 and 6 of Directive 2006/116.
author’s “own intellectual creation”; and (ii) there is “originality in the selection or arrangement of the data which that database contains”\(^64\). For the avoidance of doubt, when referring to copyright within this paper, including insofar as it may protect Big Data, this includes references to copyright which therefore may exist in a database, as confirmed in the aforementioned Football Dataco case law.

It is therefore incorrect to state that Big Data is not capable of copyright protection, even if the extent to which copyright protection may protect source text and data which is subject to TDM activities is a complex question which has been raised before the CJEU, as detailed below. Furthermore, the EU clearly saw the likelihood of copyright infringement in respect of Big Data and TDM is serious enough that it provided for it within the Proposal and ultimately, the DSM Directive. In our opinion, as long as Big Data is capable of copyright protection, legal uncertainty for TDM performed on extracts of Big Data arises, and arguably, EU law has not fully addressed the realities of the issues, which results in a risk for TDM carried out in the EU or on an EU-based source data.

3.1.4 The Copyright Monopoly

Provided the source text and data which is subject to TDM is protected by copyright, such protection grants powerful rights to the holder of the copyright(s) covering the source text and data. The copyright monopoly is comprised two categories of rights afforded to copyright holders: (i) economic rights, which allow for financial reward for the use of their works by third parties; and (ii) moral rights which protect non-economic rights (the right to be recognized as an author of the work or to object to any modification or use of a work which could denigrate the author’s reputation). Generally, moral rights tend to be more present in civil law regimes, whereas economic rights extend to both common and civil law jurisdictions. These economic rights therefore confer on copyright holders a monopoly on the exploitation of their works. For the purposes of this paper, only the copyright holder’s economic rights shall be discussed in detail.

In terms of economic rights, most copyright laws allow copyright holders to permit or prevent reproduction, performance, recording, broadcasting, translation and adaptation of their

---

64 Judgment in Case C-604/10 (Football Dataco Ltd & ors v. Yahoo! UK ltd & ors) and Football Dataco Ltd v Brittens Pools Ltd [2010] EWCA Civ 1380; Football Dataco Ltd v Brittens Pools Ltd [2010] EWHC 841 (Ch); Dataco Ltd & Ors v. Yahoo! UK Ltd [2012] EWCA Civ 1696.
copyrighted works, or indeed to receive remuneration for use of their works. Of course, it then results that such acts related to a copyrighted work which are prohibited or not authorised by the copyright holder, constitute an infringement of that holder’s copyright, unless there are any legal exceptions to copyright infringement or limitations to the copyright monopoly which mean that such authorization is not required. These exceptions and limitations insofar as they relate to TDM, and as are required to foster an innovative environment, shall be discussed below.

3.2 TDM Techniques- Tensions with Copyright

TDM is performed on “large amounts of text data, which are created in a variety of social network, web, and other information-centric applications”. According to Aggarwal et. al, “unstructured data is the easiest form of data which can be created in any application scenario” resulting in a “tremendous need to design methods and algorithms which can effectively process a wide variety of text applications”. Whilst TDM can take many forms and it “virtually impossible to provide a general and exhaustive illustration of how TDM works”, it is acknowledged there are three steps common to most TDM techniques: Firstly, access to content (Step 1); secondly, extraction and/or copying of content (Step 2); and finally, mining and knowledge discovery (Step 3). These techniques are described graphically at Appendix 1.

3.2.1 TDM - Acts of Copyright Infringement?

It is clear that “during the chain of activities enabling TDM research, some IPR relevant actions are technically necessary so that, in the absence of “specific permission within the legal framework, TDM can lead to an infringement”. The use of TDM techniques therefore might result in copyright infringement depending on the “use of the existing sources, technical tools and the extent of the mining process”.

67 Ibid.
68 Ibid, Chapters 3 to 10.
70 Ibid.
72 Ibid.
The exclusive rights laid down at Article 2(a) of Infosoc, which requires authorization from the right-holders for the direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part, of their works\(^{73}\), and Article 3(1) of Infosoc which requires Member States to provide authors of works with the exclusive right to authorise or prohibit any communication of their works to the public are particularly at risk, subject to the applicability of any exceptions or limitations.

The key exclusive rights available to copyright holders discussed in this paper which may be infringed by TDM processes are therefore limited to the so-called right to control: (i) reproduction of copyrighted materials; and (ii) communication of those copyrighted materials to the public. A stronger focus in this paper is on the issue of reproduction, which is highly likely to arise through TDM, whereas as the communication right, as noted below, is not necessarily infringed as a direct result of TDM activities themselves, but rather as a result of a human decision to communicate to the public TDM source data, or output gleaned from TDM which may then contain copyrighted works, as a method of verifying TDM results and processing.

As Dr. Rosati opines, copyright issues can arise throughout the aforementioned three-step TDM processes. Step 1, access to content implies there is free access to the text and data on which the TDM would be carried out\(^{74}\). Unauthorised access to content in itself may result in the infringement of the copyright holder’s economic rights, and even authorised access to content does not necessarily authorise TDM activities to be carried out on that content. Step 3 – the very acts of TDM discovery may also result in copyright infringement to the extent that such text is processed and analysed to ensure knowledge discovery\(^{75}\). In this sense, copyrighted works are reproduced, sometimes translated, adapted and rearranged, which may, subject to applicable limitations and exceptions, “\textit{infringe upon the right of reproduction [at Article 2(a) of Infosoc] depending on the mining software deployed and the character of the extraction}”\(^{76}\).

Part 2, the phase of extraction and copying, is likely to be to the highest risk phase of TDM as it often involves activities encroaching on the exclusive rights provided at Article 2(a) of

\(^{73}\) Ibid.


\(^{75}\) Ibid.

InfoSoc. Furthermore, TDM activities might also involve reproduction of the original selection and arrangement of the content of a database, which may also result in infringement of database copyright (in addition to *sui generis* database rights), as well as infringement of literary or software copyright. Specific techniques of TDM will include copying (for example through non-crawling activities), reproduction (extraction), translation, adaptation, arrangement, and any other alteration to Big Data. Such techniques which rely on “reproductions resulting in the creation of a copy of a protected work along the chain of TDM activities might trigger copyright infringement”77.

It is generally considered that “*TDM output should not infringe any exclusive rights as it merely reports on the results of the TDM quantitative analysis, typically not including parts or extracts of the mined materials*”78. Nonetheless, we consider that the exclusive right of communication to the public as set down at Article 3 of InfoSoc regarding text and data initially subject to mining activities may be violated by TDM techniques, if such text and data are stored as source materials and shared or communicated within the industry in order to allow verification of the exactness of the TDM output79, albeit, as noted above, this infringement of the right to control communication to the public would not generally occur as a natural or automatic consequence of the TDM activities themselves.

3.2.2 When is the Threshold for Copyright Infringement Met?

If the Big Data on which TDM is performed is free from copyright protection and not otherwise protected (by other IPRs, contract, TPMs), then there is no risk of infringement. Whether this is the case depends on the nature of the Big Data that is processed through TDM. In our opinion, most data on which TDM is performed is likely to be protected by some form of IPR, given the scope of “Big Data”. Even if TDM source data is protected by copyright, the threshold for copyright infringement may not be met if the TDM extraction techniques only reproduce parts of the work so minimal so as to fall below the threshold of copyright infringement80. This issue was dealt with by the CJEU in the Infopaq case81 which raised the “*sensitive issue of the balance*

77 Ibid.
79 Ibid.
80 Ibid, p. 6.
81 Case C-5/08 (Infopaq International A/S v. Danske Dagblades Forening)
between the protection of copyright and technological development in the information society”.

Infopaq involved a request for a preliminary ruling before the CJEU from the Danish courts in proceedings between Infopaq International A/S, a Danish media monitoring and analysis company and the professional association of Danish newspaper publishers, concerning Infopaq’s data capture process. This data capture process involved technological processing, scanning, OCR, reproduction, storage and printing of text extracts from Danish newspapers. Infopaq acted on “key words” provided by its customers to create news summaries which were emailed to customers. Infopaq’s process has strong similarities to some TDM techniques. The case considered whether the text extracts created by Infopaq were sufficient to invoke the exclusive right preventing reproduction within the meaning of Article 2 of InfoSoc, and if so, whether those automated actions themselves constitute an act of reproduction requiring consent of the right-holder (i.e., the newspaper publishers or journalists), or whether such actions fulfil the conditions of Article 5(1) of InfoSoc and are therefore permitted by EU law. For the purposes of this paper, issues relating to freedom of information and the press which arise in the Infopaq case shall not be discussed.

3.2.2.1 TDM Carried Out on Single Words

In the Infopaq judgment, the CJEU confirmed that its “main objective is to introduce a high level of protection, in particular for authors […] including at the time of reproduction of those works” and the “acts covered by the right of reproduction [should] be construed broadly”. Accordingly, the CJEU held that:

“Regarding the elements of such works covered by the protection, […] they consist of words which, considered in isolation, are not as such an intellectual creation of the author who employs them. […] Words as such do not, therefore, constitute elements covered by the protection.”

It therefore appears the threshold for copyright infringement is not met through the use of TDM techniques, even on text and data protected by copyright, when it is carried out on single words.

---

82 Opinion in Case C-5/08, para 1.
83 Ibid, para 2.
84 Judgment in Case C-5/08, para 40.
85 Ibid, para 41.
86 Ibid, paras 45 and 46.
– indeed, according to the CJEU, single words are not covered by the protection provided for in Article 2(a) of InfoSoc.

3.2.2.2 **TDM Carried Out on Strings of Words or Phrases**

In the Infopaq case, “the production of extracts from newspaper articles which are composed of search words and the five preceding and five subsequent words”\(^{87}\), totalling a minimum of 11 words in any given search. In this instance, the CJEU considered that:

> “Certain isolated sentences, or even certain parts of sentences in the text in question, may be suitable for conveying to the reader the originality of a publication […] by communicating to that reader an element which is, in itself, the expression of the intellectual creation of the author of that article. Such sentences or parts of sentences are, therefore, liable to come within the scope of the protection provided for in Article 2(a) of [InfoSoc]. […] The reproduction of an extract of a protected work […] which comprises 11 consecutive words thereof, is such as to constitute reproduction in part within the meaning of Article 2 of [InfoSoc], if that extract contains an element of the work which, as such, expresses the author’s own intellectual creation.”\(^ {88}\).

Therefore, even extracts comprising at least 11 words are capable of being protected by copyright, and thus require the consent of the copyright owner before they can be lawfully exploited (i.e., reproduced through TDM processing). This is more so the case when “reproduction of multiple extracts of protected works”\(^ {89}\) are carried out, or indeed when the “cumulative effect of those extracts may lead to the reconstitution of lengthy fragments which are liable to reflect the originality of the work in question, with the result that they contain a number of elements which are such as to express the intellectual creation of the author of that work.”\(^ {90}\).

In light of this decision, we consider that it is even likely that extracts of less than 11 words, but more than 1 word, may fall within the protection provided for by Article 2(a) of InfoSoc if the scope of the extract nonetheless contains enough content to express the author’s own intellectual creation. Therefore, in the event that this threshold is met, then conducting

---

87 Opinion in Case C-5/08, para 3.
88 Judgment in Case C-5/08, paras 47 and 48.
89 Ibid, para 49.
90 Ibid, para 50.
reproduction activities by TDM processes on such extracts of Big Data would result in copyright infringement of the protected portions of work, thus leaving the infringer liable for damages or other enforcement action. We consider that in providing such a low threshold for finding that such small or incidental extracts of works are capable of copyright protection, the CJEU has contributed to the long-standing uncertainty in the EU’s approach to TDM activities.

3.2.2.3 Workarounds

From the drafting of the applicable legislation and the approach of the CJEU in Infopaq, the EU holds copyright holders’ rights at the forefront, in order to provide legal certainty as to the strength of their position and the possibility of remuneration. The balance is clearly tipped in favour of the rightsholders, even on such small portions of text, meaning the legal framework creates a less-favourable environment for TDM in the absence of a robust exception. Maybe such fears are a non-issue, as Infopaq does provide TDM operators with guidance which would allow them to create workarounds to avoid copyright infringement – meaning that TDM activities and copyright holder rights can co-exist in harmony without any need for a specific exception or consent.

Whilst potentially stating the obvious, the preamble to the DSM Directive confirms that TDM “may also be carried out in relation to mere facts or data which are not protected by copyright and in such instances no authorisation would be required”⁹¹, however we consider that such non-copyrighted facts and data are not in issue. The crux of the problem relates to TDM activities carried out on work (including extracts) which is copyrighted, or on Big Data, which due to its “volume, velocity and variety” may contain copyrighted works. This inevitably leads to uncertainty for TDM operators who are faced with the unenviable task of ascertaining which aspects of Big Data are risk materials. This is where workarounds could come into play.

For example, TDM operators could develop technology with copyright in mind – such as ensuring their algorithms simply take data on a one-by-one or single-word basis (to avoid the content that is mined for reaching the de minimis threshold for constituting a copyrighted works). TDM could also be developed so that only scanning activities are carried out, meaning the right of reproduction is not violated.

---

⁹¹ See Recital 9 of the DSM Directive.
Indeed, “in such a case where the software only “swallows” one or two words or pixels or data or sounds at a time and then goes on to the next ones without keeping a copy of them […] then […] no copying relevant in terms of copyright takes place and that such activity does not require the consent of the rightsholder”92. Furthermore, any source data could be automatically deleted from the TDM output so as to preserve the copyright holders’ exclusive communication right. This is discussed further below in respect of exceptions applicable to TDM activities under the law applicable prior to the DSM Directive.

However, in reality, depending on the TDM mechanism used93, the most valuable insights gained from TDM involve processing source data which comprises more than 11 words and in order to do so, TDM usually requires copying of the work to be analysed94. Big Data analytics uses complex algorithms for processing data which usually involves a “discovery” phase to find relevant correlations (which can be a form of machine learning) so that algorithms can be created, which rely on vast amounts of data being processed through massive networks. Furthermore, there is a tendency to collect everything, rather than limiting analytics to random samples or statistically representative samples. This provides overall better results for TDM and AI95. As TDM is most efficient without human intervention, it would be necessary to rely on a high level of sophisticated machine learning to ascertain if within the source data extracts, the text and data that is mined, is copyrighted material in the first place, or if the extracts that are taken still express the author’s own intellectual creation in the event the source material is indeed, copyrighted.

The workarounds are therefore not realistic and would result in TDM activities which are too restrictive (thus impacting the quality of the results, whether used as output to obtain new insights for society or business or for machine learning) or which would have unverifiable output (if the source texts could not be communicated). In order to further innovation, it is arguably more important that developers focus time and investment in ensuring they are able to create the most efficient and performant TDM technology, rather than dedicate this energy to implementing copyright workarounds. Of course, a robust copyright system is necessary to

93 I.e., information extraction from text, text summarization or text clustering.
95 Big data and data protection (GDPR and DPA 2018), by Richard Kemp, Partner, Kemp IT Law and Practical Law Data Protection, Practical Law Company, Thomson Reuters, Resources ID w-017-1623, 2019.
incentivise creators to provide original content, but given the nature of TDM, the value is in the collection of several sources of work gathered from Big Data – there is no single victim of copyright infringement because there is little value in one single piece of work or one sole extract. Arguably, the value gained from TDM arises from the analysis of several works or extracts, when taken together as a collective.

Given the broad scope of work and the number of potential copyright holders, seeking consent on an individual basis is costly (if consent via licensing solutions is to be relied upon) and unworkable (for example, we would never envisage seeking consent from individual site owners before using an internet search engine). Furthermore, whilst machine learning and AI is becoming more sophisticated, it is a high burden to impose on developers of algorithms the requirement for automated tools to ascertain whether a particular work is indeed copyright protected by verifying whether any given extract of source data represents an expression of the author’s own intellectual creation. The concept of whether a work is indeed the intellectual creation of its author seems to require human intervention, since it is such a subjective issue that even the courts grapple with. Given the sheer scale of TDM and Big Data, human intervention is unworkable.

As an aside, the results of TDM are also likely to constitute works which may be protected by copyright or other IPR in their own right, and EU law to date has failed to provide a specific protection for TDM output. As such output is usually created by machines, these works are arguably not capable of protection under EU copyright law, which requires that copyrighted works must be created, and not just recorded, by a human being, in order to be deemed “original” and thus capable of copyright protection96. Member States have taken different approaches to this requirement.

UK law, which has historically had a lower originality threshold, acknowledges that works may be created when there is no human author97. Section 9(3) of the CDPA expressly provides that the author of a computer-generated work (resulting, for example, from AI or TDM output) is deemed to be the person “by whom the arrangements necessary for the creation of the work are undertaken”.

---

97 For example, Section 178 of the CDPA.
In the UK, unlike in some other Member States, the person or entity ordering the work or developing the relevant technology would be the first copyright owner of any output. Again, this highlights the uncertainty surrounding protection of TDM output by copyright across the EU, which is a further hindrance to innovation and investment in the region, particularly if from one Member State to another, TDM operators are not sure they can benefit from a monopoly to exploit the results.

Returning to the key issue, as we have seen, because IPRs can be affected whenever mining involves IP-protected subject matter\(^98\), in the absence of a specific exception for TDM, the risk of IPR infringement within this area means that investment in and development of TDM is avoided as firms wish to avoid the risk of liability.

In Europe, we therefore have a vital need for robust exceptions in respect of TDM which are clear and certain to encourage innovation to further the Commission’s Digital Agenda and remain competitive on a global scale. This is arguably where EU law to date, including the DSM Directive, has failed.

4. Pre-DSM Directive Approach to TDM Exceptions and Limitations across the EU

4.1 General Limitations and Exceptions to Copyright Infringement: International Focus

TDM requires a sufficient limitation to a copyright holder’s exclusive rights or a robust exception to avoid risk of such processes amounting to copyright infringement. A limitation or exception in copyright law refers to a mechanism which balances legitimate user rights with those of copyright holders and limits or overrides the exclusive rights granted by copyright enabling the world at large to exploit copyrighted works without consent of the copyright holder and without fear of infringement claims within the limits of the permitted limitations or exceptions.

Insofar as this relates to TDM, we could envisage exceptions which apply freely to the technical processes conducted on copyrighted works which would not require prior authorisation of the copyright holder – TDM activities could therefore become an exception to the exclusive rights conferred on a copyright owner, or indeed a limitation to such rights whereby the extent of the

right to control reproduction, for example, does not extend to reproduction carried through TDM. Whilst not specifically related to TDM, exceptions in existence prior to the DSM Directive are laid down in international treaties, at EU-level and in national laws, including those of Member States.

As signatories to international treaties, it is therefore important to consider the international position insofar as it applies to Member States. Therefore, in respect of exclusive reproduction rights at an international level, Article 9(2) of the Berne Convention provides that:

“It shall be a matter for legislation in the [signatory] countries […] to permit the reproduction of such works in certain special cases, provided that such reproduction does not conflict with a normal exploitation of the work and does not unreasonably prejudice the legitimate interests of the author”.

This provision, whilst creating an optional obligation on signatory states, confers a limitation on the scope of protection available to a right holder, to the extent that no protection or limited protection would be available for the particular activity falling within the scope of a “certain special case”. According to Cabrera Blázquez F.J et. al, the scope of such limitations is:

“[M]ainly based on the assumption that there are clear public policy grounds that copyright protection should not exist in the works in question (for example, in the interest of the public)”, whereas exceptions to protection for “certain permitted uses allow the use of protected work without the authorization of the owner”99.

International law therefore provides for the possibility to extend permitted limitations and exceptions to TDM activities, provided that it can be argued the overall output of TDM has value in terms of public interest and it is therefore overriding the private rights of authors in their works in these particular circumstances is justified100. Further provisions can be found in the TRIPS Agreement at Article 13 which provides for a general exception to the exclusive rights of copyright holders, broadly follows the aforementioned provision of the Berne Convention, however, its scope is arguably broader as it applies to all economic rights and not

100 Ibid.
only to reproduction rights. Perhaps the most important international treaty applicable to exceptions and limitations which may authorise TDM activities is the WCT, which was specifically drafted with the internet environment in mind: indeed, the WCT requires signatory states to “provide a framework of basic rights to creators […] when their works are disseminated through new technologies in the digital environment”.

Article 10 of the WCT provides an optional provision for contracting states which broadly follows the provisions of the Berne Convention and the TRIPS Agreement in that:

“[Such] limitations of or exceptions to the rights granted to authors of literary and artistic works may be provided in certain special cases that do not conflict with a normal exploitation of the work and do not unreasonably prejudice the legitimate interests of the author”.

The preamble to the WCT further states there is a “need to maintain a balance between the rights of authors and the larger public interest, particularly education, research and access to information”. The aspect of research and access to information appears to have been one of the bases on which the EU applied the TDM exception as drafted in the Proposal and ultimately as enshrined in the DSM Directive. As we argue further below, the DSM Directive’s provisions are weighted in favour of research, which is, in our view, unnecessarily restrictive and hinders commercial TDM.

To the extent that international law provisions have bearing on EU law, the provisions of the WCT are perhaps the most pertinent to our new industrial era as it expressly provides that signatory countries are able to: (i) carry forward and appropriately extend into the digital environment limitations and exceptions in their national laws; and (ii) devise new exceptions and limitations that are appropriate in the digital network environment.

Accordingly, international law, via the WCT, expressly permits national legislatures to adapt their copyright law frameworks to extend to the advent of new technologies and processes, such

---

103 See Article 10, WCT.
104 Agreed Statement concerning Article 10, WCT.
as TDM, which can go further than the limits of research provided the spirit of the Berne Convention is maintained – i.e., to ensure that such exceptions do not unduly prejudice rights holders or interfere with the normal exploitation of copyrighted works, which as submitted above, TDM activities do not, especially in light of the social value TDM output provides in our digital society (as discussed below).

In short, the restricted approach to TDM therefore seems to be a European issue, rather than an international one, since the aforementioned framework appears to permit the EU to commit to the widest possible playing field for TDM innovation.

4.2 Limitations and Exceptions Applicable to TDM: EU Focus

As well as following the framework provided for in international treaties, “the main objectives of the exceptions and limitations provided for in EU law are to achieve public policy goals, such as fundamental freedoms, education or research, or to facilitate the use of protected content in specific circumstances”\(^1\)

As far as EU legislation is concerned, exceptions and limitations to copyright-holders’ exclusive rights are provided for in various EU legislative provisions\(^2\). The majority of the exceptions provided for in the acquis communautaire are not mandatory and are entirely optional for Member States to implement. This has unfortunately resulted in a patchwork approach to exceptions and limitations across the EU, including insofar as they may relate to TDM, creating uncertainty for all stakeholders. The EU’s failure to commit to full and compulsory harmonization of copyright laws across all of its Member States is evidence that its approach to copyright in our digital society and internet age hinders innovation in the region, because innovators have, and are still, faced with uncertainty in the legal regime from one Member State to another, which means they are not able to operate on a level-playing field and benefit from legal certainty.

According to Cabrera Blázquez F.J et. al, InfoSoc “is the most horizontal of these directives as it aims at harmonising the exceptions to copyright and related rights and applies to all types of

---

106 Article 5, InfoSoc Directive; Articles 6 and 5, Software Directive; Articles 6 and 9, Database Directive; Articles 6 and 10, Rental Right and Lending Directive; Article 6, Orphan Works Directive.

Whilst exceptions and limitations applicable to TDM activities may arise under the Database Directive, namely by applying the exception related to the normal use of a database\footnote{Article 9, Database Directive.}, or by ensuring that TDM only relates to the extraction of an insubstantial part of a database\footnote{Article 6(1), Database Directive.}, for the purposes of this paper we shall focus on the InfoSoc provisions insofar as they apply to TDM activities carried out on copyrighted source data outside of the scope of personal use and public security\footnote{As an aside, the exception provided for at article 5(2)(b), InfoSoc may apply to TDM activities carried out on copyrighted works for personal use, meaning that in such cases, the permission of the copyright holder would not be required; the exception set out at article 5(3)(e), InfoSoc may apply to TDM activities carried out on copyrighted works in the interests of public security (for example, by Europol or in the interests of intelligence and counter-terrorism).}. The insistence is that prior to the DSM Directive, TDM had no firm legal foundation and relevant stakeholders had to rely on provisions which are ignorant of the realities of TDM processes. Of course, when faced with such conditions, the environment is not the most favourable in which to innovate or invest.

4.2.1 Temporary Acts of Reproduction

Copyright infringement via reproduction may be exempted from the exclusive right in Article 2 of InfoSoc (i.e., the consent of the right holder is not required) if it meets the cumulative conditions set out at Article 5(1)(a) and (b) of InfoSoc: (i) the act is temporary; (ii) it is transient or incidental; (iii) it is an integral and essential part of a technological process; (iv) the sole purpose of that process is to enable a transmission in a network between third parties by an intermediary of a lawful use of a work or protected subject-matter; and (v) the act has no independent economic significance. Article 5(1) has the potential to be powerful, since it is drafted as one of the rare mandatory exceptions – that is to say that it must be transposed into the national laws of the Member States.

The scope of Article 5(1) was discussed by the CJEU in the Infopaq case\footnote{Case C-5/08 (Infopaq International A/S v. Danske Dagblades Forening)}. The reasoning can be extrapolated to TDM. On the application of Article 5(1) insofar as it pertains to the Infopaq
case, the CJEU considered that authors need legal certainty with regard to the protection of their works\textsuperscript{112}. The CJEU therefore opined that in order for an act of reproduction to be temporary and transient, it must “not exceed what is necessary for the proper completion of a technological process”\textsuperscript{113}.

The CJEU therefore held the condition would not be satisfied unless the works subject to the reproduction are “deleted automatically, without human intervention, once its function of enabling the completion of [a technological] process has come to an end”\textsuperscript{114}. In Infopaq, this was not the case, as findings were printed out and/or emailed to end-customers – deletion therefore required human-intervention, at human will and convenience. As such, in Infopaq, the aforementioned cumulative conditions were not satisfied.

From a general TDM processing perspective, the cumulative conditions set out in Article 5(1)(a) of InfoSoc would rarely, if ever, be satisfied in a commercial TDM environment. Of course, it is easy to ascertain that reproduction is an integral part of the technological process of automated TDM procedures. Through appropriate development, technological measures could be implemented to TDM practices to ensure that automated, non-human deletion of data at the end of a technical process occurs: this would ensure compliance with the directions of the CJEU in its Infopaq ruling. However, Big Data and the results gleaned from it via TDM processing techniques is big business – for the majority of valuable development of TDM AI and processes, carrying out TDM itself and then sharing the output (which may include reproduced extracts of original works) are all steps which are monetized and/or lead to financial reward – the reproductions therefore do not fulfil the requirement of “no independent economic significance” for the purposes of Article 5(1)(a) of InfoSoc.

Additionally, the requirement regarding “transmission in a network between third parties by an intermediary” is also problematic for TDM. Whilst it could be possible to define a specific network for sharing TDM output or results, this is extremely restrictive and does not lend itself to the realities of TDM, the processes that are implemented nor how the output is aggregated. The exception at Article 5(1)(a) is more appropriate to broadcasting and telecommunications whereby works are temporarily transmitted through networks rather than to TDM activities.

\textsuperscript{112} Judgment in Case C-5/08, para. 59.
\textsuperscript{113} Ibid, para 61.
\textsuperscript{114} Ibid, para 62.
Furthermore, workarounds requiring immediate deletion, or compliance with the Infopaq judgment, does not seem feasible as “AI and machine learning need vast amounts of data for training”, which often needs to be saved or retained in some form in order to achieve the best results\textsuperscript{115}. Indeed, deleting data from machines periodically has the same effect as periodically erasing a human’s memory. If this were the case, humans would never be able to develop beyond a certain point, if we constantly had to “re-learn” things which had been erased. Memory built on acquired information and data in Machine Learning and AI development, which is used as building blocks for new machine-skills, is no different – if it is arbitrarily deleted, developers would need to start from scratch.

Finally, it is worth noting that the scope of the exceptions contained in Article 5(1)(a) would only apply to acts of reproduction and so sharing of the source data within the TDM results would still breach the exclusive right to control communication of a copyrighted work as set out at Article 3 of InfoSoc, even in the event the provisions of Article 5(1)(a) were to apply.

4.2.2 Teaching and Scientific Research

Prior to the DSM Directive, EU law contained exceptions related to teaching or scientific research in the guise of Article 5(3)(a) of InfoSoc, which, whilst not expressly stated, may apply to TDM activities conducted on copyrighted materials for which no consent or prior authorization of the copyright holder would be needed. The scope of Article 5(3)(a) would therefore potentially prevent TDM operators from infringing a copyright holder’s right of reproduction and right of communication of copyrighted works, provided that such activities are carried out:

“[F]or the sole purpose of illustration for teaching or scientific research and 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”\textsuperscript{116}.

The pre-existing exception provided for at Article 5(3)(a) is potentially sufficient for TDM activities, applying not only to the reproduction, but also to the communication of such copyrighted works within teaching and scientific research activities. When dealing with


\textsuperscript{116} Article 5(3)(a) of InfoSoc.
traditional teaching and research methods, the limitations to the exception have reasonable justification. In academia, it is important to be able to freely cite and share works, which is even more prevalent as international and disciplinary research is conducted. Whilst an in-depth discussion is outside the scope of this paper, the extent of this exception seeks to preserve the so-called moral rights, which is an important part of academic work – the right to be recognized as an author and to receive acclaim for the contribution to the given academic field. The exception also seems fairly unproblematic for TDM purposes even if it is not specifically provided for, since reproduction and communication is entirely possible, even in instances where it is not feasible to attribute the source material to an author, where, for example, such data has become so aggregated in the TDM output that it becomes impossible to allocate it to a specific individual.

However, it is noteworthy that the pre-existing exception, results in a refusal from the EU legislator to extend the exception for teaching and scientific research to reproduction or communication for commercial purposes, which seems wholly misaligned to the modern realities of academia whereby universities and researchers strive for funding and budget to carry out the most cutting-edge projects, in order to be leaders in their research fields. Furthermore, the scope of the exception is drafted in optional wording, meaning there is no obligation on Member States to transpose the provision into their national laws.

This therefore means the harmonization of the exception is limited across the EU, which provides for significant uncertainty particularly in the online environment with no borders and during intra-EU research collaborations where rights holders may be given significantly different protection depending on the law of the applicable jurisdiction.

4.2.3 Incidental Inclusion

Article 5(3)(i) of InfoSoc provides that a copyright holder's exclusive communication and reproduction rights are not infringed where dealings in the copyrighted work are merely due to the incidental inclusion of that work in other material. From a TDM perspective, this could apply to the reproduction of extracts of copyrighted material carried out during the TDM process where such reproduction is merely incidental to the overall technique applied to the copyrighted material, and ultimately also to the inclusion of source material in the TDM output.
which is created at the end of the process, or which is provided with the output simply to allow end-users to verify the veracity of the results.

This then begs the question: could Article 5(3)(i) be the solution rendering the performance of TDM techniques in Europe certain and viable? In response, it is worth noting that the exception of “incidental inclusion” is not drafted as a mandatory provision, and the scope and implementation of the provision from one Member State to another has the potential to vary greatly. IP writers argue incidental inclusion of allegedly infringed work in a work created by the defendant party is unreliable117.

English law has also transposed this exception into its national legislation in a very narrow manner, extending it only to copyright in a work is not infringed by its incidental inclusion in an artistic work, sound recording, or broadcast118. It would be hard to extend this exception to TDM. A ruling of the German Bundesgerichtshof also confirmed that “incidental inclusion of a work or other subject matter in other material” as set out at Article 5(3)(i) of InfoSoc has to be interpreted narrowly119. Whilst this case dealt with the incidental publication of copyrighted works into an online catalogue and must again be distinguished from TDM, the German ruling falls in line with settled CJEU case law, in which the judges have systematically confirmed that exceptions and limitations to copyright should be interpreted narrowly120.

In light of this restrictive approach to Article 5(3)(i) of InfoSoc by England and Germany, it is therefore appropriate to assess the approach by Member States to TDM-specific exceptions.

4.3 Limitations and Exceptions Applicable to TDM: Member State Focus

4.3.1 England and Wales

The CDPA was amended in 2014 to create a specific exception to a copyright holder’s exclusive rights in respect of TDM activities. English law provides that:

---

118 See Section 31 of the CDPA.
“The making of a copy of a work by a person who has lawful access to the work does not infringe copyright in the work provided that: (a) the copy is made in order that a person who has lawful access to the work may carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose; and (b) the copy is accompanied by a sufficient acknowledgement (unless this would be impossible for reasons of practicality or otherwise)” \(^{121}\).

Under English law, the scope of the TDM exception is narrow as it only extends to non-commercial research, must be carried out on works to which the TDM operator has already obtained lawful access and must reference the author of the source work, where possible. Interestingly, during the discussions surrounding the creation of a TDM exception in English law, the UK Government deemed it preferable not to have any particular restrictions as regards the beneficiaries of the resulting exception, i.e., it is limited to research organisations \(^{122}\), but perhaps the English-law exception was ultimately restricted by its European counterpart, Article 5(3)(a) of InfoSoc, which provides an explanation as to the limitations curtailing the scope of the exception from applying to broader TDM uses \(^{123}\), despite the UK Government’s acknowledgment that beneficiaries should not be restricted.

A criticism here, aside from the fact the scope of the exception does not cover commercial TDM practices, is that lawful access to the work is a pre-condition, however the notion of lawful access is not defined. Recital 33 of InfoSoc provides that: “a use should be considered lawful where it is authorised by the rightsholder or not restricted by law”, replications of copyright-protected works are restricted by law, which, it has been argued, creates an absurd situation whereby authorisation of the rightsholder would be necessary to rely on the exception, which undermines the entire purpose of an exception \(^{124}\). The UK Intellectual Property Office has indicated that lawful access to a work exists if the reader has a right to read it \(^{125}\). This condition therefore remains open to interpretation and creates uncertainty. In light of this, and as an example which illustrates the unharmonized nature of the national implementations of InfoSoc

\(^{121}\) See Section 29A(1)(a) and (b) of the CDPA, as amended.


\(^{124}\) Ibid.

into national laws, the provisions of Section 29A of the CDPA, as amended, go further than the initial scope of the research exception contained at Article 5(3)(a) of InfoSoc.

Firstly, Section 29A(2) of the CDPA, as amended, provides that where a copy of a work has been made pursuant to TDM activities for non-commercial research, if: (a) the copy is transferred to any other person; or (b) the copy is used for any purpose other than for non-commercial research (i.e., sold, let for hire, or offered for sale or hire), then the copy shall be held as infringing the copyright holder’s exclusive rights in the source work, save where the transfer is authorised by the copyright owner. The scope of Section 29A therefore expressly limits the English TDM exception to the exclusive reproduction right, despite the fact that Article 5(3)(a) of InfoSoc states the research exception is capable of application to both the exclusive reproduction and communication rights. As such, this means that even if TDM activities are lawfully carried out pursuant to the conditions laid down at Article 29A(1), if copies that are made during the technical TDM processes are shared or transferred, this would result in the TDM operator being liable for copyright infringement, a wholly undesirable outcome, particularly for TDM operators which wish to communicate or publish the source data with their output to enable end-users to cross-check and verify the TDM results.

Secondly, Section 29A(5) of the CDPA, as amended provides that:

“To the extent that a term of a contract purports to prevent or restrict the making of a copy which, by virtue of [TDM for non-commercial research purposes], would not infringe copyright, that term is unenforceable.”

Unlike the aforementioned provisions, Article 29A(5) does allow for greater certainty and freedom for TDM operators in the non-commercial research field, in the sense that any contractual override to the exception laid down in UK is expressly prohibited. In practical terms, this means that a contractual clause (e.g. in a license or in terms of service) that prevents or restricts TDM is unenforceable and therefore provides researchers with greater liberty to conduct TDM-based analysis, regardless as to the actual terms of any subscription or licence agreement. Despite this potential positive to the English interpretation of the exception, it is important to note that access to works may nonetheless be restricted by TPMs. Under Article 6

126 See Sections 29A(2), (3) and (4) of the CDPA, as amended.
of InfoSoc, circumvention of such measures is prohibited and constitutes copyright infringement. This therefore constitutes real restriction on the ability of researchers to carry out TDM activities within the UK, particularly in the online environment where such protection measures can be more easily applied to websites. The issue of DRM or TPMs insofar as it relates to TDM is further discussed below.

Finally, this English-law exception is only applicable to copyright in the source data, which is subject to TDM processes, and does not extend to the sui generis database right. It has been submitted that in many cases the owner of the sui generis database right may prevent TDM operations (as it has the exclusive rights of extraction and re-utilization of a database’s substantial part)\textsuperscript{128}, but further consideration of this issue is outside of the scope of this paper. Ultimately therefore, the UK legislation on TDM is not sufficient as a driver of innovation – such TDM activities are only authorised for non-commercial research, such copied (copyrighted) data may not be shared or sold and TDM activities may be severely undermined by TPMs. Whilst authorizations to rely on the exception cannot be overridden by contract, we consider the overall scope of the UK exception creates an unfavourable environment for commercial businesses wishing to carry out TDM or develop their technology in this field, and at best creates only a semi-certain environment for researchers, as it seems highly unrealistic that any research would ultimately remain non-profit, and so despite this exception in the CDPA, TDM activities in the UK continue to carry a high risk of copyright infringement for TDM operators and users of the output.

At the time of writing, it is unclear how UK law will align with its EU counterparts, as “it is a moot point whether the DSM Directive will ever be implemented in the UK or whether the timing of Brexit will prevent that happening”\textsuperscript{129}. Given the aforementioned discussions of the UK Government in 2014, we consider that English law would nonetheless tend to lean towards a pro-TDM environment for all sectors and beneficiaries of such exception.

4.3.2 France

\textsuperscript{128} Ibid.
In 2016, the French legislator introduced a provision governing exceptions for TDM activities via the Law for a Digital Republic\(^{130}\). The French National Assembly considered the purpose of this [TDM] amendment was to promote the flow of data within the research community. The French National Assembly further opined that “it should be enshrined in the law that it is essential to keep the files from the processing which then constitute research data”\(^{131}\). Accordingly, the French National Assembly considered that:

“The capitalization of curation and data formatting work […] is fundamental to researchers. This is a strong demand from the research community. Once retrieved to be searched, the documents are not immediately exploitable, it is necessary to perform processing procedures that can be quite complex. […] The Law for a Digital Republic must therefore provide for the preservation and dissemination of datasets at the end of the research”\(^{132}\).

The French legislature therefore wished to create a favourable environment for TDM. The Law for a Digital Republic amended Article L.122-5-10 of the CPI which provides as follows:

“[When a work has been published], the author cannot prohibit digital copies or reproductions which are created from a lawful source, in order to mine the text and data included in or associated with written scientific works, for the purposes of public research, providing this does not extend to any commercial purpose”\(^{133}\).

It has been noted that, the French legal provision does not refer to “lawful access” or “lawful user”, but rather focuses on the lawfulness of the source itself\(^{134}\). Furthermore, despite the French National Assembly’s apparent understanding of TDM activities and the requirement to save source data and ultimately communicate it, which would be important for commercial TDM activities and the furtherance of business activities in this field, and innovation generally, much like its European and English counterparts, the exception is limited to “research” and arguably goes a step further by specifying the notion of “public research”. This wording means the provision is even more narrow than the aforementioned counterparts as it likely refers to “research carried out by public organisations” with a further limitation to source data to be

\(^{130}\) Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique.


\(^{132}\) Ibid.

\(^{133}\) Translation by the author from the original French wording: Les copies ou reproductions numériques réalisées à partir d'une source licite, en vue de l'exploration de textes et de données incluses ou associées aux écrits scientifiques pour les besoins de la recherche publique, à l'exclusion de toute finalité commerciale.

\(^{134}\) Text and Data Mining (TDM) exceptions in the UK and France - Text and Data Mining exception in France, Clarin, available at: https://www.clarin.eu/content/clic-text-and-data-mining-tdm-exceptions-uk-and-france and accessed on 31 March 2019.
mined which is only “included in or associated with scientific publications”135. This is obviously an unsatisfactory solution for commercial TDM as well as for TDM carried out by private research institutions.

More generally, the CJEU further reminds us that “for the interpretation of each of the conditions [of Article 5 of InfoSoc], it should be borne in mind that […] the provisions of a directive which derogate from a general principle established by that directive must be interpreted strictly”136, i.e., the consent of the right holder is required as a default position, and that:

“[T]he exemption(s) must be interpreted in the light of Article 5(5) of InfoSoc, which means that the exemption is to be applied only in certain special cases which do not conflict with a normal exploitation of the work or other subject-matter and do not unreasonably prejudice the legitimate interests of the rightsholder”137.

It would therefore be a high-risk approach for TDM operators to attempt to rely on a wide and flexible interpretation of any exception under Article 5 in the hope of rendering their activities on copyrighted work lawful. The focus is on rightsholders privileges – which whilst providing certainty to creators within the EU, conflicts with the needs of innovators who must utilise copyrighted works to achieve highly developed machine learning and AI tools and useful and analytical output.

Ultimately, as technology has developed, the EU’s approach to copyright law did not follow suit. The result is that the environment to date has been unfavourable for TDM-based innovation due to non-adapted and outdated copyright laws, not least for TDM specifically, which was not even anticipated by InfoSoc back in 2001.

5. The DSM Directive – the Solution to Risk-Free TDM in Europe?

As we have seen, TDM is an area where no appropriate exception already existed in the European regulatory framework138. This has resulted in the relevant players either avoiding TDM activities and the development of and investment in the relevant technology, or which

---

135 Ibid.
136 Judgment in Case C-5/08 Infopaq, para 56.
137 Ibid, para 58.
has obliged these relevant players to resort to licensing solutions, and therefore shoulder the administration and transaction costs of the licenses\textsuperscript{139} to avoid copyright infringement claims. Considering the negative impact this has on the furtherance of the Digital Agenda, the Commission wished to “ensure wider access to knowledge by simplifying copyright rules in the areas of text and data mining for research and other purposes”\textsuperscript{140}. The Commission further affirmed the DSM Directive means that:

\[\text{“[R]esearch organisations, universities and other users will be able to make the most of the increasing number of publications and data available online for research or other purposes as they will benefit from a copyright exception to carry out TDM on large sets of data [which] will also enhance the development of data analytics and AI in Europe”}\textsuperscript{141}.

This positive statement is also reinforced in the fact that “one of the important aspects DSM Directive is that it introduces a number of exceptions which are indeed mandatory on all member states, departing from the smorgasbord approach of the previous Infosoc Directive which presented a number of exceptions from which member states could pick and choose”\textsuperscript{142}.

Given this context, one could consider the provisions relevant to TDM activities contained in the DSM Directive therefore provide the solution that had failed to exist within the EU until now, however to consider the accuracy of this assumption, it will be necessary to consider the journey of the TDM exception to its current form and provide an analysis of the finalized TDM provisions.

5.1 The Journey of the TDM Exception in the DSM Directive

The journey of the DSM Directive from the initial Proposal to the final agreed text was somewhat arduous. IP professors consider that “content has attracted substantial commentary” and “conflicting opinions amongst the EU institutions (the Commission, the Committees of the European Parliament, the Council)”\textsuperscript{143} regarding a number of provisions including the wording related to the proposed TDM exception. It is therefore necessary to analyse the scope of the

\textsuperscript{139} Ibid.
\textsuperscript{141} Ibid.
initial wording to understand the application of the present provisions as cemented in the DSM Directive.

5.1.1 The Proposal

Article 3 of the Proposal initially envisaged a mandatory exception for TDM activities which extended to “reproductions and extractions made by research organisations 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”\(^\text{144}\) only. To arrive at this initial solution to the TDM uncertainty in Europe to date, the Commission considered four options for implementing TDM exceptions which would curtail the exclusive rights of copyright holders, thus driving innovation through TDM in the EU forwards: (i) self-regulation initiatives from the industry; (ii) exceptions covering only non-commercial scientific research purpose; (iii) exceptions for commercial scientific research purpose, limited to the benefit of some beneficiaries; and (iv) exceptions for commercial scientific research purpose with no limitation on beneficiaries\(^\text{145}\).

Despite the clear business context within which the EU committed to the Digital Agenda and DSM strategy, the third option was initially considered by the legislators to be the most appropriate framework. This therefore initially translated into Article 3 of the Proposal which limited the TDM as follows:

“1. Member States shall provide for an exception to the [rightsholders’ exclusive] rights provided for in [EU legislation]\(^\text{146}\) for reproductions and extractions made by research organisations 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.
2. Any contractual provision contrary to the exception provided for in paragraph 1 shall be unenforceable.
3. Rightsholders 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.


\(^{146}\) i.e., Article 2 of InfoSec, Articles 5(a) and 7(1) of Directive 96/9/EC and Article 11(1) of the Proposal.
4. Member States shall encourage rightsholders and research organisations to define commonly-agreed best practices concerning the application of the measures referred to in paragraph 3”.

The reality of the TDM exception as initially provided for in the Proposal, therefore, was far from the Commission’s intention to drive innovation within the EU, certainly in the area of TDM and Big Data for commercial purposes. Indeed, whilst the drafting of this exception imposed a mandatory exception on Member States, which is a positive aspect when compared to the majority of optional exceptions contained in EU legislation to date, and whilst certainty is provided to TDM actors in that the scope of the legal provision is not capable of being overridden by contract, the extent of the Proposal’s TDM exception was highly limited as the scope of the exception not only required operators to have “lawful access” to the copyrighted works, which presumably meant either via open access channels, or pursuant to licence or subscription agreements, but it also limited the scope of the TDM exception to academia, and notably to “scientific research”, which is to the exclusion of start-ups and innovators which carry out TDM for commercial means.

Furthermore, the wording of the Proposal expressly permitted the application of TPMs to protect copyrighted works, which means that it would have been possible to technically prevent reproduction and extraction of copyrighted works, even when a lawful exception permits such actions. Additionally, the provision only seems to extend to an exception to the right of reproduction, and the possibility to communicate TDM source data or indeed output containing copyrighted works, does not appear to have been dealt with. These issues, which remained, are further discussed below as regards the final wording of the DSM Directive. The text of the Proposal was therefore highly criticized by academics and innovators alike. It was therefore criticised by industry leaders that Article 3 of the Proposal:

“[O]nly covered non-for-profit and public research institutions limiting the ability of private companies to carry out TDM in Europe which [would have] inevitably lead to the most promising European AI start-ups and companies relocating to the US or Japan, where they [would have] access to broader datasets, and [would have been] able to build algorithms of better quality”\textsuperscript{147}.

As drafted, such industry leaders feared that Article 3 of the Proposal would have the opposite effect of the Commission’s goals for the DSM strategy.\textsuperscript{148}

In short, whilst the Proposal’s commitment to scientific research by research organisations was admirable, it was nonetheless apparent the exception did not go far enough to support Europe’s Digital Agenda from an innovation and commercial perspective.

5.1.2 The DSM Directive – the Final Word

Given the modifications to the finalized text of the DSM Directive, it appears the EU took such criticisms on board when drafting the final version of the DSM Directive, at least to a certain extent. Does the finalized version of the DSM Directive mark a definite change in approach from the historically unfavourable TDM environment that the legal uncertainty in Europe has provided to date, and lead us towards a positive future for reduced risk TDM activities in Europe for the future and future success of the DSM strategy?

As a preliminary remark, it is important to note the legal order of Articles 3 and 4 of the DSM Directive. Indeed, the provisions of the DSM Directive are based upon and complement the pre-existing EU legislation on copyright.\textsuperscript{149} This means the exceptions contained in the pre-existing EU legislation will continue to apply provided that they do not conflict with provisions of the DSM Directive. In theory, this should mean that instances of TDM should be covered by the DSM Directive, however the preamble expressly states that in instances of TDM:

\begin{quote}
"[W]hich do not involve acts of reproduction or where the reproductions made fall under the mandatory exception of temporary acts of reproduction laid down in article 5(1) of InfoSoc [such provisions] should continue to apply to TDM techniques which do not involve the making of copies beyond the scope of that exception."\textsuperscript{150}
\end{quote}

Whilst the DSM Directive aims therefore at providing certainty for the actors involved in TDM, uncertainty still remains as operators falling outside of the scope of the DSM Directive provisions might still be able to rely on pre-existing law as a fallback argument. We consider has the potential to result in confusion. It will be important to see how this dual-regime will

\textsuperscript{148} Ibid.
\textsuperscript{149} DSM Directive, Preamble, Recital 4.
\textsuperscript{150} DSM Directive, Preamble, Recital 9.
apply to practical cases and whether there is any actual consequence for the various stakeholders.

Turning to the wording of the final TDM provisions of the DSM Directive, Article 3 remains materially unchanged from the wording of the Proposal. There remains a mandatory exception for research organisations for scientific research. Indeed, the important change contained in the DSM Directive as compared to the Proposal is the inclusion of Article 4, which provides:

“Exception or limitation for text and data mining

1. Member States shall provide for an exception or limitation to exception to the rightsholders’ exclusive rights provided for in EU legislation\(^\text{151}\) for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of TDM.

2. Reproductions and extractions made pursuant to paragraph 1 may be retained for as long as is necessary for the purposes of TDM.

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 rightsholders 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 DSM Directive now expressly provides for a mandatory exception or limitation to be implemented by Member States for TDM activities beyond the previous narrow provisions limited solely to scientific research by research organizations as initially laid down in the Proposal. This means the EU has now gone some way to resolve the issues related to commercial TDM activities and thus drive innovation in the region. In reliance upon Articles 3 and 4 of the DSM Directive, innovators, as well as researchers in Europe can now benefit from a certain amount of legal certainty in that their TDM activities involving reproductions and extractions of copyrighted source data are capable (in theory) of benefitting from a lawful limitation or exception, however, the issue of communication to the public of TDM source data and output containing copyrighted materials unfortunately does not seem to have been dealt with.

\(^{151}\) i.e., Article 2 of InfoSoc, Articles 5(a) and 7(1) of Directive 96/9/EC and Article 11(1) of the Proposal.
Further, we now have two regimes for TDM exceptions and limitations – one for scientific research and another for all other TDM activities. As we will see, the rights and obligations for each category of TDM are not aligned, and this is to be criticized - the playing field is not-level between different categories of actors and purposes, and the environment for commercial TDM therefore remains unfavourable in Europe, which is contrary to the Digital Agenda.

Nonetheless, the TDM provisions in the DSM Directive are not without their positive aspects. It is therefore appropriate to discuss the positive aspects of the DSM Directive in driving innovation in Europe for commercial and non-research purposes, before addressing the negative aspects, as in our view it appears the EU-approach to TDM still remains problematic constitutes therefore a potential hindrance to the competitiveness of the EU compared to other key jurisdictions.


5.2.1 Harmonisation

A key benefit of the DSM Directive’s provisions related to TDM is that for both scientific and non-scientific TDM activities, the provisions impose a mandatory exception or limitation for TDM activities on Member States. To a certain extent, this is likely to create a harmonization of Member States’ laws in respect of TDM. It is therefore hoped the drafting of the DSM Directive will reduce fragmentation in the approach and application of national laws for TDM activities from one Member State to another which would not only create much-welcomed certainty for the relevant actors, but would also respond to the EU’s policy goals for its Digital Agenda, namely to provide a normalised, consistent level playing field across Europe to legally carry out TDM projects\textsuperscript{152}.

As Geiger et. al submit, the harmonised framework for TDM will drive innovation in the digital single market and promote EU-wide, integrated, larger research projects\textsuperscript{153}. Whilst we agree, we do nonetheless consider the EU could have gone further by providing a broader exception to TDM (particularly for activities conducted by start-ups and independent innovators), subject


\textsuperscript{153} Ibid.
to fewer caveats and restrictions which would have resulted in a more certain harmonization and would have better served to drive innovation in the region.

5.2.2 Express Recognition of TDM and Expanded Scope

A further positive aspect of the DSM Directive’s provisions on TDM is that it is the first occasion, at an EU-level, that TDM has been expressly recognized and codified. This evidences a recognition by the EU of innovative and valuable technological tools and mechanisms within the world of data analytics and provides awareness of TDM as a process of harnessing the value of Big Data. This encourages innovators AI and machine-learning development and encourages TDM in the region, given that such activities have formal recognition, in respect of copyright implications, in official European texts. Additionally, the DSM Directive has provided a more positive environment when compared with the Proposal as the limitation / exception is now capable of covering both commercial and non-commercial uses\textsuperscript{154}, albeit, subject to certain restrictions, as set out below.

5.2.3 Rule Against Contractual Override

An important positive aspect of the TDM provisions in the DSM Directive for scientific research is that Article 7(1) expressly provides for the unenforceability of contrary contractual provisions. In practice, it would not be possible for copyright holders to expressly exclude the application of Article 3 through contract. It is arguable that this provision entrenches previous CJEU case law on the subject – for example, in VG Wort\textsuperscript{155}, it was confirmed that “the default position where contract or licence terms are not expressly allowed to limit the scope of an exception in that the exception will prevail over any rights holder authorisation”\textsuperscript{156}.

In any case, the prohibition of contractual override is critical so as not to devoide the provision of any practical utility. Previously, copyright holders, such as publishers of scientific research, would have been able to contractually exclude TDM in licence agreements whilst applying high transaction costs on TDM operators to obtain consent to mine content for research. As the nature of TDM in respect of Big Data means that it is important to mine source data belonging to

\textsuperscript{154} Ibid.
\textsuperscript{155} See CJEU cases C-457/11 to C-460/11.
\textsuperscript{156} See Voluntary Memorandum from the UK Department for Business, Innovation and Skills, available at: https://publications.parliament.uk/pa/jt201415/jtselect/jtstatin/13/1321.htm and accessed on 11 May 2019, para 5.
multiple copyright holders, such licensing activities would make TDM projects in the field of scientific research unsustainable\textsuperscript{157}, and ultimately hinders innovation.

Whilst a contractual override is possible for activities falling under Article 4 (i.e., any non-scientific research activities), this may not negatively impact the application of the TDM exception. The scope of such contractual override is limited by contract law principles - privity of contract precludes the imposition of contractual obligations on third parties. As such, when information that has been mined and reproduced from contractually protected source data is further disseminated online by a third-party (who has taken this information from the original TDM operator’s output data), the owner of the source data could be precluded from bringing a claim for breach of contract against that third-party\textsuperscript{158}. Accordingly, such contractual overrides under Article 4 of the DSM Directive would arguably only affect the initial TDM miner and not those carrying on subsequent dealings in the mined source data.

In light of these considerations, the authorisation for contractual override in respect of Article 4 of the DSM Directive is arguably of limited consequence. Nevertheless, the prohibition to such contractual override is therefore a welcome step towards a favourable environment for EU-based scientific research activities involving TDM, although, as discussed below, this leaves much to be desired for other TDM activities, despite the potential contract law limitations.

5.2.4 Certainty for Copyright Holders

Finally, the TDM exceptions and limitations in the DSM Directive create a stronger environment for rightsholders. The drafting of the provisions sets out a strict framework providing for specific instances in which TDM may be lawfully operated on copyrighted source data, without consent of the rightsholder. The specific and stringent requirements which must be met in order for the beneficiaries of these exceptions and limitations to rely on them is wholly justified in order to create a safe environment for rightsholders, and to protect them against unjustified or unfair exploitation of their works.

\textsuperscript{157} Ibid.
From this perspective, it is right that Article 4 of the DSM Directive may be so easily disapplied though copyright-holders’ reservation of their work, permitting originators of works to retain control. The fact the research exception set out at Article 3 of the DSM Directive is more lenient on TDM operators takes into account the difference between the source data – perhaps where academics have already been remunerated or have obtained funding to draft the source work, or gain notoriety from having works mined and republished, when compared to creators of more artistic content, which may still be mined for the purposes of AI learning, creating significant sums and recognition for the data scientists and developers; benefits which are not shared with owners of source work exploited, without which the innovative success would never have been possible. To illustrate, consider the recent sale by Christie’s of an AI-created artwork painted using an algorithm¹⁵⁹ which used TDM-style techniques to exploit pre-existing artistic works created by humans in order to “teach” the AI how to recreate a similar work, which sold for nearly 500,000 USD. A profit which the rightsholders’ of the source works (or their heirs) did not share.

The provisions in the DSM Directive therefore create an environment in Europe which is favourable to creators - surely a positive outcome for the region, but which conversely creates the balance away from creating a robust and desirable environment for technological innovators and furthering the Digital Agenda. It is therefore appropriate to analyse the negative aspects of the DSM Directive, from the perspective of the research and innovation communities.

### 5.3 Negative Aspects of the DSM Directive’s TDM Provisions

Despite the aforementioned positive aspects, it has nonetheless been argued the “project to allow Europeans to conduct TDM, which is crucial for modern research and the development of AI, has been obstructed with too many caveats and requirements”¹⁶⁰. We discuss these caveats and requirements in turn.

#### 5.3.1 The Harmonization is Still Open to Fragmentation

Firstly, it is important to note the wording of Article 4 of the DSM Directive for non-research TDM expressly states that the mandatory obligation on Member States to implement a

---


¹⁶⁰ The Text of Article 13 and the EU Copyright Directive has just been finalized, Julia Reda available at: https://juliareda.eu/2019/02/eu-copyright-final-text/ and accessed on 11 May 2019.
mandatory exception or limitation, whereas Article 3, in respect of research purposes, provides for an exception only. This means that for non-research TDM, i.e., commercial or innovative non-research purposes, Member States still possess some discretion as to the nature of the provision they choose to implement: TDM may be open to fragmented and unharmonized treatment from one Member State to another, depending on whether any given Member State chooses to except TDM activities from constituting an infringement of a copyright holder’s exclusive rights, or chooses to curtail those exclusive rights by implementing a limitation. Whilst this could be a case of semantics, there is potential for lack of certainty as to the precise treatment of TDM activities falling within Article 4 of the DSM Directive from one Member State to another. In terms of practical risk related to this distinction, whilst recital 13 of the preamble to the DSM Directive expressly excludes Member States from providing for compensation for rightsholders in respect of the TDM exceptions, there is no such reservation for limitations.

Arguably, this difference in treatment is undesirable and creates uncertainty as to the financial exposure of TDM operators seeking to rely on Article 4, depending on how it is implemented nationally. Of course, this is something to monitor when reviewing how Member States implement Article 4 into national laws, and then apply the law before national courts in order to ascertain whether this distinction between “exception” and “limitation” is indeed, a moot point.

Secondly, as regards the research exception contained at Article 3, Article 3(4) provides that:

“Member States shall encourage rightsholders, research organisations 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”.

This again means that each Member State has some discretion as to how to deal with the practical application of the measures applied by right holders ensure the security and integrity of the networks and databases where the works or other subject matter are hosted as well as the level and scope of storage and security of source data, at least for research organisations and cultural heritage institutions. This flexibility again increases the risk that the precise nature of the TDM exception may vary as between Member States, thus creating uncertainty for TDM operators, researchers and rightsholders across the EU. Similarly, we must monitor how
Member States define and apply such best practices as they transpose the DSM Directive into their respective national regimes.

5.3.2 Undesirable Difference in Treatment Between Research and Other TDM Activities

5.3.2.1 The Broader TDM Exception is Devoid of Function Due to the Possibility for Contractual Override

Article 4(3) of the DSM Directive provides that:

“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 rightsholders in an appropriate manner, such as machine readable means in the case of content made publicly available online”.

This provision results in a scenario whereby holders of copyrighted works are entitled to expressly disapply Article 4 for all TDM activities, save those related to scientific research (which is solely governed by the provisions of Article 3). It is appropriate to criticize the practical application on Article 4 in light of this flexibility, since this wider provision is so easy to disapply, either by technical means or indeed by contract or unilateral declaration\(^{161}\). This is confirmed by recital 18 of the Preamble to the DSM Directive which is further cemented by Article 7(1) which does not include Article 4 in the express protection against contractual override; there can be no uncertainty as to the ability of rightsholders to override Article 4 at their discretion. Practically, we can envisage rightsholders in an online environment will disapply Article 4 as in the PR Aviation/RyanAir case through the application of exclusions to screen-scraping TDM activities in website terms and conditions\(^{162}\).

Whilst the aforementioned privity of contract argument might restrain the wide-ranging applicability of such contractual override, the fact that a unilateral declaration is expressly permitted, means that a simple notice applied to a website is sufficient and that any notion of consent or acceptance in a contractual sense falls away (provided national laws so permit).

---

\(^{161}\) Recital 18 of the DSM Directive.

\(^{162}\) Judgment in Case C-30/14.
Of course, such practices are inherently difficult to monitor, TDM operated by AI is generally unable ascertain when such contractual restrictions have been applied to a website, and of course, depending on the status of the TDM operator (i.e., a research organization or otherwise) the legal application of such terms will vary meaning that in some instances the contractual override will be valid, and in other circumstances, not. The difficult application of contractual overrides or unilateral notices is akin to the application of technological restrictions, as are discussed further below, and also potentially creates further confusion for TDM operators depending on the nature of the IP rights protecting the source content – for example, when screen scraping activities include copyrighted works or works covered by the Database Directive, for which such contractual override is not permissible163, or when a notice or contractual restriction does not take into account activities which do not require consent of the relevant rightsholder.

In light of the foregoing, we consider the scope of Article 4 is therefore essentially an optional exception for broader TDM activities not falling within the remit of Article 3. Is this evidence of the EU simply playing lip-service to the industry criticisms surrounding the narrow scope of the initial TDM exception as contained in the Proposal, or just evidence of a failure to understand the realities of TDM and the likelihood of harm suffered by rightsholders? Either way, it is regrettable that despite much negotiation and effort, the DSM Directive still fails to provide non-research TDM operators with certainty as to their activities and their protection against copyright infringement actions, such exposure being depending on an individual rightsholder’s reservation – something which is almost impossible to monitor and check as data mining analytics processes huge volumes of information, often coming from thousands of source resources, meaning that such organisations may find themselves in the precarious position of relying on the provisions of Article 5 of InfoSoc, which as we have outline above, are not sufficient. As it stands, the ability of rightsholders to exclude Article 4 by various means effectively renders the provision devoid of function, leaving Europe an uncertain environment for TDM actors.

5.3.2.2 The Scope of the Broader TDM Exception Remains Unclear

163 See Articles 6(1), 8 and 15 of Directive 96/9/EC.
On reading Article 4, there is likely to be general confusion about the extent of the exception or limitation for non-research purposes. Article 4(2) states that: “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”.

In our view, “necessary for the purposes of TDM” provides for uncertainty as it could be argued by rightsholders that “necessity” is simply limited to the time required to complete the technical process, whereas operators might wish to extend that time frame to purposes which go beyond the mere technical processes. This concern was also raised by the Association of European Research Libraries which stated that:

“*We are very concerned that Article [4] only allows “temporary reproductions” which does not reflect the realities of data analytics and verification of research results. No public or private player will make the large investment required to mine data if it can only keep the material it has analysed on a temporary basis. This [provision] does not reflect the realities of TDM*”\(^{164}\).

This is to be contrasted with the provisions in the DSM Directive regarding scientific research which is not subject to the same restrictions of “necessity”. Indeed Article 3(2) of the DSM Directive states that: “*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*”\(^{165}\).

It is therefore undesirable the DSM Directive creates stark differences regarding the retention of source data depending on whether such data has been mined for research or other purposes. Unfortunately, yet again, the EU has failed to provide certainty for TDM activities across the board and the retention rights of source data is therefore very much subject to the identity of the TDM operator and the purpose that this operator seeks to achieve. Although, the difference in treatment may be justified by the fact that it is often more critical for researchers need to have recourse to source data for a longer period for verification purposes when compared to source data for *some types* of commercial or non-research TDM processes which can, *in some instances*, be more readily discarded.

---

165 Article 3(2) of the DSM Directive.
Regardless of the reasoning, it seems to us that certainty as to the interpretation of “necessary” at Article 4(2) of the DSM Directive is therefore likely to be a matter which will come before the CJEU for clarification, meaning the overall impact on the provision actually hindering or further innovative TDM-based activities in Europe remains to be seen as the legislation is implemented and applied.

5.3.3 Undesirable Restrictions for Research Organizations and Commercial Entities Alike

5.3.3.1 Qualification Restrictions for the Research Exception

Whilst recital 12 of the preamble to the DSM Directive provides for a wide notion of scientific research which extends to both natural and human sciences, and provides certainty for specific categories of beneficiaries of the exception contained at Article 3 by listing potential organisations falling within the provisions\(^\text{166}\), Articles 2(1)(a) and (b) restrict the scope of qualifying for the exception contained at Article 3 by providing a narrow interpretation of research organisations. Indeed, these provisions provide that research organisations must operate on a not-for-profit basis, or by reinvesting all the profits in its scientific research or pursue a public interest mission funded by public funds or public contracts, in order to qualify for certainty as to their TDM activities. As such, the scope of the DSM Directive is prohibitively narrow when defining the nature of a research organization, beneficiary to Article 3 by expressly stating 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”\(^\text{167}\).

This essentially restricts commercially backed research organisations from being able to benefit from Article 3, even if they are ultimately carrying out “scientific research” for the purposes of the DSM Directive, and therefore excludes certain research organisations, such as private universities. Whilst on one hand, the EU has clearly wished to ensure that scientific research

\(^{166}\) Recital 12 of the DSM Directive lists: “universities or other higher education institutions and their libraries, also entities such as research institutes and hospitals that carry out research”

\(^{167}\) Recital 12 of the DSM Directive.
carried out for TDM purposes remains neutral and independent from industry, in the current austerity climate in which public funding and investment is scarce and where regard must be had to the private sector to obtain budget for the most cutting edge research, the restriction to qualification for this exception is perhaps unwittingly a move by the EU, which will result in stunted innovation through research in the region. As argued by academics, “from a practical market-based perspective, this policy choice might cripple opportunities for start-ups and individual researchers in this area”\textsuperscript{168}.

Furthermore it has been submitted the “policy choice of excluding from the reach of the exception unaffiliated individuals and researchers – operating under the same terms as those organised in a qualifying research organisation – might fall short in terms of adequacy and proportionality”\textsuperscript{169}, thus resulting in a problematic and unfavourable TDM environment within Europe, particularly for independent data scientists or think-tank personnel who cannot rely on the TDM exception\textsuperscript{170}, thus meaning the lawfulness of TDM operations in Europe from a copyright perspective remains uncertain for these actors.

5.3.3.2 \textit{The Issue of “Lawful Access”}

The DSM Directive provides for both the research and non-research TDM the relevant limitation/exception shall only apply to operators which have “lawful access” to the copyrighted content.

Of course, from a rightsholder perspective, this allows one to retain control over copyrighted content, which is a benefit and encourages creation of original works in the traditional sense. However, from an innovation perspective, this requirement creates a potentially difficult environment for TDM operators and the quality of TDM output as applicability of Articles 3 and 4 to the source data to be mined “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”\textsuperscript{171}.

\textsuperscript{169} Ibid.
If access to the volume of source data is limited in any way, or is subject to unnecessary risk (i.e., of copyright infringement), then not only would the TDM output be less thorough and thus less valuable, the reduced amount of source data processed by machines could also have a negative impact on the development of AI as such technology requires significant amounts of data to machine-learn.

From a research perspective, it has been argued that subjecting TDM to lawful access will make TDM research projects harder to run by raising related costs, meaning that publishers of content might price TDM into their subscription fees, if only those with lawful access can perform TDM research172. It has also been argued that subjecting TDM activities to market access discriminates according to the market power of the TDM operators173. Most start-ups and research organisations will be effectively prevented from being able to gain lawful access to works due to the cost of such access, which for research organisations specifically, is problematic for TDM operators coming from less economically sound environments where public funding may be scarce. Instead of encouraging a harmonised environment to foster innovation across the EU, the requirement of lawful access effectively “widens the gap between richer and poorer research institutions and, most likely, increase the scientific and innovation divide between developed and less-developed European countries”174.

Ultimately, for both commercial and research TDM, the overall quality and value of the AI development and TDM output is likely to be put at risk whenever there is a requirement for lawful access which results in payments being made or costly subscriptions being taken out, as budget restrictions take over175. The underlying result is that where cost of conducting TDM is increased, researchers and innovators are less likely to use it, and will potentially privilege either other methods or move their activities to territories where such cost-burdens are lower, and which may be located outside of the EU altogether.

5.3.3.3 The Licensing Burden

As indicated above, lawful access is therefore likely to be subject to licence arrangements, save where such data is freely available in the public domain, or indeed when source data is not

---

173 Ibid.
174 Ibid.
175 Ibid.
protected by copyright protection and so no consent, and therefore no licence is needed to perform TDM. Nevertheless, where consent is required and the mechanism to obtain such consent falls to licensing or contractual arrangements, the situation is problematic.

Arguably, the research sector is more heavily affected by the requirement of lawful access and the potential issues related to the licensing of content to be mined since commercial TDM is often focused in some areas of online analytics (such as retail analytics) are often related to consumer movements and trends gained through the use of cookies, plug-ins or social media.

However, whilst the content to be mined for commercial purposes is therefore more often freely available, whereas this is not necessarily the case for scientific research where the source data is subject to more onerous access restrictions (online databases or private library content), the DSM Directive nonetheless assumes that profit-making firms can and should get a license to engage in TDM research from the owners of the affected IP rights, which is not necessarily the case, especially for start-ups which have limited access to financial resources.\footnote{Ibid, p. 22.}  

It is a fact that “TDM licenses may not be available on reasonable terms for start-ups and small businesses in the EU”\footnote{Ibid, p. 22.}, which means that such innovative companies at the seed stage may prefer to commence operations outside of the region. From a research perspective, the realities of licensing practices mean that whilst some publishers take an active role in developing text and data analytic technologies, and that some offer contracts that support the use of these technologies\footnote{Modernising copyright: a modern, robust and flexible framework. Government response to consultation on copyright exceptions and clarifying copyright law, HM Government, 2012, p. 37.}. The UK Government opined that “research projects may in some cases require specific permissions from a large number of publishers in order to proceed which is in some cases an insurmountable obstacle, preventing a potentially significant quantity of research from taking place at all”\footnote{Ibid.}, which again does not foster a TDM-friendly environment in Europe for research purposes. The finalized TDM provisions in the DSM Directive reinforce this environment.

5.3.4 Unresolved Issues Related to Coexistence with TPMs
A further issue reducing the effectiveness of the DSM Directive in creating a favourable environment for TDM activities in Europe is the inclusion of provisions which confer on rightsholders the possibility of limiting TDM activities via the application of technological protection measures, which result in the ability for the owners of copyrighted works to block access to operators seeking to carry out TDM. Recital 7 of the preamble to the DSM Directive states that:

“[T]he protection of technological measures established in InfoSoc remains essential to ensure the protection and the effective exercise of the rights granted to authors and to other rightsholders”.

Whilst the DSM Directive also states that such use of technological measures should not prevent the enjoyment of available exceptions and limitations, it appears that practically this is very hard to ensure since such measures are voluntary for rightsholders to apply, not least because it is for such rightsholders, or indeed Member States, to select the appropriate means of enabling the beneficiaries to benefit from the exceptions and limitations.

Of course, the recitals of the preamble are relevant to all aspects of the DSM Directive, but Articles 3 and 4 clarify the situation in respect of research and other TDM activities by confirming that respectively that:

“Rightsholders 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”,

and in the case of broad TDM activities not applicable to research-TDM, provided that copyrighted source data has not been “expressly reserved by their rightsholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online”.

Geiger et al. have noted that despite the DSM Directive’s good intention:
“[T]he application of anti-circumvention provisions might trample over users’ privileged uses […] technical protection measures might limit or prevent access to works altogether for purposes that are not restricted by authors’ rights or for uses that are actually privileged”\textsuperscript{184}.Whilst detailed discussion of this point is outside the scope of this paper, we agree with Geiger: the DSM Directive does little to balance the interests of TDM operators and rightsholders in practice as regards the application of anti-circumvention provisions on works which can be mined under a lawful exception, and it will be interesting to note how the Member States approach this dilemma from a national perspective once the provisions have been implemented and are applied to concrete cases.

Ultimately, as with other aspects of the TDM provisions in the DSM Directive, the drafting of the text as regards TPMs means there is a risk of inconsistent implementations across national jurisdictions which might effectively curtail harmonised enjoyment of the new mandatory exceptions, thus limiting the effectiveness of the DSM\textsuperscript{185}. The result of this may indirectly encourage innovators away from the EU to regions where such restrictions are less likely to impact their TDM activities or indeed which are less likely to hinder reliance on lawful exceptions or limitations, which would ultimately restrict TDM innovation across Europe.

5.3.5 Incompatibility with the International Community

Finally, the EU-approach to TDM in the DSM Directive is at odds with the international community. The EU legal framework does not align with other jurisdictions adopting broad exceptions or fair use models permitting larger numbers of players to perform TDM and promote related innovation\textsuperscript{186}. Due to the heavily caveated provisions of the DSM Directive’s TDM exceptions, “some EU firms may ship their TDM activities outside of the region in order to take advantage of less restrictive TDM rules elsewhere”\textsuperscript{187}. We agree this is a valid risk. Similarly, some non-EU firms may decide not to invest in TDM-related research in the EU due to the highly restrictive TDM provisions\textsuperscript{188}, which is certainly counterintuitive to the

\textsuperscript{185} Ibid, p. 838.
\textsuperscript{186} Ibid, p. 835.
\textsuperscript{188} Ibid.
Commission’s desire to promote Europe as an innovation centre and evidences that Europe’s approach to copyright laws continues to hinder innovative activity, particularly involving TDM, which is so core to many technological developments.

5.4 The Overall Impact of DSM Directive’s Approach to TDM

The TDM provisions are a step in the right direction\(^{189}\), but the situation remains uncertain for innovators and researchers alike. The DSM Directive’s TDM provisions demonstrate the EU’s lack of understanding of the technical realities of TDM, and the financial and contractual limitations which are faced by those performing such processes.

Dr. Rosati considers that “TDM itself is not about competing with existing content or disrupting existing models”\(^{190}\). We concur. Consequently, when TDM operators encroach on the exclusive rights afforded to copyright holders in relation to access to, extraction and copying of content for TDM purposes, it is to be noted these steps are “all incidental stages that do not ultimately result in the external re-use of protectable (expressive) parts of a work, but are rather functional to accessing those arts that are unprotected, including ideas, data, and facts considered on their own”\(^{191}\).

By applying the caveats and restrictions to the TDM provisions in the DSM Directive, and establishing a difference in treatment between research and commercial TDM activities, rather than creating an environment of innovation and collaboration favourable to innovation, the EU has “overlooked the fact that TDM is not about displacing existing content but rather extracting further knowledge from it and, in doing so, rendering it more valuable”\(^{192}\). Through these provisions, the EU has made it more difficult for such valuable output to originate from European-based operators. To drive innovation in the EU and further the commitment to the DSM, it would have been more desirable for the EU to implement a broad and all-encompassing TDM exception which is not capable of override\(^{193}\), as is the case in other key innovative jurisdictions.

---


\(^{191}\) Ibid.

\(^{192}\) Ibid.

6. Another Way: Alternatives and the Future

In the highly competitive global market for world-class AI and data science researchers, the EU may suffer from “brain drain” if its most talented researchers take job opportunities in jurisdictions where TDM is subject to fewer restrictions. In light of the global nature of the modern economy, the impact of the TDM exception on the EU’s competitive advantage compared with other top innovative economies that enable all undertakings to carry out TDM under fair use/fair dealing models (e.g. US, Canada, Israel) could have been taken into account by the EU legislature. It is therefore appropriate to briefly consider how other jurisdictions have approached TDM, as well as potential mechanisms that can be implemented to make the most of the TDM provisions in the DSM Directive to encourage innovation in the field, despite the uncertain legal framework.

6.1 Fair Use – The American Example

Fair use is a doctrine which can permit lawful use of copyrighted material without obtaining rightsholders’ consent. Examples of fair use include search engines, parody and research. Fair use is not so much as an exception or limitation to copyright but rather a defence justifying a third party’s use of the work. One could consider the beneficiaries of fair use are the parties to a copyright infringement dispute, rather than being categorized as research organisations, non-profit institutions or commercial operators, or being distinguished by the purpose of the TDM operations, as is the case in the DSM Directive.

A key jurisdiction within the innovation world which has long since introduced the doctrine of fair use is the US. 17 US. Code s.107 provides the following 4 factors must be considered in order to determine whether the use made of a copyright-protected work may be considered fair:

“(I) the purpose and character of the use, including whether such use is of a commercial nature or is for non-profit educational purposes (a key consideration in relation to this factor is whether the use is transformative, i.e. whether the new work merely supersedes

---

196 Text and Data Mining and the Need for a Science-friendly EU Copyright Reform, Science Europe Working Group on Research Data, Editor: Christoph Bruch, D/2015/13.324/1, p. 10 citation 4.
or supplants the original creation or whether, instead, it adds something new, with a further purpose or different character; (2) the nature of the copyright work; (3) the amount and substantiality of the portion used in relation to the copyright work as a whole; and (4) the effect of the use upon the potential market for or value of the copyright work.”

It likely that under US law, TDM activities for research purposes would clearly fit within the scope of this exception, since the overall TDM output is highly transformative, taking just extracts of copyrighted work, which has a large, positive impact for society, does not take away from the value of the source data, but adds to it.

However, much like the difference in treatment between research and non-research purposes at Articles 3 and 4 of the DSM Directive, the argument falls again as to whether or not it would be possible to rely on fair use for commercial purposes, since whilst not prohibited by 17 US. US Code s.107, it is nonetheless a factor to be taken into account by the courts in assessing whether to apply the doctrine in any given scenario.

6.1.1 Authors Guild v. Google, Inc – the Google Books Saga

In dealing with the commercial aspect, a US case between the Authors Guild and Google (colloquially, the “Google Books” case), presiding Judge Chin assessed the overall benefits of the project before considering the four fair-use factors. The Google Books case involved TDM-style activities conducted by Google on over 20 million books to create an electronic database of books and text which was made available for online searches through the use of extracts and keywords. Although users can search the full text of all of the books, it is not possible to view a complete copy of an extract-view book. The Judge found that Google's use of copyright-protected content was highly transformative - it is not a tool to read books, but rather a tool to search them. Furthermore, whilst full text is scanned, the amount of text displayed is ultimately limited in response to a search, and the search scans are not sold – only the output is made available, and this is not necessarily to the detriment of the authors – indeed

---

199 US District Court Southern District of New York, Opinion, 05 Civ. 8136 (DC).
201 Ibid.
the output is likely to actually increase book sales\textsuperscript{202}. Ultimately, the Judge considered that Google’s activities provided significant public benefits, and therefore ruled in favour of the tech giant\textsuperscript{203}. Following on from this decision, and in the words of Professor Matthew Sag:

\textit{“It is now absolutely clear that technical acts of reproduction that facilitate purely non-expressive uses of copyrighted works such as books, manuscripts and webpages do not infringe US copyright law”\textsuperscript{204}.}

One can clearly therefore see how this decision can support TDM activities for commercial purposes when the operators seek to rely on fair use and depending on how source data is mined and displayed. Google is clearly a company operating for profit and regardless of the motive of the Google Books project (which was not for direct commercial gain), Google would either directly, or indirectly make profit by putting such a solution on the market since it enhanced Google’s market leadership in the online-search business\textsuperscript{205}.

Perhaps in light of this decision, the EU could therefore have considered implementing a broad fair use defence for TDM activities into the DSM Directive to provide a wider scope for commercial and non-commercial TDM operations so that innovators are able to more freely operate in the EU, and also to create an environment which is on a level playing field with the US, arguably one of Europe’s main competitors in the tech-space.

\textbf{6.1.2 Fair Use – Not Without Problems}

One of the key justifications for the EU not implementing a fair use exception is the inherent uncertainty of the doctrine.Whilst the playing field for TDM activities in the DSM Directive is fraught with caveats and restrictions, which in themselves create an uncertain TDM environment, at least there is some certainty as to the legal position for all TDM stakeholders whereas the fair use doctrine is fraught with issues, one not least because fair use analysis is necessarily fact specific given the wording of the legal text on which it originates. Indeed, Harvard Professor Larry Lessig has stated that “fair use is the right to hire a lawyer”\textsuperscript{206} which

\begin{thebibliography}{99}
  \bibitem{202} Ibid.
  \bibitem{203} Ibid.
  \bibitem{204} Google Books held to be fair use, Authors Guild v. Google: library digitization as fair use vindicated, again, Matthew Sag, 14 November 2013, available at: https://matthewsag.com/googlebooks-decision-fair-use/ accessed on 27 April 2019.
  \bibitem{206} Ibid, Professor Larry Lessig.
\end{thebibliography}
indicates that each party must argue the points in each-fact specific case once litigation is contemplated, which is can be prohibitively costly for start-ups and SMEs and does not provide certainty for stakeholders.

Ultimately, a justification as to the DSM Directive’s restrictive approach to TDM when compared to fair use has been summarized by expert IP lawyer, Rosie Burbidge, who considers that:

“[T]he problem with fair use is that it’s an uncertain test which provides a lot of scope for judicial interpretation and essentially kicks the problem down the road. More restrictive standards such as those in the EU may appear to be less business friendly but actually force businesses to engage with copyright at an earlier stage and provide more certainty in the long run”207.

Therefore, when considering the furtherance of innovation in the EU, and the need for legal certainty, it is arguable the EU did indeed take the correct approach when one considers the US-doctrine of fair use as an alternative to the current TDM position as set out in the DSM Directive, which does not necessarily leave the EU as a wholly uncertain environment, but does put a high burden on the relevant actors to navigate the scope of the provisions.

6.2 A Broad-Ranging Exception – The Japanese Example

Japan is a global epicentre for innovation committed to innovation208. There are similarities between the goals of Japan and the EU, but the approaches to these goals are very different from a legislative perspective. Japan’s copyright laws have permitted Machine Learning techniques since 2009. It was the first country in the world to update its copyright laws to enable TDM by introducing Article 47(7) into the Japanese Copyright Act209. This provision specifically authorised broad TDM activities through the creation of an exception to a copyright holder’s exclusive rights, for information analysis, comparison or classification or statistical analysis, with no restriction on beneficiaries210.

209 Ibid.
Legislation almost a decade old, and it is evident that Japan benefitted from more appropriate legal environment for innovation through TDM than is currently the case in the EU. Despite Japan’s favourable TDM environment created in part through copyright legislation, many stakeholders were still concerned that original Japanese TDM-exception was too restrictive and subject to legal uncertainties and would soon be superseded by technology. This meant Japanese law was in danger of becoming obsolete due to the rapid growth of IoT, AI, Big Data and robotics, particularly as concerns reproduction, use of databases and storage of works – all elements which have also been problematic in the EU, and which the DSM Directive has also sought to address.

To respond to these concerns, the Japanese Copyright Act was updated with effect from 1st January 2019 to permit additional flexibility and legal certainty for innovators, and to enhance the already TDM-favourable environment. This is reflected in the current wording of the Japanese Copyright Act which incorporates amendments in the interests of TDM and addresses the potential risks that copyright poses for innovation by permitting all users the right to: (i) analyse and understand copyrighted works for machine learning purposes; (ii) make and retain incidental electronic copies of works; and (iii) use copyrighted works for data verification. Ultimately, this legislative amendment evidences the Japanese legislator’s understanding of the technical processes required to operate and develop TDM. Essentially, the provisions take into account the fact that most often, TDM requires accessing data or information in a form where the copyrighted expression of the works is not perceived by the user, and then processing raw data via computer programs to carry out deep learning activities, forming the basis of AI.

Given the timing of the Japanese amendment, and the global nature of innovation and technology, it is regrettable the EU did not follow in the footsteps of its Asian counterpart and use the TDM-favourable provisions as inspiration for its own legislation. Indeed, Japan has now ensured that copyright cannot be an obstacle to the furtherance of its digital society by providing provisions which acknowledge rightsholder rights, but which also acknowledge that actual

---


212 Ibid.


processes undertaken through TDM for machine learning activities do not ultimately cause any harm to the rightsholders. In comparison, the caveated provisions in EU legislation puts Europe at a competitive disadvantage as it attempts to be at the forefront of innovation – to avoid liability, it is easier for TDM operators to operate elsewhere, such as Japan. However, as the DSM Directive is yet to be implemented into Member State national laws, it remains to be seen whether this concern is merely theoretical of whether such fears will materialise.

By way of concluding remark, it is worth noting that regardless of how innovative or restriction-free laws may be in other territories (by way of wide-ranging exception in Japan, or by the application of a fair-use doctrine extended to TDM, as is the case in the US), Europe’s approach to TDM nonetheless hinders the application of international legal regimes and the effectiveness of such broad-ranging TDM exceptions. The nature of copyright means the protection may arise automatically on creation of the copyrightable work. This can occur anywhere in the world. TDM activities are often conducted in an online environment, which generally knows no jurisdictional borders. Therefore, if such TDM activities are carried out lawfully in one jurisdiction in reliance on that jurisdiction’s broad exception/limitation, there is nonetheless a risk of copyright infringement when the mined source data contains European copyrighted works, works owned by European copyright-holders or when TDM carried out on copyrighted material hosted on EU-based servers. The EU legal regime governing TDM arguably therefore curtails global innovation when international TDM involves EU-copyrighted source material, regardless of the liberal TDM-approach in other jurisdictions. Unfortunately, further in-depth analysis of the potential extra-territorial scope of EU’s TDM exceptions/limitations and its impact on non-EU TDM innovators, is outside the scope of this paper.

6.3 Making the Best of Things

Whilst it may have been desirable for the EU to implement at TDM exception akin to the provisions that we can find in Japan, we now have the DSM Directive, which whilst not perfect for ensuring the furtherance of innovation, the law in the EU is “as is”.

It will therefore be important for Member States and innovators to maximise the benefits of this law, even with its caveats and restrictions. This is more so the case given the time that it takes for the EU to legislate in such areas – the last copyright reform for the digital environment took

---

217 Ibid.
place in 2001\(^\text{218}\), and we have now been presented with the DSM Directive almost two decades later. Even in Japan, it took a decade for the already-favourable TDM provisions to be revised in accordance with current technological advances. It therefore seems that we will need to make peace with the EU’s current TDM provisions, as any further changes are unlikely to be forthcoming in the immediate to near future. This is a classic challenge with law and the speed at which technology progresses – even during the time of the legislative process, and then the implementation of EU laws by Member States, the legal provisions that have been so hotly debated and carefully drafted can quickly become redundant or out of date. The DSM Directive is not alone in this dilemma.

It therefore falls on Member States’ national laws, courts and practitioners to find appropriate and innovative ways to apply law to new facts and circumstances, which occurs faster than the creation of new or updated legislation. We can therefore look to practice in Europe and hope that courts and stakeholders take a sensible approach to TDM and copyright under Articles 3 and 4 of the DSM Directive, so that innovation is neither hindered nor prohibited in the region.

6.3.1 Practical Suggestions to Overcome Existing Barriers to TDM in Europe

6.3.1.1 Education and Raising Awareness

There is a lack of awareness regarding TDM technologies\(^\text{219}\). Whilst we accept that authors’ rights deserve protection, we agree with Jockers et al. and consider that utilising copyrighted works for non-expressive uses, such as basic searching and text mining, is a separate issue and should not be barred on the basis of concerns over copyright\(^\text{220}\). We therefore consider that it is important to educate society between the differences of TDM and other uses of copyrighted works, such as unauthorized music-file sharing, which can infringe copyright because humans experience those files as musical works. Through education, it will be possible to teach stakeholders into understanding that scanning texts for insights or AI does not interfere with the author rights\(^\text{221}\). Through awareness-raising, rightsholders may then voluntarily choose to disapply any restrictive provisions of the DSM Directive.

\(^{218}\) i.e., the InfoSoc Directive.
\(^{219}\) The Social, Political and Legal Aspects of Text and Data Mining (TDM), Murray-Rust Brook, Oppenheim, D-Lib Magazine, November/December 2014, Volume 20, Number 11/12, p. 6.
\(^{220}\) Don’t let copyright block data mining, Matthew L. Jockers, Matthew Sag and Jason Schultz, Comment, 4 October 2012, Vol. 490, p. 3.
\(^{221}\) Ibid.
6.3.1.2 Lawful Access Solutions - Closing the Value Gap

Firstly, on the basis of education and awareness, rightsholders could be encouraged to rely on open access models for the works, such as the Creative Commons or “CC” licences. CC licenses are licences for the public which can be applied by rightsholders to copyrighted works to indicate to the public what can and cannot be lawfully done with their work. We could envisage a CC licence which specifically permits TDM activities when such processes are applied automatically and are not likely to cause: (i) any detriment to the rightsholder; or (ii) any loss from the “normal” exploitation by the rightsholder of the copyrighted work, but then which contractually reserves any other exploitation which the rightsholder considers appropriate. Indeed, we have already seen an increase in the update of CC licences, and so they could feasibly be used as a mechanism to encourage a broader, lawful access to works for TDM.

A second possibility could be the imposition on rightsholders compulsory licensing to content for TDM activities, particularly in areas where the rightsholder considers that his/her content includes information which is key for the furtherance of innovation and/or is important for public interest matters or the furtherance of society in general. Quasi-compulsory licensing via declarations by rightsholders has occurred in other areas of IP involving technology and telecommunications when dealing with “standard-essential patents” which are essential for the manufacture of standard-compliant products such as smartphones via declarations by patent-owners to licence such technology by way of an irrevocable commitment on fair, reasonable and non-discriminatory terms. Whilst this is narrower than a copyright holder as against the world at large, especially since unlike patents, copyright arises automatically on creation and requires no registration, and so the existence of such “essential works” would be extremely hard to monitor, it may be possible that a system can be implemented encourage sharing of works. Yet, it would be very difficult to impose such requirements on rightsholders without balancing their interests and providing something in return.

We can therefore envisage a proportional remuneration system for the owners of copyrighted source work which is subject to TDM when the source output in reliance of a small extract of

---

223 Don’t let copyright block data mining, Matthew L. Jockers, Matthew Sag and Jason Schultz, Comment, 4 October 2012, Vol. 490, p. 5.
an owner’s work results in financial gain for the TDM operator or overall user of the output. A solution similar to the class-action settlement in 2008 in the Google Books case could be envisaged, whereby Google, as a TDM innovator, was able to continue to scan electronic books in return for sharing advertising revenue with the copyright holders. Whilst this solution may be hard to implement, this possibility would not be contrary to the spirit of the DSM Directive in general which aims to close the value gap and create a fairer playing field between the creative community and tech-giants.

As technology advances, we can also envisage a compulsory plug-in to TDM technology which allows for TDM operators to log and track the owners of mined source data and then provide for remuneration related to the exploitation of the results as linked to the specific section of work mined. Whilst the value in TDM is generally linked to the mass of data involved rather than one individual piece of work, this could be an avenue to explore to make copyright holders more comfortable in making their works freely available or indeed by providing blanket consent for TDM purposes. Technology to track contributions to music enabling rightsholders to take control and close the value gap between record labels and artists already exists, and whilst it may take time to develop and implement for disseminated copyrighted works in general, the technology could have great value if it helps to provide innovators with peace of mind as regards potential copyright infringement claims.

6.3.1.3 Managing Barriers Created by the Application of TPMs

As discussed, the implementation of TPMs which act as a technological barrier to processing copyrighted works present a barrier to TDM activities. TPMs may apply to works as a whole and prevent all manner of activities conducted on those works, even where they are permissible in certain instances, which therefore means that where a lawful exception of limitation permits a specific use of those works, that use is curtailed by the TPM applied to the work. Even if rightsholders are bound to ensure the exceptions or limitations are available to the beneficiary in spite of any TPMs, this is hard to monitor. Furthermore, third-party circumvention of those technical measures applied by copyright holders renders those third-parties potentially open to liability. Whilst an in-depth discussion of TMPs is outside of the scope of this paper, they

---

225 Don’t let copyright block data mining, Matthew L. Jockers, Matthew Sag and Jason Schultz, Comment, 4 October 2012, Vol. 490, p. 3.
227 For example, the “Auddly” solution (https://auddly.com accessed on 28 April 2019).
228 For example, see Article 6 of InfoSoc.
nonetheless merit some attention. In our view, it will be important to rely on the education of rightsholders and TDM operators alike so that they can apply TPMs accordingly and arrange for caveats to be made for TDM purposes, such as API-integrations or other technical solutions within the TPMs applied to the copyrighted works which are able to communicate, and which could automatically disapply the TPM when it recognises actions of TDM processing (as opposed to other kinds of copying or reproduction of copyrighted works) performed on source data.

Another mechanism to ensure fully operational TDM in light of TPMs is to create monitoring and enforcement bodies, such as Hadopi in France, which is an agency created to oversee technical protection mechanisms and identify copyrighted works, and to ensure the use of such TPMs does not encroach on the benefit of exceptions and limitations\(^\text{229}\). We can envisage the creation of a similar Europe-wide agency tasked with a TDM-specific mandate tailored to the research and non-research TDM provisions as set down in the DSM Directive.

Finally, the courts must apply a purposive approach to issues of copyright infringement and the application of TPMs and the circumvention of such measures in limited circumstances. The courts must play their role in the furtherance of the digital economy and should recognize explicitly that, in the digital age, copying for non-expressive purposes is not infringement of copyright\(^\text{230}\). Judges should fill the gap between the goals of copyright law, innovation and legislation in light of any technological disconnection that exists in the DSM Directive, from new technologies, or from EU’s misunderstanding of TDM during the legislative process to ensure innovation is not hindered\(^\text{231}\). We consider that this is the approach the CJEU and Member State national courts will take when dealing with issues arising out of the interpretation and application of Articles 3 and 4 of the DSM Directive generally. Whilst this is immediately undesirable since matters will be dealt with on an \textit{ad hoc} basis, which by its nature creates uncertainty, as a framework of case law develops, we will be able to more safely rely on the TDM exceptions as we commit to a future under the DSM Directive.

\(^{229}\) See Articles L. 331-13 and L. 331-31 of the CPI

\(^{230}\) Don’t let copyright block data mining, Matthew L. Jockers, Matthew Sag and Jason Schultz, Comment, 4 October 2012, Vol. 490, p. 3.

7. Conclusion

There are both positive and negative aspects to the TDM exceptions in the DSM Directive which indicate how the EU has approached technological advances and innovation generally through legislation, case law and for TDM specifically.

On one hand, the EU has strong commitment to the Digital Agenda and wishes to push the EU to the forefront on a global scale but has failed to create all-encompassing copyright framework for TDM. However, it would be unjust to state that the EU approach to TDM and copyright constitutes a full hindrance to innovation in the region as there are several factors which must work together to create a positive environment for start-ups and innovators within the research and non-research sectors, such as access to funding, available talent and knowledge sharing, as well as an appropriate copyright framework.

At the time of writing, the exact future of TDM and innovation in Europe is unknown, and further research will be required over the next few years to ascertain Europe’s market share for TDM, AI and Machine Learning on a global scale. An analysis as to whether the TDM copyright environment in Europe has had a measurable impact on Europe’s success must also be conducted - currently it is hard to confirm whether the TDM limitations and exceptions will be a furtherance or a hindrance to innovation within the EU as the DSM Directive is so recent. Such issues must be monitored closely as the DSM Directive is transposed into Member State laws.

As this paper only deals with issues related to copyright, it would also be necessary to study other related areas, such as TPMs, the application of sui generis database right232 and the GDPR233 to gain a full overview of the EU approach to, and the impact of its legislation on innovation and TDM in the region.

Further research could also be conducted on the clash between the EU-approach and the international community to see if the EU legislation has an impact on innovation globally, for example, with an analysis of the DSM Directive’s extra-territorial application should TDM be

---

232 See Directive 96/9/EC.
233 Regulation No. 2016/679/EC which came into force in all Member States on 25 May 2018.
carried out by foreign actors on source data copyrightable in the EU, hosted within the EU or owned by EU-based rightsholders.

Ultimately, Europe’s approach to copyright alone has not, and will not, continue to be a threat to innovation in the region. Such impact, to the extent it exists in a positive or negative sense, is likely a result of the EU’s highly regulated environment generally, where individual rights are held above those of start-ups or tech-giants – a positive and negative consequence depending on the viewpoint of each stakeholder.

What is clear from this paper is that Europe must be mindful to the future and to our innovators. In a closing remark from the Founder of a UK-based data and analytics company:

“To not have the freedom to access information without infringing on IPRs data science and machine learning would be detrimental to our business and quite frankly stop, or make innovation extremely hard, thus affecting the European tech and start-up economy as a whole”234.

Something which surely the EU wishes to avoid.

---

8. Bibliography

Books and Interviews

- Interview with Rosie Burbidge, IP Partner at Gunner Cooke, London, 27 April 2019
- Interview with Maryam Mazraei, Founder at Autopsy, London, 5 April 2019

Journals

- Don’t let copyright block data mining, Matthew L. Jockers, Matthew Sag and Jason Schultz, *Comment*, 4 October 2012, Vol. 490
- The Social, Political and Legal Aspects of Text and Data Mining (TDM), Murray-Rust Brook, *D-Lib Magazine*, November/December 2014, Volume 20, Number 11/12
- Big data and data protection (GDPR and DPA 2018), by Richard Kemp, Partner, Kemp IT Law and Practical Law Data Protection, *Practical Law Company*, Thomson Reuters, Resources ID w-017-1623, 2019

Scientific Papers, Publications and Government Issued Guidance


• Modernising copyright: a modern, robust and flexible framework. UK Government response to consultation on copyright exceptions and clarifying copyright law, *HM Government*, 2012


• Voluntary Memorandum from the *UK Department for Business, Innovation and Skills*, 2 July 2014, available at: https://publications.parliament.uk/pa/jt201415/jtselect/jtstatin/13/1321.htm and accessed on 12 May 2019


*Internet Articles*


• What are Creative Commons Licences? Wageningen University and Research, 2019, available at: https://www.wur.nl/en/article/What-are-Creative-Commons-licenses.htm and accessed on 12 May 2019
• Content and Influence Marketing, Measure conversation volume and sentiment, find key influencers and identify topics that resonate with target customers, all in real time, Ubermetrics, 2019, available at: https://www.ubermetrics-technologies.com/influencers-content-marketing/ and accessed on 12 May 2019
• EU Copyright Reform/Expansion, Julia Reda, 2019, available at: https://juliareda.eu/eu-copyright-reform/ and accessed on 12 May 2019
• The Text of Article 13 and the EU Copyright Directive has just been finalized, Julia Reda, 2019, available at: https://juliareda.eu/2019/02/eu-copyright-final-text/ and accessed on 12 May 2019
• Copyright Law Overview, Clarin, 2019, available at: https://www.clarin.eu/content/clc-overview-copyright-law and accessed on 12 May 2019
• Text and Data Mining (TDM) exceptions in the UK and France, Clarin, 2019, available at: https://www.clarin.eu/content/clc-text-and-data-mining-tdm-exceptions-uk-and-france and accessed on 12 May 2019

Communications and Press Releases

• A Digital Agenda for Europe, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, 2010, COM/2010/0245 f/2
• Towards a thriving data-driven economy, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, 2014, COM/2014/442 final
• A Digital Single Market Strategy for Europe, A Digital Agenda for Europe, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, 2015, COM/2015/0192 final
• Towards a modern, more European copyright framework, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, 2015, COM/2015/0626 final

Legislation and Legislative Documents and International Treaties

• Berne Convention for the Protection of Literary and Artistic Works, Article 9
• The Agreement on Trade-Related Aspects of Intellectual Property Rights, Article 13
• World Intellectual Property Organization Copyright Treaty, Article 10

EU Directives, Regulations and Related Documents:

• Directive 96/9/EC of the European Parliament and the Council of 11 March 1996 on the legal protection of databases, Articles 3(1), 6(1), 8 and 15
• Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure
• Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

National Legislation and Legislative Documents:

• UK - Copyright, Designs and Patents Act 1988, as amended (Sections 29A (2), (3) and (4), 31 and 178)
• France - Law No. 2016-1321 of 7 October 2016 for a Digital Republic (Loi pour une République numérique) (in French only)


• France - Senate, Monitoring implementation of the Law for a Digital Republic (Contrôle de l'application de la loi pour une République numérique, Sénat) 31 March 2019, available at: https://www.senat.fr/application-des-lois/pjl15-325.html (in French only).


• USA - 17 US. US Code, Section107

Case Law

CJEU Case Law and Opinions:

• Case C-476/01 (Kapper)
• Case C-36/05 (Commission v Spain)
• Case C-5/08 (Infopaq International A/S v. Danske Dagblades Forening)
• Case C-604/10 (Football Dataco Ltd & ors v. Yahoo! UK ltd & ors)
• Cases C-457/11 to C-460/11 (VG Wort)
• Case C-30/14 (RyanAir)
• Opinion in Case C-5/08 of Advocate General Trstenjak delivered on 12 February 2009

National Case Law:

• UK - Football Dataco Ltd v Brittens Pools Ltd [2010] EWCA Civ 1380; Football Dataco Ltd v Brittens Pools Ltd [2010] EWHC 841 (Ch); Dataco Ltd & Ors v. Yahoo! UK Ltd [2012] EWCA Civ 1696.
• Germany - Bundesgerichtshof ruling in case I ZR 177/13 of November 17, 2014.
• USA - US District Court Southern District of New York, Opinion, 05 Civ. 8136 (DC).
9. Appendix 1