Navigating the Legal Landscape: Technical Implementation of Copyright Reservations for Text and Data Mining in the Era of AI Language Models

by **Lisa Löbling, Christian Handschigl, Kai Hofmann and Jan Schwedhelm ***

Abstract: The profound advancements in AI-driven language models, exemplified by ChatGPT, owe their existence to vast quantities of text and data utilized in their training. However, the origins of this data and its suitability for training AI models raise considerations in the domain of Text and Data Mining (TDM) and its associated copyright requirements.

European and German regulation provide an opt-out system for TDM: Freely available works may be used for TDM if they have not been reserved by the rightsholder. A reservation of use is effective only if it is made in a machine-readable format. On the one hand, state-of-the-art language models use large amounts of text data from different domains. On the other hand, no (de facto) standard for reservations of use has yet been established. In this paper, we will therefore

- discuss the legal requirements,
- give an insight into how usage reservations are dealt with in practice and
- suggest a possible standard.

Keywords: Copyright Law, Text and Data Mining (TDM), Artificial Intelligence (AI), Data Indexing and Crawling Restrictions, Machine-Readable Standard

Recommended citation: Lisa Löbling, Christian Handschigl, Kai Hofmann and Jan Schwedhelm, Different Navigating the Legal Landscape: Technical Implementation of Copyright Reservations for Text and Data Mining in the Era of AI Language Models, 14 (2023) JIPITEC 499 para 1.

A. Introduction

1 Text and data mining (TDM) is the process of using software to automatically analyze collections of text and data to extract information and compile insights. It has become increasingly important in recent years, as the amount of digital information available is growing exponentially.[1] Alongside simple rule-based and statistical methods, TDM also entails the application of advanced algorithms and computational techniques, specifically drawn from the field of natural language processing (NLP), to identify patterns, relationships, and trends in unstructured data like text documents. This can be, e.g., journal articles, scientific papers, press releases, social media posts, and books.

* Dr. Lisa Löbling, Senior Consultant at d-fine, lisa.loebling@d-fine.com

Christian Handschigl, Web Specialist & Consultant at abnorm media, christian@abnorm.de

Dr. Kai Hofmann, Scientific Desk Officer Law at the German Centre for Rail Traffic Research, hofmannk@dzsf.bund.de

Jan Schwedhelm, Consultant at d-fine, jan.schwedhelm@d-fine.com

1 David Reinsel, John Gantz and John Rydning, 'Data Age 2025: The Evolution of Data to Life-Critical' (International Data Corporation 2017) <https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf> accessed 17.01.2024.

2 After the initial step of defining the business goal or research question and the identification of relevant data that aligns with the use case, the process of TDM unfolds as follows: TDM starts with obtaining and preparing the source material from various digital (or non-digital) sources, making it machine-readable, normalizing, structuring, categorizing, and converting it into suitable technical formats. The processed source data forms a corpus, which is then automatically analyzed using specialized software or scripts to uncover statistical frequencies or correlations within the datasets. The inclusion of annotations, which are metadata accompanying normalized and structured content, varies based on the corpus and research focus. Training machine learning algorithms to uncover hidden patterns and correlations is also considered part of TDM, requiring the preparation of training data for self-learning systems, while the quality of the processed source data significantly impacts the knowledge gained through TDM.[2]

3 In practice, TDM serves as a powerful approach to gain valuable insights from vast volumes of data across diverse fields. Today, the most prominent tool in this domain is ChatGPT, a sophisticated language model developed by OpenAI.[3] ChatGPT has garnered attention for its ability to generate human-like responses and engage in interactive conversations, making it a valuable asset for applications such as chatbots, virtual assistants, and customer support systems. The model has undergone extensive training with vast and diverse text data from various domains. The inclusion of this substantial amount of known content during training plays a crucial role in enabling the chatbot to deliver convincing and innovative responses. In addition to its attention-grabbing applications, TDM is also employed for fundamental tasks, such as extracting entities (e.g., organizations, people, places, and events) from text, identifying sentiment and emotions, and classifying texts into different categories or topics.[4]

4 TDM encompasses a range of techniques, from rule-based analysis (e.g., regular expressions) via feature-based machine learning (e.g., linear regression, support vector machines, or random forests) to representation learning (e.g., GPT-3, BERT, and variants). When selecting a model, various factors need consideration – in any case the quantity and quality of training data significantly impact the model's accuracy and effectiveness. Thus, collecting relevant data for training NLP models plays a central role in TDM projects.

5 To perform TDM, the source material is initially duplicated and organized into a corpus for subsequent analysis. This source material may be subject to copyright protection, such as literary works (Art. 2 lit. a Directive 2001/29/EC, Art. 2 Berne Convention, Section 2(1) no. 1 UrhG[5]), significant parts of databases (Art. 7 Directive 96/6/EC) or press publications (Art. 15 Directive [EU] 2019/790). The extent of protection is contingent on specific conditions, resulting in typically partial protection of the material. However, since prerequisites such as "intellectual creation"[6] (relating to Art. 2 lit. a Directive 2001/29/EC) or "substantial investment" (Art. 7 Directive 96/6/EC) cannot be checked automatically, in practice one must assume that the material is protected.

6 TDM, in principle, requires permission from the copyright holder to proceed lawfully. Some websites and platforms acknowledge this aspect and offer Application Programming Interfaces (APIs) that enable developers to programmatically access data.[7] These APIs often facilitate complex query commands for downloading targeted information in large quantities. Moreover, APIs can be utilized to manage access rights for data collection, as they permit data owners to restrict access to specific datasets and define the level of access granted to each user by distributing individual access tokens. APIs are purposefully designed for efficient, controlled, and structured information exchange, making them a preferable option from a copyright perspective. Nevertheless, setting up an API is not practically useful for many websites since it does not align with the goal and use cases that focus on providing information for human users rather than prioritizing structured data access.

7 Thus, other methods for TDM are more commonly used like web scraping, which enables the retrieval of information from websites and data collections. This technique involves utilizing software to extract data from the HTML code of a website and converting it into a structured format suitable for

2 Thomas Dreier, § 44b UrhG, in Thomas Dreier and Gernot Schulze, Urheberrechtsgesetz (7th edn, CH Beck 2022) no 5.

3 OpenAI, 'Introducing ChatGPT', <https://openai.com/blog/chatgpt> accessed 17.01.2024.

4 Daniel Jurafsky and James H Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2008).

5 German Act on Copyright and Related Rights [Urheberrechtsgesetz]

6 ECJ, ECLI:EU:C:2018:899, no. 37 et seq., see also Section 2(2) UrhG.

7 For example, X/Twitter provides a developer API that allows for programmatic access to public X/Twitter data. See <https://developer.twitter.com/en/docs> accessed 17.01.2024.

analysis with NLP methods. Web scraping has gained popularity, as it enables users from various domains, including business and science, to efficiently gather data that would otherwise be time-consuming or impractical to collect manually. Unlike APIs, web scraping does not rely on access explicitly designed for TDM purposes. Instead, this method leverages the statutory exception from copyright protection provided by Art. 4 Directive (EU) 2019/790, making it particularly relevant for this type of mining.

B. Copyright exception for TDM

8 According to the general copyright exception for TDM in Art. 4 Directive (EU) 2019/790 – and its national transposition, for German law in Section 44b UrhG, it is permitted to reproduce lawfully accessible works and other subject matters in order to carry out TDM – regardless of the purpose of the TDM. Copies are to be deleted when they are no longer needed to carry out text and data mining, Art. 4(2) Directive (EU) 2019/790.

9 The TDM exception applies to all sorts of material protected by copyright or related rights. The only prerequisite is that the material is lawfully accessible. However, the TDM user does not have to check whether the works were made accessible with the consent of the rightsholder; instead, what matters is whether the TDM user has lawful access to the source where the material is found.[8]

10 Lawfully accessible means that the TDM user himself must be able to access the material, in the case of screen scraping by crawling the web (see Section C). This is why the TDM exception does not apply to user-generated content. If an end user of a TDM-based applications (e.g., ChatGPT or DeepL) enters third-party copyrighted material into this application, it is the responsibility of the end user to ensure that the usage of such material complies with relevant legal provisions. The application provider, on the other hand, is not allowed to use this material from this source for TDM – at least not by referring to the TDM in exception in Art. 4 Directive (EU) 2019/790. If the application provider wants to include user generated content in the training of its algorithms, he must establish mechanisms[9] to refrain from utilising content for TDM for which the rightsholders have not granted authorization. However, this issue does not fall within the scope of the TDM exception.

8 Thomas Dreier, § 44b UrhG, in Thomas Dreier and Gernot Schulze, *Urheberrechtsgesetz* (7th edn, CH Beck 2022) no 8.

9 The technical approach of C2PA (see Section E.I) is heading in this direction of law enforcement.

I. Opt-out

11 The most important limitation of the exception for TDM is set out in Section Art. 4(3) Directive (EU) 2019/790: The copyright holder may reserve the use of their copyrighted material for TDM purposes. Consequently, the general TDM exception does not apply when such a reservation has been made. Under these circumstances, utilization of the copyrighted material for TDM requires explicit permission from the copyright holder, who has the discretion to either prohibit TDM use entirely or make it subject to conditions such as remuneration.

12 This opt-out approach is at the core of the TDM regulation. The process of opting out entails distinct responsibilities for both the TDM user and the copyright holder:

- The TDM user bears the onus of proof, mandated by the phrasing of paragraph 3 ("are permitted only if they have not been reserved").[10] Thus, the user is required to substantiate that the copyright holder has not opted out, necessitating active searches for and documentation of relevant opt-outs.

- Conversely, the copyright holder is accountable for properly expressing their opt-out decision. While this stipulation derives from Article 4(3) of Directive (EU) 2019/790 ("on condition that the [...] has not been expressly reserved by their rightholders in an appropriate manner"). The copyright holder assumes the risk associated with the adequacy of their chosen method to communicate the opt-out.

13 In essence, the TDM user needs only to seek opt-outs that have been appropriately conveyed. The determination of appropriateness hinges on contextual factors, encompassing how the copyrighted material is made accessible and the degree of effort required for the TDM user to verify opt-outs. Consequently, a limited set of requirements can be generalized, such as ensuring opt-outs are articulated clearly and positioned where users are likely to encounter them. Furthermore, a reservation's impact is prospective[11]; if altered subsequently, reproductions already completed remain legal within the boundaries defined Art. 4(1) Directive (EU) 2019/790. Therefore, opt-outs need

10 Benjamin Raue, 'Die Freistellung von Datenanalysen Durch Die Neuen Text Und Data Mining-Schranken (§§ 44b, 60d UrhG)' [2021] Zeitschrift für Urheber- und Medienrecht 793, 797.

11 'Explanatory Memorandum of Section 44b UrhG' 89 <https://dserver.bundestag.de/btd/19/274/1927426.pdf> accessed 17.01.2024.

only be assessed when initiating new reproductions.

II. Machine-readable opt-out, i.e. machine-interpretable opt-out

14 If the content has been made publicly available *online*, the law sets out more specific requirements. In such instances, the copyright holder must express their opt-out through "machine-readable means" (Article 4[3] Directive [EU] 2019/790) or in a "machine-readable format" (Section 44b[3] UrhG), which essentially convey the same intent. If these conditions are not met, the opt-out is deemed ineffective.

15 Through the stipulation of machine-readability, the legislator clarifies the appropriate manner for conveying opt-outs in an online context. This distribution of risk between the copyright holder and the TDM user leads to the responsibility of the latter to identify potential opt-outs. If they are discovered to lack machine-readability, it disadvantages the copyright holder, rendering the opt-out ineffective.

16 Notably, the law does not furnish a precise definition of "machine-readable". Recital 18 of the directive (EU) 2019/790 only states "... it should only be considered appropriate to reserve those rights by the use of machine-readable means, *including* metadata and terms of a website or a service." This raises a seeming contradiction, as the machine-readability of a website's terms of use is uncertain. The explanatory memorandum to the German law tries to resolve this contradiction, as it mentions the terms of use of a website only in the context of *where* to express the opt-out, but not *how* to express it: "It can also be included in the imprint or in the terms [...], *provided that* it is machine-readable there, too."[12] Although still quite unclear, this explanation at least prioritizes machine-readability. If the copyright holder expresses the opt-out in the terms of use of the website, it is effective – if it is also machine-readable.

17 In the absence of a precise definition, the term "machine-readable" is to be understood functionally. As German explanatory memorandum to the law emphasises machine-readable means must be "suitable for automated processes of text and data mining [of online accessible sources]" because "[...] the purpose of the regulation is to ensure that automated processes, which are typical criteria of text and data mining, can actually be automated in the case of content accessible online".[13] Mere discoverability and automatic legibility of the opt-out are insufficient. Machines must also be capable of interpreting the opt-out in alignment with this perspective, rendering "machine-readable" tantamount to "machine-interpretable."[14] Therefore, an opt-out in only plain text (see C) or as a pictogram[15] is most likely not legally effective.

III. Which side is responsible for proposing a machine-readable standard?

18 The distribution of responsibility within the TDM exception presumes the feasibility for rights holders to reasonably express opt-outs. Within the online environment, it specifically assumes the availability of machine-readable formats accessible to TDM users. However, the opt-out approach encounters limitations when this assumption doesn't hold. The TDM exception does not address the question of who assumes the risk in situations where established standards are absent.

19 The opt-out approach of the TDM exception is very similar to the case law on thumbnails.[16] In this context, too, an opt-out solution was established: individuals who provide text or image content freely on the internet without technically feasible restrictions should anticipate customary usage

12 ibid.: „auch im Impressum oder in den [AGB] [...], sofern er auch dort maschinenlesbar ist."

13 ibid.: „in einer Weise erfolgen, die den automatisierten Abläufen beim Text und Data Mining angemessen ist"; „[...] bezweckt die Regelung, bei online zugänglichen Inhalten sicherzustellen, dass automatisierte Abläufe, die typisches Kriterium des Text und Data Mining sind, tatsächlich auch automatisiert durchgeführt werden können."

14 Winfried Bullinger in: Artur-Axel Wandtke and Winfried Bullinger (eds), *Praxiskommentar Urheberrecht* (6th edn, CH Beck 2022) § 44b UrhG no. 10. („detected and analyzed", german:„erkannt und ausgewertet"); Raue (n 10) 797; Marco Müller-ter Jung and Lewin Rexin, 'Rechtliche Anforderungen an Intelligentes Und Automatisiertes Technologiescouting Technische Umsetzung Unter Beachtung Urheberrechtlicher Und Datenschutzrechtlicher Hürden' Computer und Recht 174 (both only mention: "detected", german: "erkannt").

15 Björn Steinrötter and Lina Marie Schauer however consider plain text and pictograms (also) as maschine-readable, in: Marek Barudi (ed), *Das Neue Urheberrecht* (1st edn, Nomos 2021) § 44b UrhG no 14.

16 BGH 29.04.2010 - I ZR 69/08, openJur 2010, 528 (thumbnails I); 19.10.2011 - I ZR 140/10, openJur 2012, 659 (thumbnails II).

under prevailing circumstances. This may lead a search engine to interpret that the rights holder has consented to the use involving works' reproductions in thumbnails.[17]

20 The ('missing') opt-out within the TDM exception can be treated as an expression of consent. It adheres to the principles of the declaration of intent and is primarily interpreted from the recipient's standpoint. In situations of uncertainty, this recipient is the objective third party, i.e., a person possessing the knowledge expected in the relevant context. Their understanding is significantly shaped by customary practices in comparable scenarios. When customs are absent, no consent is inferred. The fundamental premise remains that TDM infringes copyright and thus constitutes an unlawful act. In situations lacking established norms, the TDM user faces a disadvantage. Nonetheless, the rightsholder must still communicate the opt-out, Article 4[3] Directive [EU] 2019/790 is clear on that. In such instances, any reasonable form of opt-out would be effective. The TDM user cannot contest that the rights holder did not use a machine-readable format.

21 This outcome aligns with the underlying notion in Section 31 (5) UrhG, which stipulates that, in cases of uncertainty, usage rights are granted only if they are essential to fulfilling contractual obligations. In scenarios where usage rights are likely to remain with the author[18], it is coherent that rights holders also tend to withhold consent.

22 Finally, the question arises at what level of dissemination machine-readable formats are to be considered common, prompting rights holders to use them to ensure an effective opt-out through compliance with a commonly accepted machine-interpretable format (see. B.II). A standardization body governing web crawling does not exist. The current "standards" (see D) have been shaped over time by Google's influence within the online sphere. The legislator's intention was to not wait for this normative influence to manifest, as this would render the TDM exception regulation redundant. Thus, it must suffice that practical machine-readable formats are extensively discussed, even if they have not yet become firmly established in practice.

C. Empirical analysis on possibilities to declare opt-outs

23 To comprehensively evaluate feasible procedures with reasonable effort for both copyright holders and TDM users to declare or to search for opt-outs, we perform an empirical test to analyze viable methods based on a current sample of websites. Our aim is to understand the practicability and exertion involved in identifying potential opt-outs. Ideally, copyright holders utilize established syntax, an aspect elaborated in the subsequent section addressing search engine standards (Section D). This segment focuses on the detectability of different standards and the evaluation of the machine-readability of the website's terms of use, aligning with Recital 18 of Directive (EU) 2019/790.

24 The process of searching for an opt-out within the terms of use page involves several steps:

- Identifying the webpage displaying the website's terms of use.
- Locating the pertinent section within the webpage.
- Interpreting the identified section as an opt-out for TDM

25 We conducted an analysis on a sample of 100 websites using a subset extracted from the latest crawl of the Common Crawl[19] archive (May/June 2023). The dataset holds petabytes of data accumulated since 2008, encompassing raw website data, extracted metadata, and textual content. Given the expansive nature of the Common Crawl dataset, which negates the requirement for individualized web crawlers, it emerges as a popular resource in TDM studies, offering both efficiency and comprehensive coverage. Prominent models such as OpenAI's GPT-3 have benefited from Common Crawl during their training, underscoring the dataset's significance in advancing the state-of-the-art in natural language processing.

26 Our sample comprised European domains that feature English versions, chosen to reflect the overall distribution of European web top-level domains within the Common Crawl dataset. For instance, out of the 100 websites sampled, 20 were German, 9 French, and 1 Portuguese, among others. However, of this number, only 85 were found to be valid for our study. The exclusions were due to different reasons: first and foremost, due to the fact that they had become inaccessible at the time of our analysis. For the valid sites, we proceeded with the steps as described.

17 BGH thumbnails I, no. 35 et seq.

18 Gernot Schulze, § 31 in Dreier and Schulze (n 8) n 110.

19 The Common Crawl Foundation, a Californian non-profit organization, was established with the mission of democratizing access to web information. Their vision embodies an "open" web that facilitates free access to information, laying the groundwork for innovation in research, business, and education.

27 1.) The terms of use page can be systematically identified by detecting distinct patterns or elements within the HTML code or URL. Standard HTML text elements within the terms of use segment, containing phrases such as "terms and conditions," "terms of use," or "terms of service," can serve as keywords for identifying matching content within HTML. This process, however, has inherent limitations, as some websites may utilize unconventional terminology or phrasing, leading to challenges in precisely identifying the desired section. For 12 websites, the page containing the terms of use was automatically identified utilizing the described methodology. However, limitations arose from the potential omission of specific pages of a website in individual monthly crawls by Common Crawl. Through manual investigation, a section containing the terms of use could be identified for 40 websites. This indicates that only about half of all websites are likely to contain such a section.

28 2.) To identify the pertinent section within the terms of use webpage, a keyword-based approach similar to step 1 was followed. Identifying particular sections of interest within a webpage, such as opt-outs in terms of use, poses considerable challenges when relying solely on keyword matching. This is because the phrasing and structure of such content can vary widely across websites, with synonyms, domain-specific terminologies, and varying language nuances.[20] For all the 12 websites where the terms of use were automatically detected in step 1, basic keyword searches either missed sections of interest or mistakenly highlighted unrelated content, showing that individual text content within the terms of use is a challenge for automated identification and interpretation of opt-outs. Therefore, to achieve a high degree of accuracy in this endeavor, specialized language models, which are trained to understand the text details and nuances of such documents, are required for both the identification and interpretation of relevant TDM sections.

29 The analysis highlights the difficulties of relying on specific subpages or sections, such as the terms of use, to communicate opt-outs. When accounting for this information across multiple domains and webpages, TDM users face significant challenges, as it necessitates the use of advanced, individually designed crawlers or a method to deal with possibly incomplete webpage coverage, when relying on pre-crawled websites. To ensure full webpage inclusion, e.g., through monthly crawls of Common Crawl, each crawl would need to be inspected, which raises feasibility concerns given the substantial storage and computational demands this approach entails. Moreover, automated interpretation of unique phrasings within the terms of use usually requires sophisticated language models. The vast diversity across websites complicates the automatic extraction of statements pertaining to usage restrictions.

30 The study shows that each of the three steps involves substantial effort and costs. Effective opt-out management would require advanced NLP methods, which might still carry high error rates. This could undermine the TDM exception's effectiveness. Opting out in a website's terms of use would not be appropriate to the automated processes of TDM. Therefore, it cannot be considered a legally effective opt-out by machine-readable means.[21] Consequently, Section D discusses the use of alternative standards established for analogous purposes such as search engines. Foremost among these is the robots.txt protocol (see D.I), a widely recognized standard used by websites to communicate with web crawlers and other automated agents. From our analysis, it is evident that this standard has gained widespread adoption: 75 of the 85 valid websites we inspected had a populated robots.txt file in their root directory. Given its prevalent use and standardized nature, there is a promising potential to further harness the robots.txt standard in streamlining the processes we are addressing.

D. TDM and search engines

31 The current challenge lies in the absence of a dedicated technological standard specifically tailored for TDM to meet the aforementioned legal requisites. The exemplary investigation has revealed the lack of a defined technological standard exclusively addressing legal demands for TDM. Consequently, consideration of alternative standards established for analogous purposes becomes pertinent, potentially offering utility or even sufficiency for TDM. In this context, the established standards utilized by search engine crawlers such as Google, Microsoft Bing, Yandex, among others, stand out. These standards encompass the definition of website authorship or ownership preferences for permitting or prohibiting website crawling and indexing – namely, robots.txt and meta-tags. Hence, adaption, expansion, and alignment of pre-existing standards with appropriate distribution should be considered for TDM, tailored

20 Müller-ter Jung and Rexin (n 14) no 30.

21 Tina Gausling, 'Wie Unternehmen Online Verfügbare Daten Nutzen Können' [2021] Computer und Recht 609, 611.; Bullinger no. 10 (n 14), Raue (n 10) 797 and Müller-ter Jung and Rexin (n 14) 174 emphasize that the crawler's algorithms must be able to recognize the opt-out automatically, but are not so consistent as to exclude the possibility of opting out in the terms of use for this reason. Steinrötter and Schauer (n 15) consider plain text to be adequate and therefore opting out in the terms of use to be legally effective.

to effectively fulfil its requirements. The following sections will delve into the mentioned standards and discuss how they can be expanded and utilized for declaring usage reservations for TDM.

I. Robots.txt

32 "If you don't want crawlers to access sections of your site, you can create a robots.txt file with appropriate rules. A robots.txt file is a simple text file containing rules about which crawlers may access which parts of a site."[22] The robots.txt standard defined by the Robots Exclusion Protocol[23] (RFC9309) serves as a widely adopted means of communicating instructions to web crawlers and other automated agents. Prominent search engine providers, like those mentioned above, as well as OpenAI and its ChatGPT plugin agents, designed to respond to real-time user queries, commit to following the relevant instructions provided by website owners.[24]

33 Setting up a robots.txt file in the root directory of a domain allows a website owner to define if certain URLs, directories, file patterns or even the entire website should not be indexed. Furthermore, it enables them to specify if certain crawlers, identified by their user agent name, are allowed or disallowed.[25]

34 Crawlers interpret the absence of a robots.txt file as a generally granted invitation to index the publicly contents of a website. Thus, setting up a robots.txt file can express an opt-out.

Example robots.txt:

```
user-agent: *
disallow: /

user-agent: googlebot-news
allow: /news
```

35 In the example, the initial rule prohibits all user agents from indexing by utilizing the asterisk (*) as a wildcard character in the user-agent field. The asterisk functions as a universal placeholder for all user-agents, signifying that the instructions pertain to all web crawlers and automated agents accessing the website. The subsequent rule permits the Google "googlebot-news" crawler to index the news directory.

36 The robots.txt standard possesses the capacity to both allow and disallow specific or all user agents for certain or all URLs of a website. However, it does not offer the capability to grant or deny access for specific purposes like TDM. The limitation of purpose can solely be achieved by excluding particular user agents. The proprietary scripts and software employed for extracting information from websites for TDM typically lack identifiable user agent names, making them ineligible for disallowance. The demonstrated method above, involving disallowing all agents except those recognized as not engaging in TDM-related crawling, would be the only viable approach using the robots.txt to express an opt-out for TDM without requiring an extension of the standard.

II. Meta tags

37 Another type of annotation used by search engine crawlers are the so-called meta tags. Meta tags are invisible HTML tags integrated in the head part of a HTML document defining a website. They contain meta data offering further information about the website they are integrated in. Meta tags are on a per page basis. Therefore, they can be different for every single page of a website.

38 While meta tags can contain different types of data for different purposes, there are special meta tags for indexing:

22 Google Search Central, 'How Google interprets the robots.txt specification' <https://developers.google.com/search/docs/crawling-indexing/robots/robots_txt> accessed 17.01.2024.

23 M.Koster, G. Illyes, H. Zeller and L. Sassman, 'RFC 9309 Robots Exclusion Protocol' <https://www.rfc-editor.org/rfc/rfc9309.html> accessed 17.01.2024.

24 OpenAI, 'ChatGPT-User' <https://platform.openai.com/docs/plugins/bot> accessed 17.01.2024.

25 Google Search Central, 'How Google interprets the robots.txt specification', <https://developers.google.com/search/docs/crawling-indexing/robots/robots_txt> accessed 17.01.2024.

Example meta tags:
<meta name="robots" content="noindex">
<meta name="googlebot-news"
content="index">

39 The example disallows all robots from indexing the current page, first. Then it allows the user agent called "googlebot-news", and only this user agent, indexing of the page. This way indexing can be allowed and disallowed on any page. By default, if there are no meta tags disallowing it, indexing is allowed. Again, an opt-out is necessary to avoid indexing.[26]

40 Like robots.txt, there is no option in meta tags to only allow crawlers for specific purposes.

III. TDM as part of the search engine standard

41 Considering the resemblance between TDM and search engine operations, adopting the existing tools utilized for search engines appears rational for TDM as well. Both the robots.txt and meta tags could serve as suitable machine-readable methods to accurately convey opt-outs for TDM.[27] Conversely, a pertinent question emerges: What is the implication if a website lacks a robots.txt or meta tag conforming to the outlined scheme? Can the user then assume that the rightholder has not expressed an effective opt-out?

42 For the general TDM exception, the term TDM – i.e., the applications that are covered by it – is deliberately defined broadly:

43 Art. 2(3) Directive (EU) 2019/790: "any automated analytical technique aimed at analysing text and data in digital form in order to *generate* information which includes but is not limited to patterns, trends and correlations."

44 Section 44b(2) UrhG: "the automated analysis of individual or several digital or digitised works for the purpose of *gathering* information, in particular regarding patterns, trends and correlations".

45 In the explanatory memorandum of Section 44b UrhG, a distinction between Text and Data Mining and search engines is presented: "An opt-out [...] for a website must not lead to it being treated unequally in the context of other uses without objective justification, for example when it is displayed as a search engine hit. This is because the reservation of use should not affect other use cases.[28] However, this statement can also be interpreted more as fairness requirements applied to search engines, particularly due to their significant market influence, rather than as a definitive differentiation from TDM.

46 From a technical standpoint, search engines can be regarded as an integral component of TDM. They autonomously analyze and crawl substantial volumes of text data accessible on the internet, subsequently indexing it and utilizing algorithms to retrieve pertinent information in response to user queries. Furthermore, they employ advanced natural language processing and information retrieval algorithms to comprehend the semantic context of user queries, categorize data into applicable classifications (such as news, images, or books), and extract salient topics from texts.

47 Considering these aspects, robots.txt and meta tags can indeed be utilized for opt-out purposes, albeit with certain constraints. As demonstrated earlier, disallowing all agents except those recognized as permissible could effectively serve as an opt-out method. A TDM user who has been explicitly granted permission could rely on this arrangement.

48 However, the converse approach does not yield the same results. The rightsholder cannot be directed to employ robots.txt or meta tags in the demonstrated manner. The limitations imposed by these standards present significant challenges. The rightsholder would be obligated to individually list all authorized user agents and maintain the list's accuracy over time. Failure to do so jeopardizes the visibility of the website in prominent search results. It's worth noting that the prohibition of devaluation by search engines due to a TDM opt-out is of limited value, as search engines periodically introduce new user agents that would need to be disregarded by the website. Consequently, a TDM user who is not disallowed through robots.txt or meta tags cannot reasonably argue the absence of an effective opt-out.

26 Google Search Central, 'Robots meta tag, data-nosnippet, and X-Robots-Tag specifications' <https://developers.google.com/search/docs/crawling-indexing/robots-meta-tag> accessed 17.01.2024.

27 Gausling (n 21) 611; Björn Steinrötter & Lina Marie Schauer, § 4, in; Barudi (n 15) no 14; Müller-ter Jung and Rexin (n 14) 174.

28 'Explanatory Memorandum of Section 44b UrhG' (n 11) 89.:„Ein Nutzungsvorbehalt nach § 44b Absatz 3 UrhG-E für eine Webseite darf nicht dazu führen, dass diese im Rahmen anderer Nutzungen ohne sachliche Rechtfertigung ungleich behandelt wird, beispielsweise bei der Anzeige als Suchmaschinentreffer. Denn der Nutzungsvorbehalt sollte andere Nutzungen nicht betreffen."

E. An own machine-readable standard for TDM

49 In order to establish a standardized framework facilitating the systematic formulation of usage restrictions pertaining to TDM of websites, various methods are under consideration. Through the suggested approaches, varying levels of precision in content exclusion can be accommodated. This permits meticulous regulation of website usage for TDM, ensuring both systematic and efficient incorporation of opt-out mechanisms.

I. Standards in development

50 Presently, multiple organizations are engaged in the development of potential standards concerning the allowance and disallowance of TDM. However, as of now, none of these standards has become established and found widespread acceptance.

51 W3C proposes the TDM Reservation Protocol (TDMRep) which foresees meta tags or a JSON-LD integration of the permissions in the page's code. This way it is possible to allow or disallow certain pages. The directive "tdm-reservation" accepts either 0 (=opt-out) or 1 (=opt-int) to specify if TDM rights are reserved or not reserved. The second directive, "tdm-policy," enables the specification of a URL where additional policy-related information can be accessed. It's important to note that if the information at this URL is solely available in HTML or text formats, it is not considered machine-readable. To achieve machine-readability, policies must be articulated using JSON or JSON-LD, with W3C delineating their structure and admissible values.

52 Analogous to the robots.txt mechanism, TDM reservations may be defined in a file named tdmrep.json, which has to be placed in the domain's root folder. In this scenario, an additional directive is mandated to specify the paths to which the reservations apply.[29]

53 IPTC's RightsML standard, which was published in 2018 and is based on W3C's Open Digital Rights Language (ODRL), offers defining extensive machine-readable usage policies for any type of media. It's available as XML, RDF and JSON-LD. This standard was initially intended to facilitate the communication of intellectual property rights and usage permissions associated with media assets. Over time, RightsML has found application in conveying licensing information, copyright terms, and usage restrictions for digital content across diverse sectors. This existing standard and infrastructure could be extended to encompass opt-outs for TDM by incorporating attributes that explicitly denote TDM permissions and restrictions. By integrating TDM-specific information into the RightsML schema, a comprehensive and structured approach can be achieved for addressing opt-outs related to TDM activities.[30] Thus, RightsML has been proposed as a possible solution at the W3C Text and Data Mining Reservation Protocol Community Group.[31]

54 The Coalition for Content Provenance and Authenticity (C2PA) developed another approach. They also introduced a rights protocol that can be attached as metadata directly to content. Opt-out reservations are delineated through specific data mining entries, allowing differentiation between various forms of utilization. At the cost of being operationally more complex, it offers the advantage of cryptographic traceability for content modifications. This standard therefore goes beyond the opt-out declaration towards the enforcement through content provenance.

55 Observing the evolving landscape of AI and research applications, Google recognised the need for updated web publisher controls that accommodate these new use cases. They initiated a public discourse inviting stakeholders from the web and AI communities, including publishers, civil society, and academia, to contribute to the development of complementary protocols with robots.txt as a starting point that enhance web publisher choice and control for emerging TDM applications[32]. The discussion is still underway without preliminary results or draft implementation proposals being public yet.

II. REP – Proposal for the implementation

29 World Wide Web Consortium, 'TDM Reservation Protocol (TDMRep)' <https://www.w3.org/2022/tdmrep/> accessed 17.01.2024.

30 International Press Telecommunications Council, 'IPTC RightsML Standard 2.0' <https://iptc.org/std/RightsML/2.0/RightsML_2.0-specification.html> accessed 17.01.2024.

31 International Press Telecommunications Council, 'IPTC's RightsML at W3C Text and Data Mining Reservation Protocol CG' <https://www.iptc.org/news/iptc-rightsml-at-w3c-text-and-data-mining-reservation-protocol-wg/> accessed 17.01.2024.

32 Danielle Romain, 'A principled approach to evolving choice and control for web content' <https://blog.google/technology/ai/ai-web-publisher-controls-sign-up/> accessed 17.01.2024.

56 As described in the preceding section, a pragmatic approach involves leveraging well-established conventions of the robots.txt file (see Section D.I), which allows website owners to establish wide-ranging exceptions at both the directory and page levels. In the same way that robots.txt can be used to declare access and usage restrictions for web crawlers, it could also be extended to declare usage restrictions for AI model training and other TDM activities. This extension would involve augmenting the robots exclusion protocol (REP) to incorporate information about the approval or disapproval of content specifically for TDM purposes.

57 Technically, this proposed extension can be actualized through the introduction of an optional term "purpose" within the robots.txt file. This extension empowers website owners to precisely define access permissions and restrictions tailored to specific purposes. The "purpose" term accommodates the assignment of various values, including "search-engine," "tdm," and "other," thereby affording a finer-grained control over user-agent behaviour. To ensure comprehensive coverage, it should be mandatory for each user-agent to be assigned to at least one of the purpose groups. This condition guarantees the explicit coverage of all user-agents in terms of their intended applications.

58 In instances where the purpose term is absent from a rule specified in the robots.txt file, it defaults to encompassing all feasible values. This default behaviour contributes to inclusivity and mitigates inadvertent access or restrictions.

59 By adopting this approach, the example presented in Section D.I can be expanded to demonstrate the extension of the robots.txt file, incorporating TDM-specific usage restrictions:

Example robots.txt:

```
user-agent: *
disallow: /
purpose: tdm

user-agent: *
allow: /news
purpose: indexing
```

60 In this illustrative scenario, access for all users is denied for any TDM activities throughout the entire website. However, an exception is made for web crawlers designed to index the designated news directory. This strategic decision reflects a common consideration among website owners who aim to safeguard their content from automated gathering for NLP model training while simultaneously striving to enhance visibility in popular search engine query results, thereby increasing click rates.

61 By means of this proposed standard, website owners can effectively communicate their requirements regarding TDM and specify compliance with usage regulations for purposes beyond research.

62 The aforementioned approaches, namely the use of the robots.txt file and of HTML meta tags, focus on providing page-wide and per page opt-outs for TDM. However, in the pursuit of a comprehensive standard that allows for precise control of TDM access, we posit the necessity to introduce methods that permit the exclusion of specific sub-areas within individual web pages. There could also be instances where the need or obligation arises to permit or restrict TDM exclusively for specific segments of a single webpage, without these rules being applicable to other sections.

63 Structured data represents a potential solution that is currently employed by search engines. It encompasses machine-readable concealed supplementary content that is directly integrated into a website. It describes an element or a group of elements in a standardized form that can be interpreted by machines trained for it. While the suggested structured data format, JSON-LD, may not ideally cater to this objective due to its page-level embedding of structured data, alternatives in the form of microdata and RDFa standards are available. These extend the regular HTML code of a website creating new elements or assigning additional data to already existing elements.[33]

64 Structured data relies on standardized elements that are defined and described by schema.org. This poses challenges when attempting to introduce new elements for the purpose of excluding TDM from specific portions of a website.

65 An alternative and more straightforward approach involves the incorporation of a novel HTML data attribute, similar to the practice adopted by Google to exclude sections of a website from its featured snippets. Featured snippets are distinctive boxes that invert the format of a conventional search result, displaying the descriptive snippet before other content. These may also appear within a grouping of related questions, known as "People Also Ask."[34]

66 Referred to as "data-nosnippet," this attribute is to be applied to the HTML element whose content should

33 Google Search Central, 'Introduction to structured data markup in Google Search' <https://developers.google.com/search/docs/appearance/structured-data/intro-structured-data> accessed 17.01.2024.

34 Google Search Central, 'Featured snippets and your website' <https://developers.google.com/search/docs/appearance/featured-snippets#block-fs> accessed 17.01.2024.

not be displayed within the featured snippets.[35]

Example:

```
<body>
 <h1>Le Louvre</h1>
 <p>The Louvre in Paris is the national museum of France. It is situated within the 1st arrondissement. ...</p>
 <div data-nosnippet>
 <h2>West wing closed</h2>
  <p>Because of the ongoing renovations in the west wing, this part of the Louvre is not accessible until the end of the year. We are sorry for the inconvenience.</p>
 </div>
</body>
```

67 The provided example shows general information about the Louvre Museum in Paris. While the first part, the description, can be used by Google for the featured snippets, the second part, the information about the renovations, is excluded from being used for the snippets.

68 Introducing a new attribute "data-notdm" would enable webpage owners and content creators to exclude specific parts of a page from being used for TDM purposes. Crawlers would then have to look out for these annotations within the code and either include or exclude the corresponding HTML tag's contents when extracting information.

III. Machine-readability of the proposed standards

69 The proposed and described "data-notdm" HTML attribute, as well as, the REP approach, can be understood as machine-readable method to articulate an opt-out for TDM activities. The criteria for machine-readability (section B.II) emphasize functionality tailored to automated processes of TDM from online accessible sources. The fundamental requirement is that machines possess the capacity to comprehend and interpret the opt-out in accordance with the specified context. By adhering to this notion, both the "data-notdm" attribute and the REP framework satisfy the core prerequisites for machine-readability, as they enable precise, automated, and contextually aligned communication of TDM exclusion preferences.

F. Outlook

70 Ensuring compliance with copyright laws and adhering to legal boundaries for TDM is of paramount importance to provide website owners and content creators with continued confidence in content provisioning on websites. To achieve this, the development and widespread adoption of legally sound standards and their enforcement are desirable to maintain operational security. Current discussions about AI regulation and the European AI Act[36], primarily emphasize ensuring transparency in training data to proof that it is relevant, representative, complete and error-free. The AI Act rather establishes a form of product conformity framework for AI products to ensure their content quality depending on their risk class, rather than being a primary means of demonstrating adherence to copyright for training data usage. In the latest consolidated draft for the AI Act, however, the legislator also seeks to address this, by including an amendment that requires providers of generative AI systems to "document and make publicly available a sufficiently detailed summary of the use of training data protected under copyright law".[37]

71 Given the absence of established standards (in the meaning of customary practices) and in view of the dynamic discussion about new standards (see B.III and E), website owners are advised to adopt a pragmatic approach when expressing a machine-readable opt-out. Incorporating the proposed statement in the robots.txt – using the already accepted way to allow or disallow user-agents (see D.I) – and simultaneously integrating it with an HTML attribute within relevant webpage elements according to the now established methods, can serve as an interim solution. These can then easily be adapted for a purpose statement or changed to a "data-notdm" attribute when this way of declaring an opt-out becomes an accepted practice.

35 Google Search Central, 'Featured snippets and your website' <https://developers.google.com/search/docs/appearance/featured-snippets#block-fs> accessed 17.01.2024.

36 Tambiama Madiega, 'Artificial intelligence act' <https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-regulation-on-artificial-intelligence> accessed 17.01.2024.

37 Amendment 399 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, Article 28 b, <https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/CJ40/DV/2023/05-11/ConsolidatedCA_IMCOLIBE_AI_ACT_EN.pdf> accessed 17.01.2024.