Check for updates

REVIEW

REVISED A literature review of "lawful" text and data mining.

[version 2; peer review: 1 approved, 2 approved with reservations]

Giorgos Vrakas

Law, University of Cyprus, Nicosia, Nicosia, Cyprus

v2 **First published:** 24 Jul 2024, **4**:153
https://doi.org/10.12688/openreseurope.18013.1
Latest published: 30 Oct 2024, **4**:153
https://doi.org/10.12688/openreseurope.18013.2

Abstract

Text and data mining (TDM) is a process, typically automated, that looks for patterns in data that may otherwise remain unnoticed. In a world where data driven solutions play a progressively more important role, TDM has become a vital tool in sectors ranging from medicine, to commerce, gaining widespread attraction. Nevertheless, a variety of regulatory frameworks not always specifically attuned towards regulating TDM continue to apply concurrently. The literature within the context of regulatory frameworks governing TDM is a fragmented piecemeal of valuable insights into what "lawful" TDM resembles.

This literature review adopts a grounded theory approach analysing 88 pieces of literature, collating views regarding "lawful" TDM, ultimately providing a holistic assessment of academics' and practitioners' views and opinions regarding the regulatory framework which governs TDM. A total of 7 categories were identified and each of these are analysed. Tables are provided in the Appendix (accessible here: https://doi.org/10.5281/zenodo.12654691)highlighting which scholarly works were used for each section of the literature review, but also how those works were used.

It is ultimately concluded that the regulatory frameworks that apply to users conducting TDM are multifaceted, and ever-changing on a case-by-case basis. There is an ever-growing need for a holistic interpretation of the regulatory frameworks which apply, creating a map which would allow for users conducting TDM to navigate this complex web of legal rules.

Plain language abstract

Text and data mining (TDM) is an automated method used to find hidden patterns in information and is often used to generate new knowledge. TDM is used widely today but the laws that apply to TDM are unclear. Similarly, research examining the laws that apply to TDM offer snippets of valuable insights.

Open Peer Review

Approval Status ? ✓ ?

	1	2	3
version 2 (revision) 30 Oct 2024		✓ view	
version 1 24 Jul 2024	? view	? view	? view

1. **Thomas Margoni**, KU Leuven, Leuven, Belgium
2. **Ana Nordberg**, Lund University, Lund, Sweden
3. **Pablo Rodrigo Trigo Kramcsak**, Vrije Universiteit Brussel, Brussels, Belgium

Any reports and responses or comments on the article can be found at the end of the article.

This paper examines 88 pieces of literature which study the laws that apply to TDM, providing a more complete image of academics' views on this topic.

What this paper shows is that the laws for people doing TDM are complicated and change depending on the situation. There is a growing need for a complete understanding of these laws, to help people doing TDM find their way through this tricky set of laws.

Keywords
Text and data mining, copyright, data protection, contract law, human rights



This article is included in the Horizon 2020 gateway.



This article is included in the Marie-Sklodowska-Curie Actions (MSCA) gateway.



This article is included in the Research on Research gateway.

Corresponding author: Giorgos Vrakas (vrakas.giorgos-dimitris@ucy.ac.cy)

Author roles: Vrakas G: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101034403.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2024 Vrakas G. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Vrakas G. **A literature review of "lawful" text and data mining. [version 2; peer review: 1 approved, 2 approved with reservations]** Open Research Europe 2024, **4**:153 https://doi.org/10.12688/openreseurope.18013.2

First published: 24 Jul 2024, **4**:153 https://doi.org/10.12688/openreseurope.18013.1

REVISED **Amendments from Version 1**

In response to the valuable feedback provided by the reviewers, this paper has undergone revisions to enhance its structure, clarity, and depth of analysis.

Firstly, the introduction has been restructured to provide a clearer explanation of key concepts, which are essential to understanding Text and Data Mining (TDM). Similarly, the selection criteria for the literature review have been clarified, including language limitations and the interdisciplinary scope of the review.

The regulatory framework section has been expanded to include a broader discussion of the legal challenges surrounding TDM, particularly in relation to data protection. New authoritative sources recommended by the reviewers, such as official guidelines from supervisory bodies, have been included to enhance the discussion on GDPR and other regulatory frameworks.

The section on intellectual property rights has been renamed "TDM and Copyright and Related Rights" for greater accuracy. In response to concerns about outdated references, older sources have been justified or removed as appropriate.

The data protection section now includes more authoritative sources, and a deeper analysis of GDPR provisions such as data protection by design and default, profiling, and inferred data.

Further, the section on fundamental rights has been revised to address issues of bias and discrimination in TDM, with key literature added, while the generalist approach section has been clarified to justify its classification.

Footnotes and references have been reviewed and revised for consistency and accuracy throughout the paper.

Any further responses from the reviewers can be found at the end of the article

Introduction

"*Each day on Earth we generate 500 million tweets, 294 billion emails, 4 million gigabytes of Facebook data, 65 billion WhatsApp messages and 720,000 hours of new content added daily on YouTube*".[1]

This "information revolution" has resulted in vast amounts of data, in various formats, being continuously generated and stored and is accurately termed "big data".[2] As per the European Commission "*digital innovation, driven by the combination of Big Data, Cloud Computing, Mobile technologies and Social media, is one the most powerful drivers of change and the best opportunity for Europe to move back to a growth path*".[3] Hence, the utilisation of big data is regarded as valuable to the European economy.

For some the word "data" correlates with "*an organised set of values*".[4] Nevertheless, data is often messy, and unstructured and cannot effectively be utilised.[5] Consequently, it is often the case that to analyse "big data", unstructured messy data must be transformed into structured data.

Text and data mining (TDM) is a process, typically automated, that looks for patterns in data that may otherwise remain unnoticed, often transforming raw unstructured data into more understandable and useful information.[6] Generally TDM comes in a variety of different shapes and sizes varying on the type of data being used, data gathering techniques, analysis techniques employed, and presentation of data, but involves a common underlying objective. That is, utilising vast amounts of data to extract knowledge that would otherwise remain unnoticed.[7] Typically, TDM involves a series of stages such as document retrieval, document transformation (transforming the data into machine readable format), data extraction, (extracting the information which seems most useful), and finally mining that information resulting in new knowledge discovery.[8]

In a world where data driven solutions play a progressively more important role, TDM has become a vital tool in sectors ranging from medicine,[9] to commerce,[10] gaining widespread attraction. Nevertheless, a variety of regulatory frameworks not always specifically attuned towards regulating TDM continue to apply concurrently. As such, compliance with applicable regulatory frameworks is not always obvious, meaning unlawful TDM may be taking place in the shadows of the law.

This literature review forms part of a wider research project which aims to facilitate "lawful" TDM, by highlighting the aspects of regulatory frameworks which concurrently apply to TDM, enabling users to conduct regulatory compliant TDM. This research will ultimately result in the creation of a practical

[1] Melvin M Vopson, 'The World's Data Explained: How Much We're Producing and Where It's All Stored' (*The Conversation*) <http://theconversation.com/the-worlds-data-explained-how-much-were-producing-and-where-its-all-stored-159964> accessed 23 June 2022.

[2] Jonathan Grossman and Ami Pedahzur, 'Political Science and Big Data: Structured Data, Unstructured Data, and How to Use Them' (2020) 135 Political Science Quarterly 225. The definition of "big data" is often attributed to Doug Laney, who presented the idea of the three V's. Volume (amount of data), Velocity (rate at which data is created) and Variety (variance of data format). In accordance with this definition data can be termed "big" if all three V's are high. See Doug Laney, '3D Data Management: Controlling Data Volume, Velocity, and Variety' (META Group 2001).

[3] European Commission, 'European Data Market Study (SMART 2013/0063)' <https://digital-strategy.ec.europa.eu/en/library/european-data-market-study-update> accessed 24 June 2022.

[4] Grossman and Pedahzur (n 2).

[5] ibid.

[6] James Caudwell, 'LibGuides: Text & Data Mining: What Is TDM?' <https://libguides.cam.ac.uk/tdm/definitions> accessed 7 November 2023; 'Text and Data Mining at Springer Nature' <https://www.springernature.com/gp/researchers/text-and-data-mining#:~:text=TDM%20(Text%20and%20Data%20Mining,can%20provide%20valuable%20information%20needed> accessed 7 November 2023.

[7] Caudwell (n 6); Marti A Hearst, 'Untangling Text Data Mining', *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (Association for Computational Linguistics 1999) <https://dl.acm.org/doi/10.3115/1034678.1034679> accessed 18 January 2024; Sonali VijayGaikwad, Archana Chaugule and Pramod Patil, 'Text Mining Methods and Techniques' (2013) 85 International Journal of Computer Applications.

[8] Caudwell (n 6).

[9] Thilla Rajaretnam, 'Data Mining and Data Matching: Regulatory and Ethical Considerations Relating to Privacy and Confidentiality in Medical Data' (2014) 9 Journal of International Commercial Law and Technology 294.

[10] Douglas M Kochelek, 'Data Mining and Antitrust' (2009) 22 Harvard Journal of Law & Technology 515.

guide and tool enabling users who conduct TDM to navigate the complex web of legal rules which make up the regulatory framework governing TDM.

Just as the regulatory framework, the literature within the context of regulatory frameworks governing TDM is a fragmented piecemeal of valuable insights. This literature review collates views regarding the regulatory frameworks governing TDM, providing a holistic assessment of academics' and practitioners' views and opinions regarding the regulatory framework which governs TDM.

To collate the relevant literature, keyword searches, using search operators were used on Google Scholar, Westlaw and LexisNexis. Namely, in October 2023 the keyword search used was ("Text and data mining") OR (("Text" OR "Data") AND ("mining" OR "mine")) AND ("law" OR "regulation" OR "legal").

This keyword search was used with the aim of encompassing all published literature which examined the regulatory framework governing TDM. Given the inherently legal nature of the topic, legal libraries like Westlaw and LexisNexis were used. Consequently, this review includes primarily legal literature. Nevertheless, in an attempt to widen this review's reach in order to provide a holistic approach, some non-legal literature was also analysed as part of this review. For instance, as will emerge from the discussion below, primarily the section examining "Data Protection and TDM", some computer science literature was also caught by this keyword search. However, given the interdisciplinary nature of this topic, and with the aim of providing a holistic assessment of academics' and practitioners' views and opinions regarding the regulatory framework which governs TDM, disciplinary boundaries were not set as selection criteria for this literature review.

The only selection criteria used for this literature review was the search query. One inherent limitation of doing so, is the exclusion of publications, not curated by the aforementioned libraries, which may result in some publications not being collected. Some literature which was not caught by the search query, was recommended by reviewers of this literature review. These pieces of literature were added so as to include as much relevant literature as possible. Additionally, this method of collecting literature to review, also excludes literature that has been published in the official languages of each jurisdiction. Nevertheless, given the difficulties and limitations associated with reviewing translated documents,[11] the English language was set as a selection criterion.

Considering the nature of TDM, it is inevitable that users conducting TDM encounter multijurisdictional regulatory frameworks. As such, Flynn *et al.* call for a harmonised international approach towards the regulation of TDM, asking for the World Intellectual Property Organization (WIPO) to play a "*constructive role in establishing a consistent international baseline*" for the regulation of TDM.[12] Nevertheless, as it stands users conducting TDM must navigate a complex web of regulatory frameworks spanning across multiple jurisdictions. Given the inherent difficulties of comprehensively examining the inner workings of lawful TDM on a global scale, this literature review will provide scholarly insights from the United States (US), European Union (EU), and United Kingdom (UK) providing insights where relevant from other jurisdictions.[13] Similarly, the practical guide and tool aimed at facilitating lawful TDM, created as part of this research's wider purpose, will take into account and comparatively analyse these jurisdictions.

The collected literature was updated periodically, up until May 2024 using the "sort by date" function identifying any newly published literature. Although literature examining the regulatory frameworks which apply to TDM is a relatively new topic of concern, this yielded a plethora of results. This is attributed to the returning of false positives such as mining in the geological sense. After qualitatively assessing results by examining their title and, if need be, their abreact, determining if the content examined regulatory frameworks which apply to TDM, a total of 88 pieces of literature were collected. These were then analysed manually adopting a grounded theory approach individually analysing and categorising them. For a list of all pieces of scholarly work analysed in this review please see Table 9 in the Appendix (accessible here: https://zenodo.org/records/13880018).

As indicated by Table 1 below, a total of 7 distinct categories were identified. Literature was not always exclusively under one of the categories, meaning that some could fall under two or more of the categories. Tables for each of the category, listing which pieces of scholarly work were used is included in the Appendix (Tables 2 – 8) accessible here: https://zenodo.org/records/13880018.

The structure of this literature review will follow the 7 categories identified. It begins by providing scholars' views regarding what TDM entails, and how their view of TDM ultimately shapes their respective analysis of the regulatory frameworks This review then presents academics' and practitioners' views of the regulatory framework governing TDM. Namely TDM's interactions with copyright and related rights, data protection and privacy rules, fundamental rights, contract law, and competition law are presented respectively. Finally, several authors adopt a broader, "generalist approach" to the legal analysis

[11] Bogusia Temple and Alys Young, 'Qualitative Research and Translation Dilemmas' (2004) 4 Qualitative Research 161; Elizabeth Peña, 'Lost in Translation: Methodological Considerations in Cross-Cultural Research' (2007) 78 Child development 1255.

[12] Sean Flynn and others, 'Implementing User Rights for Research in the Field of Artificial Intelligence: A Call for International Action' [2020] Joint PIJIP/TLS Research Paper Series <https://digitalcommons.wcl.american.edu/research/48>.

[13] See for instance, Pamela Samuelson, 'Text and Data Mining of In-Copyright Works: Is It Legal?' (2021) 64 Communications of the ACM 2.

Table 1. Categories of literature identified and the amount of literature within each of the categories.

Category	Description	Amount of Literature included in each section
Taxonomy of TDM	Identifying what TDM is and what it entails.	9
TDM and Copyright and Related Rights	How TDM interacts with copyright and database rights.	44
TDM and Data Protection	How TDM interacts with data protection regulations.	18
TDM and Fundamental Rights	How TDM interacts with fundamental human rights	6
TDM and Contract Law	Hurdles for TDM caused by contracts and licenses.	11
TDM and Competition Law	How TDM interacts with competition law	2
Generalist Approach	A generalist approach where authors provide a more holistic approach toward the regulatory framework applicable to TDM	8

of TDM. Rather than focusing on one particular area of law, these pieces of literature explore the intersection of multiple legal regimes and how they collectively influence the practice of lawful TDM. Hence the final section of this literature review presents literature that decided to present the applicable regulatory frameworks more holistically, adopting a so called "generalist approach".

A taxonomy of TDM

As aforementioned, TDM is a process, typically automated, uncovering knowledge from data that may otherwise remain unnoticed, often transforming raw data into more understandable and useful information.[14] In the literature, there is an undeniable articulation of the benefits that TDM has to offer, whether that be within the context of libraries,[15] scientific research and innovation,[16] medicine,[17] and generally its ability to yield economic advantages.[18] Wyner *et al.*, even introduce the idea of using TDM to extract information from case law databases thereby enhancing the legal profession, especially in common law systems which rely on case law.[19] Margoni argues that "*TDM, together with other types of data-driven analytical tools, deserves its own autonomous methodological classification as computational legal methods*".[20] Nevertheless, there does not seem to be a common understanding of TDM amongst scholars, both within and across different disciplines.

Some scholars make a distinction between text mining and data mining.[21] Data mining is sometimes seen as a subset of text mining since text mining is the computational process of analysing unstructured messy data, whilst data mining is the computational process of analysing structured data.[22]

Other scholars focus on a taxonomy and categorisation of TDM as a way of explaining what regulatory frameworks a user conducting TDM might encounter. For instance, Colonna emphasises the fact that data mining involves separate processes, and in different contexts, each of them with varying legal implications.[23] Similarly, Tóth argues that the process of TDM involves several stages each of which could potentially

[14] Caudwell (n 6); 'Text and Data Mining at Springer Nature' (n 6).

[15] Marinos Papadopoulos and others, 'Empirical Research on Web Harvesting in the Process of Text and Data Mining in National Libraries of EU Member States' (2019) 10 Open Journal of Philosophy 88; Liane Colonna, 'Opportunities and Challenges to Utilizing Text-Data Mining in Public Libraries: A Need for Legal Research' (2018) 65 Scandinavian Studies in Law 191.

[16] Nicolas Jondet, 'The Text and Data Mining Exception in the Proposal for a Directive on Copyright: Why the European Union Needs to Go Further than the Laws of Member States.' (2018) 67 Propriétés Intellectuelles 25; Eleonora Rosati, 'Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and Its Role in the Development of AI Creativity' (2020) 27 Asia Pacific Law Review 1.

[17] Rajaretnam (n 9); Sean M Fiil-Flynn and others, 'Legal Reform to Enhance Global Text and Data Mining Research' (2022) 378 Science 951.

[18] Jackson Bertón (n 13); C Handke, L Guibault and JJ Vallbé, *Is Europe Falling Behind in Data Mining? Copyright's Impact on Data Mining in Academic Research* (AmsterdamIOS Press 2015) <https://dare.uva.nl/search?identifier=1b29a7cf-aa67-47d9-80c7-b2fd790ffd87> accessed 8 November 2023.

[19] Adam Wyner and others, 'Approaches to Text Mining Arguments from Legal Cases' (2010).

[20] Thomas Margoni, 'Text and Data Mining in Intellectual Property Law: Towards an Autonomous Classification of Computational Legal Methods' [2020] CREATe <https://www.create.ac.uk/blog/2020/05/01/new-working-paper-text-and-data-mining-in-intellectual-property-law-towards-an-autonomous-classification-of-computational-legal-methods/> accessed 1 December 2023.

[21] Maarten Truyens and Patrick Van Eecke, 'Legal Aspects of Text Mining' in Nicoletta Calzolari and others (eds), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (European Language Resources Association (ELRA) 2014) <http://www.lrec-conf.org/proceedings/lrec2014/pdf/452_Paper.pdf> accessed 8 November 2023; Matthew Sag, 'The New Legal Landscape for Text Mining and Machine Learning' (2019) 66 Journal of the Copyright Society of the U.S.A. 291.

[22] Caudwell (n 6); VijayGaikwad, Chaugule and Patil (n 7).

[23] Liane Colonna, 'A Taxonomy and Classification of Data Mining' (2017) 16 SMU Science and Technology Law Review 309.

conflict with regulatory frameworks.[24] Carroll also methodically works through each type of copying that happens in the course of collecting, formatting, processing, and storing data for TDM purposes, analysing relevant case law for each stage.[25] Fiil-Flynn *et al.* also claim that TDM projects potentially face regulatory restraints at different stages such as digitisation and compilation of data and the subsequent application of algorithms.[26] Equally, Rosati identifies three general TDM processes each of which with their own legal consideration:

"*1) Access to content,*

2) Extraction and/or copying of content, and;

3) Mining of text and/or data for knowledge discovery".[27]

Likewise, Guadamuz and Cabell state that "*it is difficult to generalise on what exactly the method for content mining is, as there are different algorithmic and model structures depending on the subject, the type of database, and the type of analysis being performed*".[28] Nevertheless, the authors argue that most content mining roughly follows these steps:

"*1. Individual content is created.*

2. Content is placed into data set, repository or collection.

3. Miner gains access to the data.

4. Mining tools applied to the data set.

5. Analysis of the processed data.

6. New knowledge".[29]

Guadamuz and Cabell then argue that legal conflicts will arise at stages 3 and 4.[30] Hence, it can be deduced that the literature although divided on the substantive procedural aspects of TDM stages, there is a cohesive undertone in the understanding that TDM involves a series of steps, each of which have their own legal considerations.

Regulatory frameworks which govern TDM

Flynn *et al.* highlight the importance of being able to conduct TDM without disproportionate legislative barriers but argue that a legislative patchwork of regulatory frameworks spanning various jurisdictions may in fact hinder the progress of TDM.[31] This section aims to outline academics' and practitioners' views and opinions regarding the different regulatory frameworks which govern TDM. This is intended to act as a stepping-stone towards uncovering what "lawful TDM" i.e., TDM which would be considered as compliant with the applicable regulatory frameworks, resembles.

A variety of regulatory frameworks are examined when discussing "lawful TDM" such as copyright and sui generis database rights, data protection and privacy, contract law, competition law and fundamental rights. Some scholars take a generalist approach examining a variety of frameworks which interact with the act of TDM whilst others adopt a more targeted approach examining one regulatory interaction. This review will present authors' views systematically, presenting interactions with distinct regulatory frameworks separately. Subsequently, it will collectively discuss the views of authors who adopt a generalist approach.

TDM and copyright and related rights

Above, TDM was defined as a process, typically automated, that looks for patterns in data that may otherwise remain unnoticed, often transforming raw unstructured data into more understandable and useful information. Article 2(2) of the Directive on Copyright in the Digital Single Market (CDSM) defines TDM as "*any automated analytical technique aiming to analyse text and data in digital form to generate information such as patterns, trends and correlations*".[32] Likewise, Recital 8 of the CDSM defines TDM as "*the automated computational analysis of information in digital form, such as text, sounds, images or data' enabled by new technologies*".[33]

From these definitions it can be deduced that the act of TDM may conflict with intellectual property rights where the work or database being reused is protected by intellectual property rights. In the EU, such protection is afforded by Directive 2001/29 (InfoSoc Directive),[34] Directive 2009/24 (Software Directive),[35] or Directive 96/9 (Database Directive).[36] Other jurisdictions have their own regulatory frameworks for affording intellectual property rights. This section aims to present authors' views, drawing insights from a variety of jurisdictions beginning with the EU approach, as this seems to make up the bulk of scholarly work. Namely 29 out of the 42 pieces of literature analysed under this category examined how TDM interacts with EU copyright and database rights.

[24] Andrea Katalin Tóth, 'Algorithmic Copyright Enforcement and AI: Issues and Potential Solutions through the Lens of Text and Data Mining' (2019) 13 Masaryk University Journal of Law and Technology 361, 378.

[25] Michael W Carroll, *Copyright and the Progress of Science: Why Text and Data Mining Is Lawful* (UC Davis School of Law 2019) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3531231> accessed 8 November 2023.

[26] Fiil-Flynn and others (n 17).

[27] Rosati (n 16).

[28] Andrés Guadamuz and Diane Cabell, 'Data Mining in UK Higher Education Institutions: Law and Policy' (2014) 4 Queen Mary Journal of Intellectual Property 3, 5.

[29] ibid 5–6.

[30] ibid 6.

[31] Fiil-Flynn and others (n 17).

[32] Directive (EU) 2019/790, Article 2(2).

[33] Directive (EU) 2019/790, Recital 8.

[34] Directive 2001/29 on the harmonisation of certain aspects of copyright and related rights in the information society [2001] OJ L167/10.

[35] Directive 2009/24 on the legal protection of computer programs [2009] OJ L111/16.

[36] Directive 96/9 on the legal protection of databases [1996] OJ L77/20.

It is often argued that since data itself is not protected by intellectual property rights, the act of TDM should in theory not interfere with any intellectual property rights.[37] Recital 9 of the Directive on Copyright in the Digital Single Market (CDSM) acknowledges this by stating that TDM "*can also be carried out in relation to mere facts or data that are not protected by copyright, and in such instances no authorisation is required under copyright law. There can also be instances of text and data mining that do not involve acts of reproduction or where the reproductions made fall under the mandatory exception for temporary acts of reproduction provided for in Article 5(1) of Directive 2001/29/EC, which should continue to apply to text and data mining techniques that do not involve the making of copies beyond the scope of that exception*".[38] This however, is not always the case. In some instances, the act of copying involved with TDM will often encroach upon a rightsholders right to reproduce a work afforded to them through copyright protection.[39] For instance, when creative works are extracted, it is often the case that the entirety of that work is reproduced, resulting in infringement of the rightsholder's exclusive right of reproduction.[40] Nasr *et al.*, were able to show that generative Artificial Intelligence (AI), which utilises data mining mechanisms sometimes regurgitates its training data in its entirety.[41] Similarly, Baris who examined derivative musical works created by AI, states that "*where an AI-generated output reproduced protectable parts of a work that was used in the training of the AI system",* this would likely amount to copyright infringement.[42] On the other hand Senftleben, argues that at an international copyright law level, the act of TDM is not prohibited.[43] Awarding copyright protection is ultimately a balancing exercise of all stakeholders' interests. As will emerge from the discussion below, this balancing exercise becomes more complex in the realm of TDM.

The act of TDM may in some instances consist of an act of extraction and/or re-utilisation of a database protected in the EU by the sui generis database right.[44] As stated by Rosati "*if the content extracted and/or copied is included in a database, then both copyright and the sui generis (database) right – these being two rights that are independent of each other and may subsist together at the same time on the same database) – might come into consideration"*.[45] Hence, depending upon the circumstances under which data is accessed, how data is processed and presented, it is possible that intellectual property rights are infringed.

The majority of scholarship in this domain emanates from the applicability of copyright exceptions to the act of TDM, most notably with the recent introduction of the two new mandatory copyright exceptions within the EU jurisdictions under Articles 3 and 4 of the CDSM.[46]

In a pre-CDSM era, several Member States had already introduced their own TDM-specific copyright exceptions. The UK for instance, was the first Member State (in a pre-Brexit era) to have introduced their own TDM-specific exception back in 2014.[47] Section 29A of the Copyright, Designs and Patents Act 1988 introduced a text and data analysis exception for the purposes of non-commercial research.[48] Similarly, France introduced their own TDM exception applying to both works and databases.[49] Estonia and Germany also introduced their own TDM-specific exceptions.[50] Greece had also introduced its own rules regulating the act of web harvesting and archiving.[51] As per Geiger *et al.*, "*this regulatory patchwork* […] *led to a fragmented legal environment in the EU*", which ultimately led

[37] See for instance, Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko, 'Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?' (2018) 49 IIC - International Review of Intellectual Property and Competition Law 814, 817; Martin Kretschmer and Thomas Margoni, 'Data Mining: Why the EU's Proposed Copyright Measures Get It Wrong'; Thomas Margoni and Martin Kretschmer, 'A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology' (2022) 71 GRUR International 685; Reto Hilty and Heiko Richter, 'Position Statement of the Max Planck Institute for Innovation and Competition on the Proposed Modernisation of European Copyright Rules Part B Exceptions and Limitations (Art. 3 – Text and Data Mining)' (Max Planck Institute for Innovation & Competition 2017) Research Paper No. 17-02 para 3; Jonathan Griffiths, Tatiana Synodinou and Raquel Xalabarder, 'Comment of the European Copyright Society Addressing Selected Aspects of the Implementation of Articles 3 to 7 of Directive (EU) 2019/790 on Copyright in the Digital Single Market' (2023) 72 GRUR International 22.

[38] Directive (EU) 2019/790, Recital 9.

[39] See CJEU, C-5/08, *Infopaq International A/S v. Danske Dagblades Forening* (16 July 2009). ECLI:EU:C:2009:465, paras. 54–55 (finding that even an excerpt of 11 words might be protected).

[40] Signatories to the Berne Convention, which include all but 20 countries of the world will grant rightsholders the right to reproduce their work. See Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979), Article 9(1). In the EU this right is enshrined under Article 2 of Directive 2001/29.

[41] Milad Nasr and others, 'Scalable Extraction of Training Data from (Production) Language Models' (arXiv, 28 November 2023) <http://arxiv.org/abs/2311.17035> accessed 4 December 2023.

[42] Antonios Baris, 'AI Covers: Legal Notes on Audio Mining and Voice Cloning' [2024] Journal of Intellectual Property Law & Practice jpae029, 3.

[43] Martin Senftleben, 'Compliance of National TDM Rules with International Copyright Law: An Overrated Nonissue?' (2022) 53 IIC - International Review of Intellectual Property and Competition Law 1477.

[44] Directive 96/9 on the legal protection of databases [1996] OJ L77/20.

[45] Rosati (n 16).

[46] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.

[47] Regulation 3 of the Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014, No. 1372, adding Art. 29A to the Copyright, Designs and Patents Act 1988.

[48] Copyright, Designs and Patents Act 1988, section 29A.

[49] Art. 38 of Law No. 2016-1231 for a Digital Republic added paragraph 10 to Art. L122-5 and paragraph 5 to Art. L342-3 of the Intellectual Property Code (*Code de la proprie´te´ intellectuelle*) (CPI).

[50] Estonian Copyright Act, Art. 19(3), and Urheberrechtsgesetz § 60d respectively.

[51] For a discussion of this, see Maria Kanellopoulou - Botti and others, 'Legal and Technical Issues for Text and Data Mining in Greece' (2019) 2019 Computer Ethics - Philosophical Enquiry (CEPE) Proceedings <https://digitalcommons.odu.edu/cepe_proceedings/vol2019/iss1/11>.

to the introduction of the CDSM exceptions so as to create a harmonised level playing field in the EU.[52]

Scholars at this stage had called for the broadest possible EU TDM-specific exceptions so as to "*boost the international competitiveness of its knowledge economy*".[53] Once the dust had settled and the final version of the CDSM was published, Article 3 and 4 introduced two new mandatory exceptions arguably creating a new avenue for "lawful TDM". Article 3 mandates Member States to allow for cultural heritage and research institutions to freely use protected works for scientific text and data mining research purposes.[54] Article 4 necessitates the introduction of a general TDM exception, allowing other entities to develop algorithms without the burden of authorisation and remuneration payment, provided the right has not been expressly reserved by the rightsholder.[55] These exceptions were welcomed to a certain extent, for their potential to harmonise a collective EU-wide approach towards "lawful TDM", ultimately fostering competition.[56] For instance Papadopoulos *et al.*, who examined libraries' views and opinions of web-harvesting (a form of TDM), and found that most libraries already conduct such forms of TDM and those that do not indicate a desire to do so, concluded that the EU CDSM TDM-specific exceptions create "*a favourable legal foundation for the deployment of Web-harvesting and archiving operations through the national libraries of the EU Member States*".[57]

Nevertheless, authors have not strayed away from criticising these new set of mandatory EU TDM-specific exceptions. For instance, Ducato and Strowel present their arguments in a table which outlines all that these two exceptions got right, wrong, and also those sections which raise more questions.[58] They ultimately argue that Article 3 is too narrow with its limited scope of cultural heritage and research institutions whilst Article 4 is limited in the sense that it can be overridden by contracts.[59]

This limiting of scope in Articles 3 and 4 of the CDSM to content that is "lawfully accessed" is interpreted in Recital 14 as including content that is openly accessible online and materials made available to users through the conclusion of a contractual agreement.[60] Whilst Article 3 cannot be overridden by contracts, Article 4 can.[61] This has resulted in commentators characterising this as giving rise to an opt-out mechanism.[62] The European Copyright Society, in its comment addressing selected aspects of the implementation of Articles 3 to 7 of the CDSM, argued that as a result of this provision "*publishers might price TDM into their subscription fees and, as a consequence, few research organisations would be able to acquire licences for all databases that are relevant for a TDM research project*".[63] It may even result in biased AI systems where it becomes more "*economically attractive for developers to train their algorithms on older, less accurate, biased data, or to import AI models already trained on unverifiable data*".[64] This, arguably greatly limits the applicability of the exception, putting the EU at a competitive disadvantage when compared to jurisdictions like the US which will be discussed in greater detail below, and undermines the "*widespread assumption that the right to read should be the right to mine*".[65] In a blog post, Keller argues that the way in which the proposed, at the time, EU AI Act is being phrased explicitly reinforces this opt-out mechanism, with some of the big names in generative AI also having shown evidence that they intend to implement some form of opt-out mechanisms for rightsholders.[66] Keller however, argues that this creates a fragmented framework for rightsholders who have to individually opt out of all new generative AI machines.[67] Nevertheless, the author paints the EU's approach towards regulating AI, which inevitably regulates TDM as balanced, implying that this could in turn result in it being the next global standard.[68] Similarly, Senftleben argues that the EU AI Act could serve as a template for other regions promoting a global standard supporting human creativity in the age of AI.[69] Namely, Senftleben argues that the primary objective of the EU AI Act is to "*ensure that authors are properly remunerated for the use of their works in AI training processes*".[70] Senftleben sees this as a necessary

[52] Geiger, Frosio and Bulayenko (n 37) 830.

[53] Jondet (n 16); Kretschmer and Margoni (n 37).

[54] Article 3 of Directive (EU) 2019/790.

[55] Article 4 of Directive (EU) 2019/790.

[56] See for instance, Maria Bottis and others, 'Text and Data Mining in Directive 2019/790/EU Enhancing Web-Harvesting and Web-Archiving in Libraries and Archives' (2019) 09 Open Journal of Philosophy 369; Tóth (n 24).

[57] Papadopoulos and others (n 15).

[58] Rossana Ducato and Alain Strowel, 'Ensuring Text and Data Mining: Remaining Issues with the EU Copyright Exceptions and Possible Ways Out' (2021) 43 European Intellectual Property Review 322, 331.

[59] Ducato and Strowel (n 58).

[60] Recital 14 of Directive (EU) 2019/790.

[61] Directive (EU) 2019/790, Article 7(1) states that "*any contractual provision contrary to the exceptions provided for in Articles 3 […] shall be unenforceable*." See also, Luisa Gatti, 'The European Solution for Text and Data Mining: A Focus on the Opt-Out System Provided by Article 4 of the DSM Directive' (2021) 43 European Intellectual Property Review 767.

[62] Gatti (n 61); Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko, 'Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU' (2019); Nicolas Binctin, 'TDM: A CHALLENGE FOR ARTIFICIAL INTELLIGENCE' [2019] Revue In-ternationale du Droit d'Auteur.

[63] Griffiths, Synodinou and Xalabarder (n 37) 11–12.

[64] ibid 12; Margoni and Kretschmer (n 37).

[65] Geiger, Frosio and Bulayenko (n 62).

[66] Paul Keller, 'Generative AI and Copyright: Convergence of Opt-Outs?' (*Kluwer Copyright Blog*, 23 November 2023) <https://copyrightblog.kluweriplaw.com/2023/11/23/generative-ai-and-copyright-convergence-of-opt-outs/> accessed 27 November 2023.

[67] ibid.

[68] ibid.

[69] Martin Senftleben, 'AI Act and Author Remuneration - A Model for Other Regions?' (24 February 2024) <https://papers.ssrn.com/abstract=4740268> accessed 29 May 2024.

[70] ibid 1.

step in order to support human creativity, thereby preventing the market from being overrun by AI-generated content.[71]

Another issue that may be brought back to the limelight by this opt-out mechanism, is the interplay between copyright and technical protection measures (TPMs). TPMs allow for rightsholders to place architectural constraints on their works to regulate access. Under Article 6(1) of the InfoSoc Directive, TPMs are not to be circumvented, whilst under Article 6(4), rightsholders are to make exceptions for those uses which may be able to circumvent TPMs by falling under a copyright exception.[72] TPMs may be implemented as a way of utilising a rightsholder's ability to opt out from TDM.[73] Nevertheless, this will undoubtably create uncertainty as to when TDM is being carried out for the purposes of research, which may inadvertently result in an over-blocking of TDM for research purposes.[74] A study conducted by the Max Planck Institute for Innovation and Competition suggests that whilst rightsholders should be able to utilise TPMs to restrict TDM, "*such measures must not go beyond what is required, i.e. technical protection measures must be precluded from making TDM unnecessarily more difficult or even de facto impossible*".[75] Havlikova who examined the interplay between the opt-out provision in Article 4 of the CDSM and web scraping, a form of data collection from web sources which is often the starting point of some forms of TDM, found that rightsholders may struggle to reserve their rights.[76] Havlikova attributes this to the fact that for content published online, a rightsholder's reservation must be made using machine-readable means and "*according to Recital 18 of the CDSM Directive, such machine-readable means may include metadata and terms and conditions of a website or a service*".[77] Havlikova argues that as a result of such a wide and unstructured means afforded to a rightsholder allowing for them to reserve their right "*could be the Achilles heel of the TDM exception*".[78] Havlikova ultimately argues that "*Robots.txt, which is a simple text file containing rules on which crawlers may access which parts of a site*",[79] could be one feasible way of reserving one's rights in a machine-readable means, but that this would also have its drawbacks, especially when it comes to indexing websites.[80] As it stands, the author calls for a more standardised method of reserving one's rights through machine readable means.[81]

As aforementioned, awarding copyright protection is a balancing exercise of all stakeholders' interests. One way of balancing the extent of protection afforded is through the applicability of exceptions. These mandatory exceptions, whilst a step forward in the context of a harmonised approach towards the regulation of TDM, have their limitations which may ultimately tip the balance in favour of rightsholders.

The limitation of scope of these TDM-specific EU exceptions has been regarded as a major pitfall in the EU's attempt to become more competitive in the digital era.[82] For instance, Rosati argues that they "*might have a negative impact on the (unlicensed) development of AI creativity*".[83] As per Manteghi "*the vague and narrow nature* [of the EU CDSM TDM-specific exceptions] *could force users to consider alternative solutions to avoid potential infringement claims*".[84] One of these solutions, Manteghi claims could be found under Article 35 of the proposed Data Act which "*provides a safe harbour for users needing to access, use or share only databases made of data generated by the use of a product or related service (IoT data)*".[85] This avenue towards avoiding potential infringement claims is limited, to certain types of TDM which will utilise certain specific types of databases, meaning it would only apply to potential sui generis database rights infringements and not copyright. Nevertheless, it is possible that certain users conducting TDM may find refuge under this proposed Article. Another alternative may be offered by Article 5(1) of the InfoSoc Directive which allows for transient copies to be made.[86] Nevertheless, as per Margoni and Kretschmer whilst Article 5(1) of the InfoSoc Directive "*retains a significant potential for TDM activities and computational uses, the cumulative, occasionally narrow and partially uncertain nature of its conditions and the fact that it only covers temporary reproductions, does not offer a clear and comprehensive solution within which not only science but virtually any human activity employing text and data analytics can operate confidently*".[87]

The extent of these exceptions' reach is inevitably dependent upon the national implementation of Member States. Pereira in their examination of the Portuguese implementation of these

[71] ibid 4.

[72] Article 6, Directive 2001/29 on the harmonisation of certain aspects of copyright and related rights in the information society [2001] OJ L167/10.

[73] Andrée Glancia MADINDA, 'THE TEXT AND DATA MINING EXCEPTION IN THE DIRECTIVE ON COPYRIGHT IN THE DIGITAL SINGLE MARKET: A STORMY OUTLOOK FOR AUTHOR'S RIGHTS' [2022] Winner of the 2022 edition of the GESAC/ALAI European Authors' Right Award.

[74] 'Europe's TDM Exception for Research: Will It Be Undermined By Technical Blocking From Publishers?' (*LIBER Europe*) <https://libereurope.eu/article/tdm-technical-protection-measures/> accessed 7 February 2024.

[75] Hilty and Richter (n 37) para 29.

[76] Stepanka Havlikova, 'Web Scraping and Text and Data Mining Exception: Could the Cdsm Directive, Designed to Support the Reuse of Publicly Available Data, Have Had the Opposite Effect?' (18 October 2023) <https://papers.ssrn.com/abstract=4605551> accessed 29 May 2024.

[77] ibid 14.

[78] ibid 14–15.

[79] ibid 17.

[80] ibid.

[81] ibid 24.

[82] Geiger, Frosio and Bulayenko (n 62).

[83] Rosati (n 16).

[84] Maryna Manteghi, 'Overcoming Barriers to Text and Data Mining in the Era of ChatGPT: The Proposed Data Act as a Game-Changer' [2023] GRUR International.

[85] ibid.

[86] Article 5(1) of Directive 2001/29. See Margoni and Kretschmer (n 37) 691.

[87] ibid 694.

exceptions, which "*has been faithful*" to the CDSM, argues that these exceptions seek to balance the interests of all relevant stakeholders.[88] Other commentators have however, called for a wider approach. For instance, Micke called for Finland to take an expansive transposition thereby guaranteeing the freedom to mine.[89] Similarly, Calabrese questions why the EU legislature did not exempt TDM users from infringing a rightsholder right of communication to the public in the TDM exceptions, and praises Italy for doing so in their national transposition.[90] Calabrese concludes by stating that the Italian approach "*can be appreciated as an attempt of 'better'—or at least 'not worse'—harmonization in terms of free communication to the public for TDM, not forgetting that the EU has also to face the regulatory competition of third-country systems which are more open to innovation in this field*".[91] When stating this Calabrese was referring to the Japanese approach which was the first nation in the world which introduced a TDM-specific copyright exception allowing for TDM researchers to utilise copyright protected works for machine learning purposes.[92] Dermawan similarly argued that EU Member States shall consider adopting the Japanese approach towards a more favourable TDM exemption in their national implementation of the CDSM.[93] Nevertheless, Rosati, when examining the extent to which national member states have freedom to diverge from Articles 3 and 4 of the CDSM, stated "*both Articles 3 and 4 require a minimalistic transposition technique on the side of national legislatures, with the result that the normative content of either provision cannot be compromised at the Member State level. In turn, a national* [exception] *with a broader scope of application than what either provision allows would be incompatible with EU law*".[94] At the time of writing this review, Poland had yet to implement the TDM-specific copyright exceptions into national law. The Polish government argues that their delayed implementation of the CDSM enabled a thorough examination of the implications of generative AI for copyright protection, ultimately deciding that training generative AI models on copyright protected works should fall outside the scope of the TDM-specific copyright exceptions.[95] How this rationale plays out remains to be seen, but as noted by Paul Keller in a blog post, it is "*clear that any attempt to exclude from the scope of the TDM provision the reproductions made in the context of training generative AI models would, prima facie, result in a non-compliant implementation*".[96] Hence, it is unlikely that such a divergence is accepted, but it does highlight the fact that national implementation of these exceptions remains fragmented. In their master's thesis, André Stéphan, in analysing the national implementation of France, Germany, the Netherlands, Spain and Ireland argues that "*if the objective was to create a digital single market, where stakeholders could benefit from the same rules in each Member State, it has not been reached*".[97] This implies that there is still a long way to go before true harmonisation is reached.

Another jurisdiction under which TDM researchers may find refuge is that of the US. There are numerous scholars which approach TDM through the lens of the US's fair use copyright exceptions. Namely, as did EU scholars, several authors who examine the US's regulatory framework for TDM assert that TDM is and should be lawful as a matter of US copyright law.[98] In other words such authors claim that the US fair use copyright exceptions would encompass the act of TDM. Most justify this reasoning by presenting the lawsuits that commercial authors and publishers lodged against Google,[99] and the HathiTrust Digital Library.[100] These cases involved the digitisation of printed books, allowing for users to search through this collection of literature. In accordance with the Author Guild's perspective the systematic copying of copyright protected authorial works amounted to copyright infringement. However, Google and HathiTrust argued that their re-use was "fair use" as it was socially beneficial, but also since copyright protected works were not displayed in full. Only small extracts of the copyright protected works were presented to those who used the search function. These arguments ultimately swayed the court to rule in favour of Google and HathiTrust in finding that their reuse amounted to fair use.

Sag argues that TDM would have always found refuge under the US's fair use exceptions and the US courts have now solidified this position.[101] Sag makes a distinction which the courts failed to do in these cases between a non-expressive use and transformative use.[102] Sag argues that TDM would fall under

[88] Alexandre L Dias Pereira, 'TDM Copyright for AI in Europe: A View from Portugal' (2023) 18 Journal of Intellectual Property Law & Practice 900.

[89] Micke Lindholm, 'Text and Data Mining under Finnish Copyright Law before and after the DSM Directive' (2020) <https://helda.helsinki.fi/items/42a78c1d-da42-4461-a7f1-37b66c4d3cb6> accessed 14 November 2023.

[90] Bernardo Calabrese, 'Scientific TDM Exception and Communication to the Public: Did Italians Do It Better … or at Least Not Worse?' (2022) 17 Journal of Intellectual Property Law & Practice 399.

[91] ibid.

[92] Samuelson (n 13).

[93] Artha Dermawan, 'Text and Data Mining Exceptions in the Development of Generative AI Models: What the EU Member States Could Learn from the Japanese "Nonenjoyment" Purposes?' The Journal of World Intellectual Property <https://onlinelibrary.wiley.com/doi/abs/10.1111/jwip.12285> accessed 18 January 2024.

[94] Eleonora Rosati, 'No Step-Free Copyright Exceptions: The Role of the Three-Step in Defining Permitted Uses of Protected Content (Including TDM for AI-Training Purposes)' (2023) 123.

[95] Paul Keller, 'TDM: Poland Challenges the Rule of EU Copyright Law' (*Kluwer Copyright Blog*, 20 February 2024) <https://copyrightblog.kluweriplaw.com/2024/02/20/tdm-poland-challenges-the-rule-of-eu-copyright-law/> accessed 21 February 2024.

[96] ibid.

[97] Stéphan André, 'The Legal Evolution of the Text and Data Mining Copyright Exceptions in the European Union - From Fragmentation to Uniformity?' (2022) 63.

[98] Carroll (n 25); Edward Lee, 'Technological Fair Use' (2010) 83 Southern California Law Review 797; Jerome Reichman and Ruth Okediji, 'When Copyright Law and Science Collide: Empowering Digitally Integrated Research Methods on a Global Scale' (2012) 96 Minnesota Law Review 1362; Matthew Sag, 'Copyright and Copy-Reliant Technology.' [2009] Faculty Publications & Other Works.

[99] Authors Guild, Inc. v. Google, Inc., 804 F.3d 202 (2d Cir. 2015). See for instance, Benjamin Goh, 'Two Ways of Looking at a Printed Book' (2022) 85 The Modern Law Review 697.

[100] Authors Guild, Inc. v. HathiTrust, 755 F.3d 87 (2d Cir. 2014).

[101] Sag (n 21) 293.

[102] ibid 302.

the category of a non-expressive use, whereby the act of TDM does not communicate original forms of expression for the purposes of them being read.[103] Sag assimilates it to photocopying a magazine and throwing it in a fireplace, which would not be regarded as infringing.[104] It is asserted that in Google Book saga cases non-expressive reuse was treated as a subset of transformative uses falling under "fair use".[105]

Carroll, argues that in the US, "*a researcher can legally download all or a portion of the Sci-Hub collection* [a website which contains copious copyright infringing academic literature] *solely for TDM research*".[106] Whilst, such an argument may attract greater controversy in that it tips the balance in favour of users to the detriment of rightsholders, it is actually supported by others.[107] In a recent blog post about an upcoming journal article, Thomas Margoni puts forward this same idea, but within the remits of an EU jurisdiction.[108] Margoni argues that the lawful access requirement found under Articles 3 and 4 of the CDSM which ultimately creates a TDM realm where content which is accessed needs a license (discussed in greater detail below), should "*only cover the behaviour of the beneficiary of the exception and not extend to the status of the accessed source*".[109] Whether this liberal interpretation will indeed be the case remains to be seen.

Similar to its EU counterpart, the US allows for TPMs restricting TDM.[110] Nevertheless, Dombrowski and Tilton highlight the historical development of an exception for circumvention of TPMs for research purposes in US.[111] The authors state that "*per the final decision, researchers residing in institutions of higher education could bypass DRM for research under a series of conditions. This included staff and students if they are a part of a research team or as a part of teaching. The university must own the source of the data (i.e., not accessing it via a subscription service), and the researchers must take "effective security measures" to protect the data. As a result, researchers at institutions of higher education can conduct TDM with sources such as DVDs and ebooks, which is a significant development for the Digital Humanities*".[112] Nevertheless, the authors argue that while the ruling expands TDM access for "well-resourced" institutions, it falls short when considering wider TDM users and activities.[113]

It can be deduced that authors generally view intellectual property rights as restricting one's freedom to conduct TDM. Whilst the recent introduction of the EU CDSM TDM-specific copyright exceptions provides a new avenue towards "lawful TDM", their narrow scope acts as a barrier to free mining. The wider approach of the US arguably creates a more welcoming market for TDM in the US. In a 2022 Research Paper Sean Flynn *et al.* categorised global copyright exceptions for research purposes focusing on TDM exceptions.[114] The authors provide an informative world map comprising of seven colours categorising countries as having a general research exception, restriction on sharing, restrictions to private reproductions, restrictions to institutional users, restrictions on types of works, TDM being restricted and countries that were not mapped.[115] The authors then comparatively analyse each of their categorisations. They also provide a figure comparing specific TDM exceptions that have been introduced.[116] Sean Flynn *et al.* ultimately "*show that although every copyright law in the world has at least one exception that promotes research purposes, there is a wide degree of variation between countries*" ranging across the spectrum of categorisations they identified with some like the US, being more favourable than others.[117] As per Rosati "*the US copyright regime has been considered more favourable to TDM practices than what appears to be the case under European laws*".[118] This is also evidenced by an empirical study conducted by Handke *et al.*, in 2015 who empirically examined bibliometric data from 43 large economies, including the 15 largest EU Member States, spanning from 1992–2014.[119] The authors found that data mining based articles were more common in countries where mining is "probably allowed", under copyright law like that of the US, "*which suggests that a more permissive copyright framework is associated with more data mining research*".[120] On the other hand, Kretschmer *et al.*, who examined machine learning, natural language processing and computer vision for computer moderation as case studies where TDM is vital, found that US fair use jurisprudence is likely to set global trends. The legal uncertainties created by fair use exceptions, may lead to practices "*where commercial AI developers are told by their legal departments to "mine everything and then destroy the training material" since it will be very difficult to reverse-engineer the trained model, go back to the training material*

[103] ibid.

[104] ibid 301–302.

[105] ibid 320.

[106] Carroll (n 25) 954.

[107] Pamela Samuelson, 'A Bold Take on Copyright Implications of Text & Data Mining' [2021] JOTWELL <https://ip.jotwell.com/a-bold-take-on-copyright-implications-of-text-data-mining/> accessed 10 November 2023.

[108] Thomas Margoni, 'Saving Research: Lawful Access to Unlawful Sources under Art. 3 CDSM Directive?' (*Kluwer Copyright Blog*, 22 December 2023) <https://copyrightblog.kluweriplaw.com/2023/12/22/saving-research-lawful-access-to-unlawful-sources-under-art-3-cdsm-directive/> accessed 18 January 2024.

[109] ibid.

[110] U.S. Code § 1201.

[111] Quinn Dombrowski and Lauren Tilton, 'Access and Advocacy: Text & Data Mining and DMCA §1201' (2024) 13 Digital Studies / Le champ numérique <https://www.digitalstudies.org/article/id/9658/> accessed 29 May 2024.

[112] ibid 12.

[113] ibid 14.

[114] Sean Flynn and others, 'Research Exceptions in Comparative Copyright' [2022] Joint PIJIP/TLS Research Paper Series <https://digitalcommons.wcl.american.edu/research/75>.

[115] ibid 17.

[116] ibid 36.

[117] ibid 37.

[118] Rosati (n 16).

[119] Handke, Guibault and Vallbé (n 18).

[120] ibid 18.

and prove infringement".[121] Thus, the balancing of stakeholder interests in the realm where copyright and related rights, and TDM intersect arguably still needs refining.

TDM and data protection

Given the inherent processing of data involved in the act of TDM, it is inevitable that users conducting TDM encounter data protection related regulatory frameworks.[122] Nonetheless, how TDM interacts with data protection frameworks seems under-researched. Colonna states that in existing literature "*while the applicability of copyright law to TDM is being considered, there is very little research being done concerning the applicability of data protection law to this new technology"*.[123] This section aims to nevertheless, provide an outline of existing research in the field examining the interplay between TDM and data protection regulatory frameworks.

One of the most influential regulatory frameworks in this regard is the EU General Data Protection Regulation (GDPR).[124] As per Lynskey, unlike jurisdictions like the US, which is characterised as having a "sectoral" regulation of data protection and privacy, meaning certain aspects of data protection and privacy are regulated on a granular level, the GDPR takes an "omnibus" approach towards the all-encompassing regulation of data protection and privacy.[125] Under Article 2, the GDPR regulates the processing of personal data.[126] As such "*to fully understand the material scope, one needs to look at the definitions of "personal data",* [and] *"processing"*."[127]

Article 4(1) defines personal data as "*any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person"*.[128] The notion of "any information," implies that the term "personal data" should be interpreted as broadly as possible.[129] Similarly, the CJEU in case C-434/16, stated that the notion of "relating to" found under Article 4(1) of the GDPR is "*is satisfied where the information, by reason of its content, purpose or effect, is linked to a particular person*".[130] As per Bygrave and Tosoni, this "*puts another gaping hole in the few borders of the personal data concept, although the reference to 'a particular person' would seem to exclude that data relating to an aggregate of persons (e.g. a household) are personal data, irrespective of the size of the aggregate"*.[131] Hence the concept of "personal data" is arguably rather broad. One must also consider the higher level of protection afforded to so called "special categories of personal data" under Article 9 of the GDPR which restricts the processing of "*personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation"*.[132] This may prove particularly challenging for users conducting TDM, particularly where inferences can be made using particular data.[133] The GDPR implications caused to users conducting TDM by inferences are discussed below.

Article 4(2) defines processing as "*any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction"*.[134] This definition, like the definition of "personal data" is also

[121] Martin Kretschmer, Thomas Margoni and Pinar Oruç, 'Copyright Law and the Lifecycle of Machine Learning Models' (2024) 55 IIC - International Review of Intellectual Property and Competition Law 110, 125.

[122] Lita van Wel and Lambèr Royakkers, 'Ethical Issues in Web Data Mining' (2004) 6 Ethics and Information Technology 129.

[123] Colonna (n 15).

[124] Pattaraporn Sangaroonsilp and others, 'Mining and Classifying Privacy and Data Protection Requirements in Issue Reports' (arXiv, 2015) <http://arxiv.org/abs/2112.13994> accessed 21 November 2023.

[125] Orla Lynskey, *The Foundations of EU Data Protection Law* (Oxford University Press 2015) 15.

[126] Article 2, 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) [2016] OJ L 119/1.

[127] Herke Kranenborg, 'Article 2 Material Scope' in Christopher Kuner and others (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press 2020) 66 <https://doi.org/10.1093/oso/9780198826491.003.0004> accessed 3 September 2024.

[128] Article 4(1), 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) [2016] OJ L 119/1.

[129] Lee A Bygrave and Luca Tosoni, 'Article 4(1). Personal Data' in Christopher Kuner and others (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press 2020) 109 <https://doi.org/10.1093/oso/9780198826491.003.0007> accessed 4 September 2024.

[130] Case C-434/16, Nowak.

[131] Bygrave and Tosoni (n 129) 110.

[132] Article 9(1), 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) [2016] OJ L 119/1.

[133] Sandra Wachter and Brent Mittelstadt, 'A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI' (2019) 2019 Columbia Business Law Review 494, 71.

[134] Article 4(2), 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) [2016] OJ L 119/1.

broad, essentially covering "*any data processing operation*".[135] Thus, taking a synoptic view of the notions of "personal data" and "processing" it can be deduced that the act of TDM is likely to interact with the GDPR.

Apart from its "omnibus" approach toward the regulation of data protection and privacy, the GDPR also has wide territorial reach. As per Dove "*one might consider the GDPR to be a Europe-centric law of little global consequence […] on the contrary, the territorial scope of the GDPR follows the data that it protects and therefore has direct bearing on the activities of organisations based in countries around the world*".[136] This cross-border reach of the GDPR may in fact restrict certain forms of TDM. As stated by Colonna, "*the ability to conduct meaningful research utilising TDM requires not only being able to access data remotely but also being able to share and further process results within the Digital Single Market and beyond*".[137] The territorial reach of the GDPR is governed by Article 3 which extends the regulations' reach beyond the territorial boundaries of the EU. As clearly outlined by Svantesson, "*Article 3 may be broken down into three parts. The first (Article 3(1)), ensures that the GDPR applies to the processing of personal data by a controller or a processor with an establishment in the Union. The second (Article 3(2)), extends the GDPR's application to a controller or a processor that lacks an establishment in the Union, under certain defined circumstances. The third (Article 3(3)), addresses specific situations where Member State law applies by virtue of public international law*."[138] As such, the GDPR not only applies to personal data that is processed in the EU, but also beyond, following the data that it protects. Given the global scale at which TDM is often conducted, it is consequently expected that users conducting TDM will only escape the territorial scope of the GDPR in very limited circumstances.

To legally process personal data, a user conducting TDM would have to abide by the principles relating to lawful processing of personal data under Article 5, meaning the personal data would have to be "*processed lawfully, fairly and in a transparent manner in relation to the data subject*" (Article 5(1)(a)), "*collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes"* (Article 5(1)(b)), "*relevant and limited to what is necessary in relation to the purposes for which they are processed"* (Article 5(1)(c)), "*accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay"* (Article 5(1)(d)), "*kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed"* (Article 5(1)(e)), "*processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures"* (Article 5(1)(f)).[139]

One issue, raised in the literature examined for this literature review, which users conducting TDM could encounter and is discussed in greater detail below, is the issue that comes along with data minimisation under Article 5(1)(c). Namely, minimising data can sometimes lead to a loss of context, potentially causing more harm than good in certain data mining scenarios. In their book chapter which predates the introduction of the GDPR, Van der Sloot, instead of data minimisation, which was also a requirement in the Data Protection Directive which predates the GDPR, introduce the concept of "data minimumisation".[140] Meaning that a minimum set of data "*are gathered, stored and clustered when used in practice*".[141] Van der Sloot argues that doing so, could alleviate some of the data protection and privacy concerns that arise from a loss of context due to minimising data.[142] This balance between data manipulation and data utility is discussed in greater detail below.

Additionally, to legally process data, users conducting TDM would need to specify a legal basis for the data processing under the closed list of six legal bases found in Article 6 of the GDPR.[143] The first legal basis for processing data is that of consent (Article 6(1)(a)). Obtaining consent may be cumbersome for users conducting TDM. For instance, taking the use of social media data as an example, Bremert argues that "*it is more than doubtful, that a clause, hidden within the terms of service or the privacy policy of an online service is a clear indication of the data subject's wish nor can these terms specify legal consequences for plain actions such as publishing content on a social media platform"*.[144] Bremert ultimately

[135] Luca Tosoni and Lee A Bygrave, 'Article 4(2). Processing' in Christopher Kuner and others (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press 2020) 119 <https://doi.org/10.1093/oso/9780198826491.003.0008> accessed 4 September 2024.

[136] Edward S Dove, 'The EU General Data Protection Regulation: Implications for International Scientific Research in the Digital Era' (2018) 46 The Journal of Law, Medicine & Ethics 1013, 1013.

[137] Colonna (n 15).

[138] Dan Jerker B Svantesson, 'Article 3 Territorial Scope' in Christopher Kuner and others (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press 2020) 3 <https://doi.org/10.1093/oso/9780198826491.003.0005> accessed 4 September 2024.

[139] Article 5, 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) [2016] OJ L 119/1.

[140] Bart van der Sloot, 'From Data Minimization to Data Minimummization' in Bart Custers and others (eds), *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases* (Springer 2013) <https://doi.org/10.1007/978-3-642-30487-3_15> accessed 5 September 2024.

[141] ibid 285.

[142] ibid 286.

[143] Article 6, 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) [2016] OJ L 119/1.

[144] Benjamin Bremert, 'Legal Aspects of Text Mining Publicly Available Data' [2017] Unabhängiges Landeszentrum für Datenschutz 4.

concludes that such consent should be regarded as a "sham consent" thereby not being regarded as "legal basis" for the processing of data.[145] Similarly, where the data being processed as part of mining activities, is being processed for purposes which differ from the purposes the data was initially collected for, users conducting TDM may struggle to comply with Article 6(4) of the GDPR.[146] To comply, a compatibility test would have to be performed "*which considers the link between the original purposes and the secondary purpose, the context of processing, the type and nature of the data, the possible consequences of the further processing, and the presence of the appropriate safeguards for processing"*.[147]

Notably, under Article 89 the GDPR acknowledges the importance of facilitating different forms of research.[148] A baseline level of adherence with the GDPR is set for scientific and research purposes whilst balancing the rights and freedoms of the respective data subjects using appropriate safeguards, such as pseudonymisation or anonymisation "*though the list of safeguards mentioned in the article is non-exhaustive"*.[149] It permits derogations from certain data subject rights including, if fulfilling them would impair the research. Article 89 also allows the further processing of personal data for research even if not originally collected for that purpose, and data can be stored for extended periods if solely used for research. Thus, provided TDM is carried out for research purposes, it is possible that the such users may find refuge under Article 89 of the GDPR with exceptions that are designed to balance privacy with the benefits of research.

As aforementioned, one of the first steps for TDM is the collection of data. Arguably, new methods of data collection, namely "data scaping", also referred to as "web scraping", which allows users to systematically and efficiently collect data from the web, have allowed for more efficient and wider reaching data collection.[150] However, it is argued that users who conduct TDM and utilise such data collection methods are likely to encounter restrictions posed by the GDPR.[151] Namely, in the Polish case of *Bisnode* it was decided that using data scraping, may result in a breach of Article 14 (1–3) of the GDPR which mandates users who scrape publicly available data to provide the data subject with specific details, including the types of data being processed, the legal basis and purposes for the processing, their rights concerning personal data, and the contact information of the data controller.[152]

While the GDPR provides protections to users whose data are being collected, TDM in some instances may be used to draw inferences and predictions "*about the behaviours, preferences, and private lives of individuals"*.[153] The GDPR's application to inferential data i.e., the insights and predictions derived from data mining, remains ambiguous and insufficient. According to Wachter and Mittelstadt, inferences drawn from data are often treated as "economy class" personal data under the GDPR, meaning individuals have little control or oversight over how these inferences are used in decision-making processes.[154] This "accountability gap" is especially concerning given that inferences can be privacy-invasive, discriminatory, or damaging to one's reputation, yet they are often unverifiable and based on non-transparent algorithms.[155] As such, Wachter and Mittelstadt argue in favour of a new "right to reasonable inferences" that would require data controllers to justify and disclose the reasoning behind high-impact inferences before they are used.[156] This gap reveals that while the GDPR aims to regulate the collection and processing of data, it falls short in addressing the broader implications of algorithmic decision-making and profiling in TDM-related activities.

Whilst the GDPR, is generally viewed as an advancement in data protection law, encouraging responsible scientific research by introducing a plethora of stringent pre-processing and procedural requirements such as data protection by design, requirements of reporting data breaches, the appointment of data protection officers, the enforcing of hefty fines, and the right to be forgotten, it is argued that it does not necessarily represent a radical shift in data protection laws.[157] Nevertheless, it also argued that some businesses face challenges in going from GDPR non-compliance to compliance.[158] "*Organisations and businesses have been processing personal data of people, in cases without explicit consent, up to the extent required for their businesses, retaining it for longer periods, and even*

[145] ibid 5.

[146] Catherine Altobelli and others, 'To Scrape or Not to Scrape? The Lawfulness of Social Media Crawling under the GDPR' (2021) 175.

[147] Altobelli and others (n 146).

[148] Christopher F Mondschein and Cosimo Monda, 'The EU's General Data Protection Regulation (GDPR) in a Research Context' in Pieter Kubben, Michel Dumontier and Andre Dekker (eds), *Fundamentals of Clinical Data Science* (Springer 2019) <http://www.ncbi.nlm.nih.gov/books/NBK543521/> accessed 4 September 2024.

[149] Christian Wiese Svanberg, 'Article 89 Safeguards and Derogations Relating to Processing for Archiving Purposes in the Public Interest, Scientific or Historical Research Purposes or Statistical Purposes' in Christopher Kuner and others (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press 2020) 1247 <https://doi.org/10.1093/oso/9780198826491.003.0136> accessed 4 September 2024.

[150] Zuzanna Gulczynska, 'Scraping Personal Data from Internet Pages?: A Comparative Analysis of the Polish Bisnode Decision and the US hiQ Labs v LinkedIn Corp Judgment' (2020) 45 EUROPEAN LAW REVIEW 857.

[151] ibid.

[152] Decision of 15 March 2019, Ref. ZSPR.421.3.2018 (President of the Polish Data Protection Office); English translation: https://uodo.gov.pl/en/file/314 [Accessed 15 March 2020].

[153] Wachter and Mittelstadt (n 133).

[154] ibid 6.

[155] ibid 17–18.

[156] ibid 89–93.

[157] Dove (n 136).

[158] Rashid Zaman and Marwan Hassani, 'Process Mining Meets GDPR Compliance: The Right to Be Forgotten as a Use Case' [2019] ICPM Doctoral Consortium 2019 <http://www.scopus.com/inward/record.url?scp=85071764645&partnerID=8YFLogxK> accessed 22 November 2023.

subletting it to other businesses to be used for their own processing purposes".[159]

In existing literature examining the interplay between TDM and data protection and privacy rules, a lot of emphasis is placed on privacy by design, or privacy preserving measures.[160] GDPR non-compliance may result in hefty fines being placed.[161] Hence, an emphasis on privacy preserving measures in existing literature may be correlated to an increasing need for GDPR compliance.

Privacy by design is enshrined under Article 25 of the GDPR which asks of data controllers and processors to "*both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects*."[162] It is often recommended that techniques for preserving privacy are adopted at early stages of a research project to align with respective regulatory frameworks.[163]

The European Data Protection Board, in addition to the early adoption of privacy preserving measures, makes a series of further recommendations for adherence to Article 25 of the GDPR.[164] They encourage the active involvement of a Data Protection Officer, if available.[165] They also recommend the certification of processing operations as these offer added value and can enhance trust and competitiveness.[166] If certification is not available, controllers should ensure producers and processors comply with privacy by design requirements through other guarantees.[167] Specific protections for children and vulnerable groups should be considered, and producers and processors should support controllers in meeting privacy by design obligations by facilitating implementation and staying updated on technological advancements.[168] Controllers should include contractual clauses to keep up with changes and use performance indicators to assess compliance.[169] Harmonised, sector-specific guidance should be sought, and transparency in demonstrating effective privacy by design implementation is crucial.[170] Privacy-enhancing technologies (PETs) should be used as appropriate, and legacy systems must comply with privacy by design obligations or be discontinued.[171] SMEs can facilitate compliance by conducting early risk assessments, starting with small-scale processing, seeking guarantees from producers and processors, and utilizing available guidance and professional advice.[172] Adopting privacy by design means that rules mandated by regulatory frameworks can be "hardwired" into a TDM process.[173]

A number of scholars in this domain, attempt to provide practical guides on how to implement privacy by design. For instance, Sangaroonsilp *et al.*, identified "*an emerging need to translate complex privacy concerns set out in regulations and standards into requirements that are to be implemented in software applications*".[174] The authors consequently "*developed a comprehensive taxonomy of privacy requirements for software systems by extracting and refining requirements from the widely-adopted GDPR and ISO/IEC 29100 privacy framework as well as the newly developed Thailand PDPA and the region-specific APEC privacy framework*".[175] Hence, Sangaroonsilp *et al.* took a multijurisdictional approach to the applicability of data protection and privacy related regulatory frameworks developing a holistic taxonomy of the requirements a business which collects and/or processes data will have to adhere to. This could equally apply to a user conducting TDM. The authors broke these requirements down into 7 categories, including themes like lawfulness, purpose limitation, data minimisation, accuracy, storage limitation, integrity and confidentiality, and accountability, with a total of 71 requirements.[176]

Other scholars provide more technical guidance on implementing a privacy by design approach. The majority of such literature predates the GDPR. Nevertheless, they still provide valuable technical guidance, for implementing techniques which would

[159] ibid.

[160] Nils Gruschka and others, 'Privacy Issues and Data Protection in Big Data: A Case Study Analysis under GDPR', *2018 IEEE International Conference on Big Data (Big Data)* (2018) <https://ieeexplore.ieee.org/document/8622621> accessed 21 November 2023.

[161] Article 83(40, EU General Data Protection Regulation (GDPR): 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), OJ 2016 L 119/1.

[162] Article 25, EU General Data Protection Regulation (GDPR): 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), OJ 2016 L 119/1.

[163] Gruschka and others (n 160); Daniel Solove, 'Why I Love the GDPR: 10 Reasons' (*TeachPrivacy*, 2 May 2018) <https://teachprivacy.com/why-i-love-the-gdpr/> accessed 21 November 2023.

[164] EDPB, 'Guidelines 4/2019 on Article 25 Data Protection by Design and by Default | European Data Protection Board' 29 <https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-42019-article-25-data-protection-design-and_en> accessed 5 September 2024.

[165] ibid.

[166] ibid.

[167] ibid.

[168] ibid 30.

[169] ibid.

[170] ibid.

[171] ibid.

[172] ibid.

[173] Bart Schermer, 'Risks of Profiling and the Limits of Data Protection Law' in Bart Custers and others (eds), *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases* (Springer 2013) 150 <https://doi.org/10.1007/978-3-642-30487-3_7> accessed 5 September 2024.

[174] Sangaroonsilp and others (n 124).

[175] ibid.

[176] ibid.

allow for GDPR compliant mining techniques. Clifton and Marks emphasise the need to balance between the benefits of data sharing and the risks associated with data mining, advocating for careful control of access to data to ensure efficient sharing without compromising security and privacy.[177] Whilst this paper is from the year 2000, the authors acknowledge the evolving nature of the field and call for ongoing research to address emerging challenges in preventing unwanted data mining.[178]

Schermer examines the data mining approaches in profiling which involves discovering correlations in databases to represent individuals or groups.[179] Schermer explores two data mining approaches: descriptive, which uncovers relations between data objects without a specific target, and predictive, which predicts events based on established patterns.[180] Schermer identifies an number of privacy related risks associated with profiling using data mining techniques such as discrimination, de-individualization, and information asymmetries.[181] Schermer argues that traditional privacy protection strategies alone are insufficient, and a multidisciplinary approach involving increased accountability, transparency, and algorithmic safeguards is crucial.[182] Schermer argues that for descriptive data mining, awareness among data subjects and controllers about the possibilities, limits, and risks of data mining is essential.[183] Similarly predictive data mining requires transparency to prevent inaccurate applications, emphasizing the importance of distinguishing between direct and indirect indicators in decision-making.[184] Whilst this paper also predates the GDRP, it highlights, and provides guidance on some of the difficulties highlighted by Wachter and Mittelstadt, which were discussed above.[185] Namely that GDPR's application to inferential data remains ambiguous and insufficient.

Clifton *et al.*, in a pre-GDRP era highlight the growing interest in privacy-preserving data mining, but the existence of ambiguity in defining privacy.[186] Whilst this paper predated the GDPR, it provides some technical guidance on privacy preserving techniques which could be implemented by users conducting TDM today. Clifton *et al.* underscore the need for diverse measures of privacy preservation, providing four distinct measures for doing so.[187] First, the concept of bounded knowledge is introduced which involves data obscuration techniques where information about a protected attribute may be revealed within certain bounds.[188] Clifton *et al.* provide a metric for doing so which adds noise to certain variables.[189] Second the so-called "need-to-know" concept is introduced which implies that data should be processed only if necessary for a specific purpose, aligning with legal and privacy regulations.[190] Secure Multiparty Computation is suggested as a basis for data mining that adheres to the "need to know" standard.[191] "*If the results of the data mining are required to accomplish an allowable task, then learning those results should be allowed under the "need to know" standard under which most privacy regulations allow release of information. Data mining approaches that are Secure Multiparty Computations can be proven not to disclose anything except the results"*.[192] Third Clifton *et al.* argue that certain items, including individual data items and rules, may need protection from disclosure categorising such techniques for achieving this under the category of "protected from disclosure".[193] Existing techniques in the database security community for inference prevention are discussed, but challenges in defining appropriate thresholds persist.[194] Finally Clifton *et al.* present anonymity techniques as a concept for privacy preserving data mining techniques.[195]

Lindell and Pinkas address the challenge of two parties with private databases wanting to collaborate on computing a data mining algorithm without revealing the contents of their databases to each other.[196] The goal is for each party to only learn information from the output of the data mining algorithm.[197] Lindell and Pinkas propose the use of decision tree learning.[198] One concern that is raised is the concept of the so called "semi-honest adversary", where malicious parties follow the protocol but attempt to learn additional information from the communication transcript, by altering their input, e.g. leaving their input parameter blank returning the data of the other party which they wish to retrieve.[199]

Agrawal and Aggarwal aim to protect user information while still enabling data mining.[200] The proposed approach involves perturbing data values using a known distribution, allowing the

[177] Chris Clifton and Don Marks, 'Security and Privacy Implications of Data Mining' [2000] ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery.

[178] ibid.

[179] Bart W Schermer, 'The Limits of Privacy in Automated Profiling and Data Mining' (2011) 27 Computer Law & Security Review 45.

[180] ibid.

[181] ibid 48.

[182] ibid 52.

[183] ibid 50.

[184] ibid.

[185] Wachter and Mittelstadt (n 133).

[186] Chris Clifton, Murat Kantarcioglu and Jaideep Vaidya, 'Defining Privacy for Data Mining' (2003).

[187] ibid 198.

[188] ibid.

[189] ibid.

[190] ibid 199.

[191] ibid 200.

[192] ibid.

[193] ibid.

[194] ibid 200–201.

[195] ibid 202.

[196] Lindell and Pinkas, 'Privacy Preserving Data Mining' (2002) 15 Journal of Cryptology 177.

[197] ibid.

[198] ibid.

[199] ibid 2.

[200] Dakshi Agrawal and Charu C Aggarwal, 'On the Design and Quantification of Privacy Preserving Data Mining Algorithms', *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (Association for Computing Machinery 2001) <https://dl.acm.org/doi/10.1145/375551.375602> accessed 23 November 2023.

reconstruction of aggregate distributions rather than individual records.[201] Similarly, in their book, Vaidya, Zhu and Clifton explore techniques such as data perturbation and Secure Multiparty Computation to address potential misuse of data.[202] This highlights the aforementioned balance that must be struck between data manipulation and data utility.

Kantarcioğlu *et al.*, raise concerns about existing techniques for privacy-preserving data mining, such as adding noise to data or encryption-based approaches.[203] Kantarcioğlu *et al.* emphasise the need to evaluate the privacy impact of data mining models.[204] The authors use a "medical diagnosis" scenario to illustrate the complexity of privacy issues, discussing the challenges of determining whether a classifier violates privacy when using public and private information.[205] The authors propose a framework for evaluating privacy loss in data mining results, offering precise definitions and formal analysis.[206] Section 2 presents a classification model, and Section 3 introduces a metric for privacy loss, providing examples and methods for calculation.[207] Their work aims to contribute to understanding and addressing privacy concerns in the context of data mining models.[208]

In conclusion, literature examining the interplay between TDM, and data protection is sparse. There seems to be a focus on creating mining mechanisms which implement privacy by design and several scholars provide practical and technical solutions for achieving this. The common underlying message seems to be the manipulation of private information to avoid data breaches. Nevertheless, there seems to be a correlation between data manipulation and data utility and a balance must be struck between the two to allow for privacy preserving TDM research to be carried out.

TDM and fundamental rights

Another regulatory framework which may come into play when conducting TDM, which is sometimes examined by scholars, is that of fundamental rights. Fundamental rights, in this context, refer to the basic rights and freedoms guaranteed to individuals, typically protected through various legal mechanisms. In the United States, these are primarily enshrined in the Constitution and its amendments. In the European Union, they are protected at multiple levels: through the EU Charter of Fundamental Rights, national constitutions and laws of member states, and the European Convention on Human Rights.

These rights form the cornerstone of democratic societies and often include freedoms such as expression, privacy, and access to information. As TDM technologies continue to advance, they increasingly interact with and sometimes challenge these fundamental rights, raising important questions about the balance between technological progress and individual protections.

Some scholars indirectly examine the interplay of TDM and fundamental rights, by exploring how TDM regulatory frameworks like those examined above, might interfere with fundamental rights. Others, directly examine the act of TDM and how that might interact with fundamental rights. Several human rights may be at risk, depending on the type of TDM being carried out, and under which circumstances. This section aims to present the literature that examines both the direct and indirect impacts of TDM on fundamental rights in the U.S. and the EU.

Indirect. A series of academic papers examine the recent EU CDSM TDM-specific exceptions, presented above, but under the guise of how such exceptions interact with fundamental rights. For instance, Manteghi argues that the balancing mechanisms regulating their scope of application, may restrict one's freedom of information and research enshrined under the freedom of expression.[209] Manteghi argues that the restriction of Article 3 of the CDSM to public sector entities, excluding private actors like journalists, individual researchers, and businesses raises concerns about restricting the freedom of information and research.[210] In an earlier paper, Manteghi argues that Article 3 and 4 of the CDSM which introduce the TDM-specific exceptions "*are likely to create uncertainty about their coverage and scope, which possibly may undermine the users' right to information*".[211] Similarly, Geiger argues that the EU's recent strategies for regulating AI resembles, a "*football team that would be left without any strikers to score successfully and to win any of the competitions with other jurisdictions which may have the advantage of more flexible legal provisions allowing broader TDM activities*".[212] Geiger argues that by restricting the scope of Article 3 of the CDSM, it may interfere with journalists' freedom of expression highlighting the

[201] ibid.

[202] Jaideep Vaidya, Yu Michael Zhu and Chris Clifton, *Privacy Preserving Data Mining*, vol 19 (Springer US 2006) <http://link.springer.com/10.1007/978-0-387-29489-6> accessed 23 November 2023.

[203] Murat Kantarcioğlu, Jiashun Jin and Chris Clifton, 'When Do Data Mining Results Violate Privacy?', *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (Association for Computing Machinery 2004) <https://dl.acm.org/doi/10.1145/1014052.1014126> accessed 23 November 2023.

[204] ibid.

[205] ibid.

[206] ibid.

[207] ibid.

[208] ibid.

[209] Maryna Manteghi, 'In Search of Balance: Text, Data Mining and Copyright in the Digital Single Market Directive from a Fundamental Rights Perspective - UTU Research Portal - UTU Research Portal' (2023) 48 European Law Review 443.

[210] ibid 448–449.

[211] Maryna Manteghi, 'Text and Data Mining in the EU: Managing a Conflict Between Copyright and the Right to Information' (2021) 43 European Intellectual Property Review 701.

[212] Christophe Geiger, 'The Missing Goal-Scorers in the Artificial Intelligence Team: Of Big Data, the Fundamental Right to Research and the Failed Text and Data Mining Limitations in the CSDM Directive' [2021] Joint PIJIP/TLS Research Paper Series 5.

"Panama Papers" scandal as an example which may fall outside the scope of this new exception.[213] Geiger added that the Commission "*risked creating an ineffective and therefore rapidly obsolete provision, in particular regarding the development of artificial intelligence, but also with regards to other activities of essential research and innovation not conducted by public bodies"*.[214] Scholars' views regarding the narrow scope given to the EU TDM-specific exceptions were presented above, whereby their application is limited to content that is "lawfully accessed". Manteghi argues that this concept should be interpreted in light of the term "lawful use" thereby ensuring that any harm caused by this scoping is proportional to the benefits that these exceptions are likely to provide in terms of copyright protection.[215] Furthermore, Synodinou argues that "lawful access" requires the adoption of a broad meaning within the confines of TDM extending to *"any use which is not restricted by law"*.[216] Hence, any means by which lawful use has been obtained will inevitably allow for the application of those EU CDSM TDM-specific exceptions, arguably proportionally safeguarding one's right to information and research.[217]

Direct. Berendt and Preibusch, conducted an empirical study designed to investigate how people make decisions with the assistance of data mining tools and how discrimination arises in these semi-automated decisions.[218] The authors argue that TDM may in some instances create or perpetuate discriminatory outcomes ultimately interfering with one's freedom from discrimination, which in the EU is enshrined in the GDPR.[219] This is one example of how TDM may directly interfere with fundamental rights. Berendt and Preibusch point out that data-driven decisions can impact individuals' rights by reinforcing existing biases or introducing new forms of discrimination, especially when based on sensitive attributes like race, gender, or nationality.[220] The authors advocate for a combination of algorithmic fairness tools, human oversight, and organisational accountability to reduce discrimination in data mining processes.[221] Similarly, but in the context of the US, Barocas and Selbst, argue that "*algorithms are only as good as the data they work with"* and if the data reflects historical biases or social inequalities, the algorithm applied to such data will likely reproduce those biases in its decision-making processes.[222] In the US, discrimination can be proven through "disparate impact", which focuses on outcomes rather than intent, and a seemingly neutral practice disproportionately affects a protected class.[223] Data mining can cause disparate impact by unintentionally discriminating against vulnerable groups due to biased data or flawed feature selection.[224]

Tien, examining the US jurisdiction, considers the extensive information held by the US government on individuals, stored across various databases, raising concerns about civil liberties such as privacy, freedom of association, and freedom of speech questioning whether the Fourth Amendment, designed to protect against unreasonable searches and seizures, constrains government data mining activities.[225] "*The Fourth Amendment's basic purpose is to safeguard the privacy and security of individuals against arbitrary invasions by governmental officials"*.[226] The author argues that the identification of patterns by mining across several databases, constitutes a search and thus cannot be knowingly exposed, and therefore, subjects of such searches deserve Fourth Amendment protection.[227]

In summary, several scholars examine how the recent EU CDSM TDM-specific exceptions may interfere with one's freedom of information and research, whilst another optic angle is provided by authors who assert that TDM itself might interfere with one's right to privacy or freedom from discrimination. As aforementioned, the interplay between TDM and fundamental rights is potentially multifaceted, depending on the type of TDM being carried out, and under which circumstances. For instance, one's right to property which acts as the human rights foundation for the protection of intellectual property rights, enshrined in the EU under Article 17(2) of the Charter of Fundamental Rights, which states that "*intellectual property rights shall be protected*", may be interfered with when conducting TDM. Nevertheless, the CJEU clarified the scope of this right in stating that whilst "*protection of the right to intellectual property is indeed enshrined in Article 17(2) of the Charter of Fundamental Rights of the European Union ('the Charter'). There is, however, nothing whatsoever in the wording of that provision or in the Court's case-law to suggest that that right is inviolable and must for that reason be absolutely protected"*.[228] Hence, one's right to intellectual property may be interfered with under certain circumstances. For this to

[213] ibid 6–7.

[214] ibid 7.

[215] Manteghi, 'In Search of Balance: Text, Data Mining and Copyright in the Digital Single Market Directive from a Fundamental Rights Perspective - UTU Research Portal - UTU Research Portal' (n 209) 452.

[216] Tatiana Eleni Synodinou, 'Lawfulness for Users in European Copyright Law: Acquis and Perspectives' (2019) 10 JIPITEC para 20 <https://www.jipitec.eu/issues/jipitec-10-1-2019/4876>.

[217] Manteghi, 'In Search of Balance: Text, Data Mining and Copyright in the Digital Single Market Directive from a Fundamental Rights Perspective - UTU Research Portal - UTU Research Portal' (n 209) 453.

[218] Bettina Berendt and Sören Preibusch, 'Toward Accountable Discrimination-Aware Data Mining: The Importance of Keeping the Human in the Loop—and Under the Looking Glass' (2017) 5 Big Data.

[219] ibid.

[220] ibid.

[221] ibid.

[222] Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2016) 104 California Law Review 671.

[223] ibid 701–712.

[224] ibid 674.

[225] Lee Tien, 'Privacy, Technology and Data Mining' (2004) 30 Ohio Northern University Law Review 389.

[226] ibid 400.

[227] ibid 408–415.

[228] C-70/10. *Scarlet Extended v. SABAM,* [2011] ECR I-11959 para. 43.

happen, the CJEU requires a "*balance to be struck between the various fundamental rights protected by the European Union*".[229] Thus, it can be concluded that another area of law which may come into play when conducting TDM is that of fundamental rights. The right interfered is dependent upon the type of TDM being carried out, and under which circumstances.

TDM and contract law

So far, the arguments presented above, to a great extent, assume that no contract applies. It is, however, possible that contracts exist between hosts of databases and datasets and those users which want to access that data for mining purposes. Such contracts may restrict TDM from taking place.[230] For instance, many scientific publishers offer licenses which explicitly prohibit TDM from taking place.[231] Others apply their contracts through "terms of use" on their websites, which often also restrict TDM from taking place.[232] This section aims to present scholars' views on the area where contract law and TDM meet.

The CDSM has arguably brought about changes in this intersection between contract law and TDM.[233] Above, author's views regarding Article 4 of the CDSM were presented which arguably allows for rightsholders to opt-out from TDM activities to a certain extent restricting TDM through contracts. Nonetheless, Article 7(1) of the directive explicitly restricts contracts from overriding the mandatory exception found under Article 3 of the CDSM.[234] However, as discussed in previous sections and as pointed out by Meys, "*in terms of material scope, the exception is limited to subject matter to which the beneficiaries have lawful access",* which creates a contradiction as it conceptually implies that this exception will only apply to TDM that makes use of content that is freely available online or accessed based on "*open access policies or contractual arrangements such as subscriptions*".[235] Bearing in mind the discussion above, this lies on the opposite side of the legal spectrum when compared to its US counterpart where Carroll argued that in the US, users would, in theory be able to access, download and use works from Sci-Hub for the purposes of TDM.[236]

Ducato and Strowel, examined the real-world impact contracts have on users who wish to conduct TDM.[237] The authors examined the terms and conditions of "*21 online platforms, equally distributed among three sectors: mobility (carpooling and car sharing); accommodation (including services for sharing office space); and food (i.e. initiatives for the recuperation of unsold or unused food, sharing or delivery of home-cooked meals, etc.)*".[238] Ducato and Strowel found that almost all (20 out of 21) websites examined either expressly or indirectly excluded TDM from taking place.[239] The authors assert that their analysis "*shows a trend toward the general contractual ban of TDM*".[240]

McCracken and Raub who examined the difficulties researchers may face when conducting TDM in the US assert that "*in some ways, TDM licensing still seems like a Wild West, where too many vendors are taking too many approaches, causing too many librarians to have to figure out far too many different licensing options*".[241] The authors attempted to establish "*access for a collection of faculty and graduate students from a range of disciplines, departments, and even universities".*[242] When doing so the authors encountered significant and remarkable limitations in how much they were able to achieve with different content providers.[243]

In a blog post Hugenholtz argues that the EU CDSM TDM-specific exceptions "*effectively create and legitimise a derivative market for text and data mining, which right holders may wish to control, license or even entirely prohibit".*[244] Manteghi asserts that this may in turn "*suppress the ability of users to access, share, or express protected content across the*

[229] C-469/17 *Funke Medien NRW GmbH v. Bundesrepublik Deutschland,* ECLI:EU:C:2019:623, para. 53.

[230] See for instance, Rosati (n 11) 12 who discusses that contracts may restrict TDM at the point that she calls the second step of TDM: Extraction and/or copying of content or Dorian Cougias and others, *Data Protections and Licenses Affecting Text and Data Mining for Machine Learning* (2022) who argue that the complex interplay of copyright and the resulting need to license unreasonably restrict TDM in the research context.

[231] Truyens and Van Eecke (n 21).

[232] ibid.

[233] Romain Meys, 'Data Mining Under the Directive on Copyright and Related Rights in the Digital Single Market: Are European Database Protection Rules Still Threatening the Development of Artificial Intelligence?' (2020) 69 GRUR International 457.

[234] Directive (EU) 2019/790, Article 7(1).

[235] Meys (n 233) 465.

[236] Carroll (n 25) 954–959. See section "1. TDM and Intellectual Property Rights".

[237] Rossana Ducato and Alain Strowel, 'Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility"' (2019) 50 IIC - International Review of Intellectual Property and Competition Law 649.

[238] ibid 668.

[239] ibid 674.

[240] ibid.

[241] Peter McCracken and Emma Raub, 'Licensing Challenges Associated With Text and Data Mining: How Do We Get Our Patrons What They Need?' (2023) 11 Journal of Librarianship and Scholarly Communication 13 <https://www.iastatedigitalpress.com/jlsc/article/id/15530/> accessed 30 November 2023.

[242] ibid 3.

[243] ibid.

[244] P Bernt Hugenholtz, 'The New Copyright Directive: Text and Data Mining (Articles 3 and 4)' (*Kluwer Copyright Blog*, 24 July 2019) <https://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/> accessed 28 November 2023.

public and private spheres".[245] The author highlights the difficulties involved in obtaining rights clearance, which will indirectly benefit very large commercial companies or internet giants, whilst hindering start-ups and SMEs ability to progress in the field of AI.[246] As per Geiger "*take it or leave it contractual conditions make access conditional upon accepting providers' terms, while obtaining specific permission from various publishers to carry out TDM research can be extremely complex*".[247] Manteghi highlights the need for Member States to deviate from this opt-out mechanism in a way in which is beneficial for start-ups and SMEs involved in AI development.[248]

On the contrary, Senftleben argued that the opt-out mechanism for TDM under Article 4 of the CDSM could be transformed into frameworks that enhance the collection of descriptive and ownership data in favour of rightsholders.[249] Senftleben argues that these provisions provide an opportunity for the EU to enhance licensing opportunities for rightsholders.[250] If Europe fails to develop large-scale, low-cost licensing solutions, AI training may shift to regions with better infrastructure, risking significant revenue loss for the European creative industries.[251] Therefore, the creation of a comprehensive and interoperable data infrastructure is urgent to capitalize on these emerging opportunities.[252] Hence, unlike what other scholars have argued, where complex licensing agreements may ultimately stifle TDM research, Senftleben argues that Article 4 of the CDSM provides the opportunity to enhance metadata, streamlining licensing processes, allowing for rightsholders to more easily license their content for TDM purposes. On a similar note but using a different line of argument for a proposed solution, Tylec *et al.*, highlight the difficulties involved with identifying works used in training AI, and ensuring rightsholders are equitably remunerated.[253] Tylec *et al.* propose a solution similar to the private copying levy in the InfoSoc Directive.[254] With their proposed solution, AI developers would be allowed to use copyrighted works for training algorithms without needing individual licenses for each work. Instead, they would pay a statutory fee to compensate copyright holders, and the collected fees would be managed by collective management organisations.[255]

In conclusion, contracts, and licensing agreements between hosts of datasets and those who wish to access said data for mining purposes, currently act as a barrier to free mining. To conduct lawful mining, users must navigate complex contractual terms and agreement procedures. Whilst some scholars, propose solutions for streamlining licensing agreements for TDM, there is still a gap in formal black letter law in this regard.

TDM and competition law

Vesala proposes that one way in which these contractual hurdles for TDM may be addressed is through competition law.[256] This takes us onto the next section of this literature review which presents author's views on the interplay between TDM and competition law. Vesala argues that EU competition law can assist in overcoming the obstacles created by contract law "*by requiring copyright holders to provide access to works and grant licences for their use"*.[257] Nevertheless, Vesala acknowledges the fact that such legal remedies offered by legal frameworks which regulate competition, may only be sought "*when the practices concerned constitute abuse of a dominant position or a restrictive agreement, which is not automatically or even generally the case when access to or use of copyright protected materials is limited*".[258]

Beyond the remedies offered by EU competition law in overcoming contractual hurdles users conducting TDM may encounter, Kochelek examined the potential of using TDM to act anticompetitively, examining how such practices may come into conflict with US antitrust laws.[259] Kochelek argues that when data mining practices are employed against aggregated customer data, thereby enabling the identification of a consumers' so-called "pain-point" i.e., the most undesirable value at which a party will still engage in a transaction, this in turn enables price discrimination.[260] The author ultimately argues that enforcing US antitrust rules against data-mining based price discrimination aligns with the US Sherman Act principles, which prescribe the rules governing free competition.[261]

Hence, competition law may act as a bridge in overcoming contractual restrictions for users conducting TDM using datasets

[245] Maryna Manteghi, 'The Insufficiency of the EU's Text and Data Mining Exceptions for Using Artificial Intelligence' [2022] European Intellectual Property Review 660–661 <https://www.westlaw.com/Document/I08AA9B604C1B11EDBF0AD27084713A21/View/FullText.html?transitionType=Default&contextData=(sc.Default)&VR=3.0&RS=cblt1.0&sp=wlnintell-000> accessed 28 November 2023.

[246] ibid 661.

[247] Geiger, Frosio and Bulayenko (n 37) 826.

[248] Manteghi, 'The Insufficiency of the EU's Text and Data Mining Exceptions for Using Artificial Intelligence' (n 245) 661.

[249] Martin Senftleben, 'Copyright Data Improvement for AI Licensing – The Role of Content Moderation and Text and Data Mining Rules' (4 May 2024) <https://papers.ssrn.com/abstract=4817796> accessed 29 May 2024.

[250] ibid 6–8.

[251] ibid.

[252] ibid.

[253] Grzegorz Tylec and others, 'Is It Possible to License Works Used in the Learning Process of Artificial Intelligence Algorithms?' (16 February 2024) <https://papers.ssrn.com/abstract=4729495> accessed 29 May 2024.

[254] ibid.

[255] ibid 12–13.

[256] Juha Vesala, 'Developing Artificial Intelligence-Based Content Creation: Are EU Copyright and Antitrust Law Fit for Purpose?' (2023) 54 IIC - International Review of Intellectual Property and Competition Law 351.

[257] ibid 368.

[258] ibid.

[259] Kochelek (n 10).

[260] ibid 520–523.

[261] ibid 526–534.

held by hosts which abuse their dominant position. Similarly, TDM may be used to enable anticompetitive practices.

Generalist approach

In addition to examining specific regulatory frameworks like copyright and the sui generis database right, data protection, contract law, fundamental rights, and competition law, several authors adopt a broader, "generalist approach" to the legal analysis of TDM. Rather than focusing on one particular area of law, these pieces of literature explore the intersection of multiple legal regimes and how they collectively influence the practice of lawful TDM. Their work often highlights overarching legal challenges such as inconsistencies across legal frameworks, the fragmentation of regulations, or the cumulative impact of multiple legal constraints, faced by those seeking to navigate between "free mining" and lawful, legally compliant TDM. This section presents literature that approaches TDM regulation to a certain extent, more holistically, offering cross-cutting insights into how different legal regimes converge to either facilitate or hinder TDM activities. In contrast to the prior sections, which analyse regulatory frameworks in isolation, these works provide a more integrated view of the legal landscape.

In a pre-Brexit era, Brook highlighted the EU regulatory framework governing TDM drawing examples from the UK's approach.[262] The author stresses that copyright protection and the subsequent potential for copyright infringement acts as a barrier to users who wish to conduct TDM, examining the recent, at the time, introduction of a TDM-specific UK copyright exception for TDM carried out for the sole purpose of research for a non-commercial purpose, provided copying is accompanied by sufficient acknowledgement, except where this would be impossible.[263] The author nonetheless, argues that with the introduction of that exception "*numerous constraints have, for UK researchers, been removed at a stroke*".[264] The author then highlights the difficulties database rights create for researchers wishing to conduct TDM. The author whilst only briefly describing the issues associated with database rights concludes that "*as with copyright, the law here is often unclear and acts as a chilling effect and barrier to many*".[265] Brook then moves onto discuss the issues created by complex licensing procedures highlighting the importance of creating open-access scientific literature through the use of creative commons licensing.[266] In an era preceding the CDSM, the author calls for a collective harmonised approach at EU level.[267] Hence, it can be deduced that Brook, in a pre-Brexit and CDSM era, highlighted some of the same difficulties that more recent scholarly work reviewed above, had also discussed.

In a 2014 study funded by the European Commission, when the UK was on the cusp of introducing their TDM specific exception, Triaille *et al.*, examined the copyright issues relevant to TDM in the EU in-depth, providing first a working definition of TDM and access to data choosing to use the term "data analysis" as an all-encompassing term.[268] The authors then analyse the exclusive rights afforded by copyright protection and how TDM might interfere with those, but also the potentially applicable exceptions at the time.[269] Triaille *et al.* subsequently examine other legal provisions that may be of relevance in the context of TDM, briefly highlighting issues relating to data protection, contract law and unfair competition.[270] The authors conclude by suggesting the introduction of a "new "data analysis exception" applying solely to TDM, which greatly resembles the CDSM exceptions.[271]

Truyens and Van Eecke provide an analysis of the regulatory frameworks applicable to text mining, distinguishing it from data mining.[272] The authors examine how text mining interacts with copyright law analysing the criteria for assessing originality and the applicability of exceptions.[273] The authors then describe how text mining might come into conflicts with database rights, and contract law,[274] before providing their assessment claiming that a legal issue within the context of text-mining is fact-dependant and that "*only for some delineated areas (very old texts, legal statutes, texts in the public domain) strong legal certainty can be obtained without case-by-case assessments*".[275] The authors suggest that rigid legal rules in the EU might hinder scientific research and technological developments related to text mining.[276]

Guadamuz and Cabel, focus on an analysis of content mining in research institutions and the challenges faced by both users and creators and as aforementioned, claim that legal issues arise at the point at which a user conducting TDM obtains access to their dataset and apply mining tools to said dataset.[277] Guadamuz and Cabel also focus on an analysis of copyright law, database rights and licensing, calling for an increase in open access policies enabling greater access to mining.[278]

[262] Michelle Brook, Peter Murray-Rust and Charles Oppenheim, 'The Social, Political and Legal Aspects of Text and Data Mining (TDM)' (2014) 20 D-Lib Magazine <http://www.dlib.org/dlib/november14/brook/11brook.html> accessed 30 November 2023.

[263] ibid.

[264] ibid.

[265] ibid.

[266] ibid.

[267] ibid.

[268] De Wolf & Partners and others, *Study on the Legal Framework of Text and Data Mining (TDM)* (Publications Office of the European Union 2014) <https://data.europa.eu/doi/10.2780/1475> accessed 27 June 2022.

[269] ibid 28–84.

[270] ibid 89–93.

[271] ibid 116–117.

[272] Truyens and Van Eecke (n 21).

[273] ibid 2182–2183.

[274] ibid 2184–2185.

[275] ibid 2185.

[276] ibid.

[277] Guadamuz and Cabell (n 28).

[278] ibid.

Meys, examines the existing limitations on data reusability imposed by database rights and contracts, whilst analysing the EU CDSM TDM-specific exceptions.[279] The author ultimately concludes that copyright, sui generis database rights and contracts pose obstacles to data re-usability for AI development.[280]

Rosati who as mentioned in the first section of this review, identifies three general TDM processes "*1) Access to content, 2) Extraction and/or copying of content, and 3) Mining of text and/or data for knowledge discovery*" each of which with their own legal consideration, also adopts a generalist approach highlighting several areas of law that may make up the regulatory framework governing TDM.[281] Namely, when analysing the first TDM process – "access to content" – Rosati discusses contractual issues that may arise in terms of obtaining a license to mine.[282] The added difficulties that may arise when dealing with orphan works and out-of-commerce works are also discussed.[283] When examining the second TDM process – "*extraction and/or copying of content"* – Rosati highlights the legal concerns that may arise within the context of copyright and sui generis (database) rights.[284] The author then finally notes that other areas of law may be relevant at this stage such as data protection/privacy laws, and contract law.[285] In the third stage of TDM – "*mining of text and/or data and knowledge discovery"* – the author provides a procedural analysis of what a typical TDM activity would involve, in terms of knowledge discovery.[286] Upon analysing these TDM processes, Rosati then moves onto a jurisdictional comparison of "*jurisdictions that display a more frequent and advanced use of TDM"*, examining the US's fair use, the UK's TDM exception and the EU's CDSM TDM-specific exceptions.[287] Rosati concludes by stating that as EU Member States transpose the CDSM into their laws, caution is urged to ensure that EU businesses engaging in TDM are not disadvantaged compared to non-EU counterparts, considering TDM's potential impact on AI and AI-generated creativity.[288]

Hammon *et al.* discuss the rapid growth of AI applications particularly since the release of ChatGPT in November 2022, emphasising the need for businesses to consider various legal and compliance risks associated with the development of AI.[289] The authors provide an overview of the regulatory frameworks such businesses are likely to encounter, such as contractual protection and AI specific regulations such as the EU's AI Act discussed above.[290] Hammon *et al.* then separate their analysis into input and output risks, examining regulatory risks associated with what the AI is fed as input and what output an AI is likely to produce respectively.[291] In terms of input risks the authors highlight that there are likely data protection and intellectual property risks.[292] Similarly with regards to output risks the authors highlight the intellectual property risks associated.[293] Hammon *et al.* then examine what they call "overarching considerations", highlighting the environmental risks related to the energy consumption of AI systems, social risks related to the takeover of a business's workforce by AI, and the importance of ethical policies, audits, and alignment with corporate values in AI adoption is emphasised.[294]

Finally, in a recent paper, Sean M Fill-Flynn summarises the key points that have emerged out of the wider project on the Right to Research in International Copyright. First, the author highlights the importance of TDM which involves analysing digitized information and is crucial for various AI tools, including generative AI models whilst it is also widely used in academic research across different fields.[295] Second, the importance of TDM is emphasised, demonstrating how it is linked to fundamental rights like receiving and imparting information and benefiting from culture and science, often referred to as the "*right to research*".[296] Finally, the author exemplifies the barriers that copyright created for TDM research. First, a patchwork of frameworks across multiple jurisdictions create barriers for TDM research especially in cross-border projects.[297] Second, licensing models for regulating access to text and data hinder research and economic activity, particularly affecting resource-dependent researchers and AI industries, especially in developing countries.[298] Open research exceptions could benefit TDM research, and claims of harm to creators from open exceptions lack empirical support.[299] Finally, there is a trend towards open exceptions especially for research purposes and the right to research should be advocated for in international copyright law.[300]

Conclusion

Currently, users who conduct TDM must navigate this complex web of legal rules without a map. This research as a whole,

[279] Meys (n 233).

[280] ibid.

[281] Rosati (n 16).

[282] ibid.

[283] ibid.

[284] ibid.

[285] ibid.

[286] ibid.

[287] ibid.

[288] ibid.

[289] Catherine Hammon and others, 'What Risks Need to Be Considered by a Business Using Artificial Intelligence?' (2023) 12 Compliance & Risk.

[290] ibid.

[291] ibid.

[292] ibid.

[293] ibid.

[294] ibid.

[295] Sean Fiil-Flynn, 'Copyright and Generative AI: Insights from the Project on the Right to Research' [2023] Joint PIJIP/TLS Research Paper Series 2 <https://digitalcommons.wcl.american.edu/research/112>.

[296] Fiil-Flynn (n 295).

[297] ibid 2–3.

[298] ibid 3.

[299] ibid 3–4.

[300] ibid 4.

which this literature review forms part of, aims to create a practical guide and tool facilitating lawful TDM to take place.

What this literature review makes evident is that literature examining "lawful TDM" is fragmented. Authors have examined how TDM interacts with copyright and the sui generis database right, data protection, contract law, fundamental rights, and competition law each providing valuable piecemeals of insights into what regulatory frameworks users conducting TDM may encounter.

Whilst some attempt to provide a more general interpretation of how one can conduct "lawful TDM", this literature review illuminates the fact that they are not all-encompassing. The regulatory frameworks that apply to users conducting TDM are multifaceted, and ever-changing on a case-by-case basis, which heavily depends upon the type of mining being conducted, where mining takes place, the dataset involved, how such data is accessed, and how the output of mining processes will ultimately be used. There is an ever-growing need for a holistic interpretation of the regulatory frameworks which apply, creating a map which would allow for users conducting TDM to navigate this complex web of legal rules.

Data availability

Extended data

Zenodo: Appendix for Review Paper Entitled: "A literature review of "lawful" text and data mining." (https://zenodo.org/records/13880018) (Vrakas 2024)

The project contains the following extended data:

Appendix

Data is available under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0).

Open Peer Review

Current Peer Review Status: ? ✓ ?

Version 2

Reviewer Report 04 November 2024

https://doi.org/10.21956/openreseurope.20215.r45724

© **2024 Nordberg A.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✓ **Ana Nordberg**
Lund University, Lund, Sweden

I have no further comments.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Intellectual Property Law; Data protection; Technology regulation.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 12 August 2024

https://doi.org/10.21956/openreseurope.19468.r42476

© **2024 Trigo Kramcsak P.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

? **Pablo Rodrigo Trigo Kramcsak**
Vrije Universiteit Brussel, Brussels, Brussels, Belgium

I. Overview

The paper offers a valuable literature review on Text and Data Mining (TDM), mapping the complex regulatory framework surrounding these processes and shedding light on what

constitutes lawful TDM. However, several areas require improvement for the paper to reach its full potential. This report provides some comments and suggestions to strengthen the paper's structure, argumentation, and use of sources.

II. Comments

1) Introduction

- Structure improvement: The introduction could be better structured to present the study's subject matter clearly. The first two paragraphs are not sufficiently clear, particularly in introducing the concept of "big data." It would be beneficial to discuss the characteristic elements of big data and differentiate between structured, unstructured, and semi-structured data. This is crucial for readers to grasp the subject thoroughly. For reference, see Jonathan Grossman and Ami Pedahzur, [Ref:1]

- Clarification of concepts: The sentence 'data "is the new oil and lifeblood of economic development" and just as oil needs to be burnt for it to be useful, raw data must be processed to realise its true potential' is problematic and contested. This analogy oversimplifies the complexities of data processing in the digital economy. See, e.g., Clemens Mewald, 'Why Data Is *Not* the New Oil and Data Marketplaces Have Failed Us' (Medium, 13 July 2023) https://towardsdatascience.com/why-data-is-not-the-new-oil-and-data-marketplaces-have-failed-us-b42dd87a0ba0; David Moschella, 'Data Isn't the New Oil; That Might Be a Good Thing' (2023) https://itif.org/publications/2023/05/30/data-isnt-the-new-oil-that-might-be-a-good-thing/

2) Regulatory Framework

A broader perspective is needed: The paper should more comprehensively address the regulatory challenges of TDM, particularly from a data protection standpoint. This includes not only the technical challenges of big data processing but also the legal issues associated with scraping and mining of publicly available data. Including opinions from public agencies and supervisory bodies would offer a more holistic view of the regulatory landscape. These perspectives could be instrumental in identifying the main challenges to individual rights and public interests.

3) Data Protection Issues

- Insufficient coverage of GDPR issues: The paper does not sufficiently address the complexities and challenges that data mining poses for data subjects. The paper relies on some sources of questionable quality. It would benefit from a more careful selection of references, particularly those recognised as authoritative in the field. I recommend citing more reliable sources, especially when describing the fundamental elements of GDPR, such as Christopher Kuner and others (eds), [*Ref:2],A Commentary* (New York, 2020; online edn, Oxford Academic); Indra Spiecker genannt Döhmann and others (eds), *General Data Protection Regulation: Article-by-article Commentary* (München, 2023; Beck).
- Concept of "meaningful research": The paper should elaborate on the idea of "meaningful research" and how data protection regulations address data processing for scientific, research, or statistical purposes. This is a critical aspect that currently needs more depth.

- Data Protection by Design and Default: When discussing the implementation of the principle of data protection by design and default, it is important to reference guidelines issued by supervisory authorities. For instance, the EDPB Guidelines 4/2019 on Article 25 Data Protection by Design and by Default [Ref 5]; Commission Nationale d'Informatique et des Libertés (CNIL), GDPR guide for developers (2022); Agencia Española de Protección de Datos (AEPD), A Guide to Privacy by Design (2019).

- Data mining, profiling and inferences: The paper should more comprehensively address the challenges and risks of data mining, especially concerning profiling and the processing of inferred data. The relevance of purpose limitation and data minimisation principles needs more emphasis. Key references include Bart Custers and others (eds), [*Ref*:3,4]

- Missing legal aspects. Data scraping: The paper currently lacks a discussion on the legal aspects of data scraping at scale and text mining of publicly available data, which are increasingly relevant, especially for training AI models. It would be beneficial to incorporate an analysis of the legal challenges posed by these practices, e.g., Benjamin Bremert, *Legal Aspects of Text Mining Publicly Available Data*, Unabhängiges Landeszentrum für Datenschutz (ULD), 2017; Zuzanna Gulczynska, 'Scraping personal data from internet pages - a comparative analysis of the Polish Bisnode decision and the US hiQ Labs v LinkedIn Corp judgment', European Law Review, vol. 45, no. 6, pp. 857–869, 2020; Catherine Altobelli and others, 'To Scrape or Not to Scrape? The Lawfulness of Social Media Crawling under the GDPR', in Jean Herveg (ed), *Deep Driving into Data Protection: 1979-2019: Celebrating 40 Years of Privacy and Data Protection at the CRIDS* (Larcier, 2021); Autoriteit Persoonsgegevens, Richtlijnen scraping door private organisaties en particulieren (2024)

4) TDM and fundamental rights

Effectively addressing bias and discrimination in data mining is critical to safeguarding fundamental rights. Failure to do so can result in inequitable outcomes that disproportionately harm marginalised groups, perpetuating existing social inequalities. See Bettina Berendt and Sören Preibusch, 'Toward Accountable Discrimination-Aware Data Mining: The Importance of Keeping the Human in the Loop-and Under the Looking Glass', Big Data. 2017 Jun;5(2):135-152; Solon Barocas and Andrew Selbst, 'Big Data's Disparate Impact', 104 California Law Review 671 (2016); Solon Barocas, 'Data Mining and the Discourse on Discrimination', Conference on Knowledge Discovery and Data Mining (2014).

5) Several footnotes and cross-references need to be revised and improved for accuracy and clarity.

References

1. Grossman J, Pedahzur A: Political Science and Big Data: Structured Data, Unstructured Data, and How to Use Them. *Political Science Quarterly*. 2020; **135** (2): 225-257 Publisher Full Text
2. The EU General Data Protection Regulation (GDPR). 2020. Publisher Full Text
3. Discrimination and Privacy in the Information Society /authors.
4. A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI/authors.

Is the topic of the review discussed comprehensively in the context of the current

literature?
Yes

Are all factual statements correct and adequately supported by citations?
Partly

Is the review written in accessible language?
Partly

Are the conclusions drawn appropriate in the context of the current research literature?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Personal data protection, data governance, AI regulation

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 02 Oct 2024

Giorgos Vrakas

Thank you very much for your constructive and insightful comments. I have thoroughly revised the paper in light of your feedback, and I believe these changes have significantly strengthened the review. Below, I address each of your comments and explain the adjustments made.

1. **Introduction: Structure and Clarity**

I have restructured the introduction to improve clarity. The concept of "big data" is now introduced with a more thorough explanation of its characteristic elements, including a clearer distinction between structured and unstructured data. I have also reconsidered the analogy of data being “the new oil,” and, following your advice, I have removed such statements.

1. **Regulatory Framework: Broader Perspective**

The section on the regulatory framework has been expanded to offer a broader and more comprehensive discussion of the legal challenges associated with TDM, particularly from a data protection standpoint. I have added references to opinions from public agencies and supervisory bodies to provide a more holistic view of the regulatory landscape. **Data Protection Issues: More Depth and Careful Source Selection** The data protection section has been substantially reworked to include more authoritative and recognised sources, as you recommended. Additionally, I have expanded on the concept of "meaningful research," elaborating on how data protection regulations address the processing of data for scientific, research, or statistical purposes. I have also incorporated references to the EDPB Guidelines 4/2019 on Article 25 (Data Protection by Design and Default), as you suggested. **Data Mining, Profiling, and Inferences** I have significantly expanded the discussion on the challenges and risks of data mining, particularly regarding profiling and the processing of

inferred data. **Legal Aspects of Data Scraping** I have added a paragraph discussing the legal aspects of data scraping citing the sources that you had recommended.

1. **Fundamental Rights: Bias and Discrimination**
    - I have reworked the section on fundamental rights to more thoroughly address the critical issues of bias and discrimination in data mining.
2. **Footnotes and Cross-References**
    - I have revised the footnotes and cross-references to improve their accuracy and clarity. This includes correcting any inconsistencies and ensuring that all references are properly cited and linked to the relevant sections of the paper.

Once again, thank you for your detailed feedback. I am confident that the revisions made in response to your comments have strengthened the paper both in terms of structure and substantive content.

Competing Interests: No competing interests were disclosed.

Reviewer Report 07 August 2024

https://doi.org/10.21956/openreseurope.19468.r42474

© 2024 Nordberg A. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Ana Nordberg
Lund University, Lund, Sweden

The paper provides a highly valuable overview of the literature on the topic of TDM. Exploring the literature in a meaningful manner and highlighting scholarly work conducted on this topic.
The following comments and suggestions are submitted for the author's consideration:

Methodology
The sources included in the review are mostly relevant and the analysis is sound however some aspects of the methodology section would benefit from further development:

1. The sources selected consist mostly of well-known authors and reputable publication venues. However, the characteristics of journals or other venues of publications either were not considered or are not described. Given the proliferation of low-quality, biased or otherwise unreliable publications, it should be addressed whether or not the study included a qualitative approach to the selection of sources, and if relevant how.
2. Table 1 categorizes the literature by topics (most of them of a legal nature), however, some of the sources appear to present other disciplinary perspectives. It is unclear whether the field of research of the literature analyzed was considered and how such influenced the qualitative analysis of the materials.
3. A large part of the literature analyzed is legal literature where the jurisdictional aspect is relevant, the methodology section does not clearly state the jurisdictional focus of the study

and whether comparative law or other methodologies were used in the qualitative analysis. Instead, the section 'jurisdictional variance' provides some information in this regard which could be more detailed.

4. Arguably, there is a strong possibility that a considerable body of legal literature has been published in the official languages of each jurisdiction, in particular in those countries that have enacted national legislation on TDM that predates Directive (EU) 2019/790 CDSM. However, the literature reviewed only included publications in the English language. If language was a selection factor it should be described and treated as an important methodological limitation or has a subject scope delimitation.

Readability

1. The text is well written with occasional less clear sentences. Some sentences and paragraphs are excessively long and complex.
2. The use of the plural 'they' to avoid gender pronouns can be confusing. Gender neutrality is important and can be better achieved with the person's name or neutral expressions, such as 'the author' or equivalent.
3. A few footnotes require revision and the entries in the appendix are inconsistently cited, with many being incomplete.

These points can easily be improved with the usual final text review, which is recommended.

Substantive content

1.Section 'Regulatory frameworks which govern TDM'
Would benefit from a clear structure and a definition of what is understood by "lawful TDM". The term lawful entails a presumption of compliance with a set of legal rules, thus it should be accompanied by clear specification of the said set of rules (including eventual conditions and exceptions).

2. Section 'TDM and Intellectual Property Rights'
The section does not address all IPRs, perhaps it would benefit from being more accurately named or at least the type of IPRs considered could be enunciated in the section's introductory sentences. Some statements in the text are excessively vague or not sufficiently explored/supported. For example, the statement below introduces several points of comparative law that are left 'hanging in the air', and are not explored further in the section. I would suggest either erasing ('kill the darling') or further elaborating.
'Jackson Bertón, who examined the applicable copyright frameworks of the five largest economies of South America in a pre-CDSM era, claimed that even then, the EU was falling behind other more competitive jurisdictions, but ultimately found that South American frameworks were not prepared for digital research techniques.'

3. Section ' TDM and Data Protection'
Here the reader faces difficulties with the jurisdictional aspect. Although the information is implied, it should be made more clear what is the jurisdictional origin and coverage of the legal instruments mentioned (e.g. GDPR). It is not always clear what jurisdiction was considered in the papers analysed.
Some of the papers analysed in this section were written in the early 2000's. The temporal aspect could be better scrutinized as to whether their findings remain unchanged by subsequent technological and legal developments.

4. Section 'TDM and Fundamental Rights'
What is understood by fundamental rights and what legal framework(s) are considered is not defined. The expressions 'fundamental rights', 'human rights' and 'fundamental human rights' are sometimes used interchangeably making those sentences confusing.

5.Section 'Generalist approach'
The qualification of the papers analysed under this section as 'generalist' is not very convincing. Perhaps the author could better justify their classification, or otherwise include this literature in previous sections.

Is the topic of the review discussed comprehensively in the context of the current literature?
Yes

Are all factual statements correct and adequately supported by citations?
Partly

Is the review written in accessible language?
Partly

Are the conclusions drawn appropriate in the context of the current research literature?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Intellectual Property Law; Data protection; Technology regulation.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 02 Oct 2024
Giorgos Vrakas

Thank you for your constructive feedback, which has been invaluable in helping me to strengthen this literature review. I have carefully addressed each of your comments and made the necessary revisions to the paper. Below, I provide a detailed response to your suggestions.

1. **Methodology: Selection Criteria and Language Limitation**

I have clarified the selection criteria for the literature included in this review. The introduction now explicitly discusses the section criteria used. I have also acknowledged that the study focused on publications in English due to the scope of the review, and I have identified this as a methodological limitation. Additionally, I have noted the possibility that significant legal literature may exist in other languages, particularly in countries with pre-existing TDM legislation, and this is now treated as a subject scope limitation.

1. **Use of 'They' and Readability**

Following your suggestion, I have replaced the use of the plural 'they' with specific references to 'the author' or the author's name to avoid confusion. I have also restructured several sections, shortening complex and overly long sentences to improve readability and clarity. A review of the footnotes has been conducted to correct inconsistencies, and the appendix has been updated to ensure all entries are complete and correctly cited.

1. **Definition of "Lawful TDM" in Regulatory Frameworks Section**

I have revised the section on regulatory frameworks to include a clearer definition of "lawful TDM." The term is now defined as "TDM that is considered compliant with applicable regulatory frameworks". This should make the concept more transparent for the reader.

1. **TDM and Intellectual Property Rights Section**

Based on your feedback, I have renamed the section 'TDM and Intellectual Property Rights' to 'TDM and Copyright and Related Rights' to better reflect its content. Furthermore, I have removed the portions that you suggested did not add sufficient value to the discussion, such as the statement regarding Jackson Bertón's analysis of South American copyright frameworks. This has helped to eliminate vague or unsupported statements, ensuring the section remains focused and well-supported.

1. **Fundamental Rights Section**

I have revised the section on TDM and Fundamental Rights to clarify what is meant by fundamental rights. The introductory paragraphs now define the legal frameworks considered, and I have ensured that the terms 'fundamental rights', 'human rights', and 'fundamental human rights' are used consistently to avoid confusion. This should provide a more structured and coherent discussion.

1. **Generalist Approach Section**

I have revisited the 'Generalist Approach' section, and I now provide a clearer justification for the classification of the papers analysed under this category. I explain why they are considered 'generalist' in their approach to TDM, distinguishing this category from other sections of the review. Once again, thank you for your insightful suggestions. These revisions have contributed significantly to improving the clarity, structure, and overall quality of the paper.

Competing Interests: No competing interests were disclosed.

Reviewer Report 30 July 2024

https://doi.org/10.21956/openreseurope.19468.r42477

© 2024 Margoni T. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Thomas Margoni
KU Leuven, Leuven, Belgium

This is a very relevant piece of research reviewing the literature in the field of TDM. As such it is important and, within the limits of the initial stage of the research project, original and well developed. A few observations can be found below that may be taken into considerations by the author in the next stages.

"Lawful tdm": not clear what it is, lacks definition.

Literature review: not clear in which disciplinary domains it was conducted. Are you assessing mainly legal sources, or also technical literature (computer science; digital humanities, etc)? There seems to be a need to define better the boundaries of the review and which concepts are assumed.

“Some scholars make a distinction between text mining and data mining.[18] Whilst this distinction may be vital on the technical side of TDM, altering the analysis techniques used, it arguably does not necessarily play a pivotal role in the regulatory frameworks which may apply. Data mining is sometimes seen as a subset of text mining since text mining is the computational process of analysing unstructured messy data, whilst data mining is the computational process of analysing structured data.[19] ”

General observation: It is not always clear where you are reviewing current literature and where you are taking a position in the debate. This should be made clearer. As a lit. rev. I would assume the former, but I have doubts here and there.

E.g.,: One could of course argue that text is more copyright focused, while data is less copyright and, if applicable, more sgdr/neighbouring rights with different regulatory impact. Is this something that you are addressing?

Lit rev. may need to be temporally structured: e.g., sources of 2022-2024 are discussed together with sources of 2014 (e.g., around fn 20 – 25 but also in other parts), but the underlying technology and regulatory framework have evolved considerably. Are there category commensurate?

“Bearing in mind the definition of TDM presented above” why not using the CDSM definition?

“One of these solutions, Manteghi claims could be found under Article 35 of the proposed Data Act which “*provides a safe harbour for users needing to access, use or share only databases made of data generated by the use of a product or related service (IoT data)”.*[83]“
Unclear. Perhaps clarify that Art. 43 (ex 35) Data Act, only refers to SGDR not copyright.

“On the other hand, Kretschmer *et al.*, who examined machine learning, natural language processing and computer vision for computer moderation as case studies where TDM is vital, found that legal uncertainties like those created by fair use exceptions, can lead to practices exploiting these uncertainties for commercial gain, particularly by large tech corporations.“
Same as above.

General observation: It seems that certain statements are somehow decontextualised from the original source and used in a way that changes the meaning. In the specific example above, it seems that Kretschmer is against fair use for tdm. If I recall it correctly, they are in favour,

however, they suggest that legal uncertainty (which can be found in various legal mechanisms including fair use) has, among its consequences, that of favouring larger players who have the financial resources to litigate/resist in court.

It may be necessary to revise that all the excerpts are correctly contextualised in order to ensure that the cited literature is assessed properly.

The part on fundamental right seems a bit rushed in comparison to copyright and DP and to conflate EU/US?

The “generalist approach” category and section is not clear. How is it different from the sections analysed above, particularly copyright? It also seems to rediscuss some of the same sources already analysed.

Conclusions: again it seems you are expressing an opinion about TDM, but can you do that on the basis of a lit. rev.? Perhaps, at this stage you should only focus on what emerges from lit. rev. (which you do in part at the end of the section).

Is the topic of the review discussed comprehensively in the context of the current literature?
Yes

Are all factual statements correct and adequately supported by citations?
Partly

Is the review written in accessible language?
Partly

Are the conclusions drawn appropriate in the context of the current research literature?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: IP, copyright, TDM.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 02 Oct 2024

Giorgos Vrakas

Thank you for your valuable feedback. I have carefully considered each of your comments and made revisions accordingly to improve the clarity and strength of the paper. Below, I

provide a point-by-point response to your suggestions.

1. **"Lawful TDM" is unclear, lacks definition:**

I have now briefly defined “lawful TDM” as “TDM which would be considered compliant with the applicable regulatory frameworks” in the relevant section, ensuring that this term is clearer for the reader.

1. **Literature review: disciplinary domains unclear:**

I have clarified in the introduction that no disciplinary boundaries were imposed in the selection criteria for the literature reviewed, to reflect the interdisciplinary nature of TDM. This literature spans legal, technical, and policy domains. This revision aims to better communicate the broad scope of the review.

1. **Clarification of position vs. review:**

I have made clearer distinctions between my own interpretations and the positions found in the literature. Where I offer commentary, it is now more clearly framed as a response or evaluation of the literature, rather than presenting it as a standalone position. This is particularly important for a literature review, and I appreciate your observation in this regard.

1. **Temporal structuring of the literature:**

Although I have chosen not to temporally structure the review, I have made it clearer when older sources are discussed alongside newer ones. I have provided explicit justification for the inclusion of these older sources, explaining why their contributions remain relevant despite technological and regulatory developments.

1. **Use of CDSM definition:**

I have revised the relevant section to incorporate the CDSM definition of TDM, as you suggested. This revision ensures greater coherence with existing EU legal frameworks.

1. **Clarification regarding Article 43 (formerly Article 35) of the Data Act:**

I have clarified that Article 43 (formerly 35) of the Data Act refers specifically to SGDR, not copyright, as you rightly pointed out.

1. **Contextualisation of Kretschmer et al.'s position on fair use:**

I have revisited the discussion of Kretschmer et al.'s work to ensure that it is accurately contextualised. The revision better reflects their nuanced position on fair use, highlighting their concern about legal uncertainty benefiting larger players, while also ensuring that their general support for fair use is not misrepresented.

1. **Fundamental rights section and conflation of US/EU perspectives:**

I have revised the fundamental rights section to avoid conflating EU and US perspectives. The revision now presents these frameworks more distinctly.

1. **Generalist approach and redundancy in the review:**

I have restructured the “generalist approach” section to avoid redundancy, ensuring that it builds upon prior sections rather than repeating the same sources or arguments. I have clarified how it differs from the copyright-focused discussions, making the overall structure of the paper more coherent.

1. **Conclusions:**

I have revised the conclusion to ensure that it focuses strictly on summarising the key findings from the literature review. I have removed any personal opinions or evaluative commentary that could be construed as going beyond the scope of a literature review.

Thank you once again for your insightful comments, which have contributed significantly to refining the paper.

Competing Interests: No competing interests were disclosed.