Artificial Intelligence and the Law of Machine-Readability

A Review of Human-to-Machine Communication Protocols and their (In)Compatibility with Article 4(3) of the Copyright DSM Directive

by **Hanjo Hamann** *

Abstract: Many legal scholars critique the supposed ineffectiveness of European copyright regulation regarding commercial text and data mining. At the same time, tech-savvy entrepreneurs keep proposing new standards to effectuate them at a rate that has been described as "exponential". The present paper reconciles these complementary perspectives. In the first (doctrinal) part, it develops a framework for article 4(3) of the Copyright DSM Directive by arguing that: (1) Web-scraping for AI training is a use case of TDM. (2) European TDM regulation seeks to protect fundamental rights and to uphold incentives of both AI developers and rightholders. (3) To ensure balanced protection, the legislator provided for a "reservation of rights" as an exception similar to one found in the Berne Convention. (4) This reservation instrument gets criticized on account of being either unduly effective or largely ineffective – a tie that can only be broken by clarifying the doctrinal hurdles raised by the Directive. (5) The Directive establishes two standards that reservations need to fulfil simultaneously: They must be explicit (specific for a given content and use) and automatable (employing a well-defined technical protocol). In the second half of the paper, it uses these standards to assess seven communication protocols commonly proposed to reserve TDM rights. It concludes that only some qualify as "machine-readable" in a legal sense at all, and that the proliferation of standards currently precludes any effective reservation of TDM rights. This may, however, come with a silver lining.

Keywords: AI web-scraping, text and data mining, machine-readability, rights reservation, H2M communication

Recommended citation: Hanjo Hamann, Artificial Intelligence and the Law of Machine-Readability: A Review of Human-to-Machine Communication Protocols and their (In)Compatibility with Article 4(3) of the Copyright DSM Directive, 15 (2024) JIPITEC 102 para 1.

[*] Prof. Dr. Dr., JSM (Stanford), assistant professor for civil law, commercial and intellectual property law, in particular the law of digitalisation and legal linguistics, at the Wiesbaden University of Business and Law (EBS Law School). The basic argument of this paper was presented at Humboldt University in Berlin (19 June 2023) and at a conference on "Generative AI Through the Lens of Copyright Law" that I co-hosted with Katharina de la Durantaye, Franz Hofmann and Benjamin Raue in Berlin (23 Feb 2024). Besides these colleagues, I thank Emre Bayamlıoğlu, Anna Bernzen, Péter Mezei, Alexander Peukert, Jonathan Pukas, Eleonora Rosati, Alain Strowel, and Maren Wöbbeking for feedback and helpful suggestions on various drafts for this paper, as well as Simon Weyhofen for editorial assistance.

A. The Little Spider who tried to Save the Web

1 The following story is based on true events.[1]

2 Once upon a time in Europe, there was a small computer program. It got sent on a mission to collect text so that its master's could train a Large Language Model. It was told to follow a simple protocol: Go to a website on the Internet, copy its contents into a database, then follow each hyperlink to other websites, and start over. Since the program "crawled" the web in this manner, some called it a "spider". (Others admired its robot-like discipline and called it a "bot".) The crawling spider did a good job, although its mission protocol was not as simple as it appeared at first:

3 Whenever the spider approached a website that it sought to enter, it had to identify itself to the virtual butler ("server") by telling him its name. For instance, our spider might have called itself "CCbot" or "GPTBot" or "anthropic-ai". One beautiful morning, the spider approached a server and (following good old robot-spider manners) started by asking for the rules of the house. The server responded that he knew them and was ready to hand them to the spider, which in machine language sounded like this:[2]

4
```
HTTP/2 200
server: myracloud
date: Mon, 04 Mar 2024 02:01:00 GMT
accept-ranges: bytes
tdm-policy: https://rsw.beck.de/beck-online-service/tdm-vorbehalt
tdm-reservation: 1
content-security-policy: […] etag: […] x-content-type-options: […] X-Firefox-Spdy: h2
```

5 Along with this response, the server delivered the requested list of rules as a text file ("robots.txt"), which our spider instantly read. It said:[3]

6
```
User-agent: CCBot
User-agent: GPTBot
User-agent: ChatGPT-User
Disallow: /
```

7 The spider already knew this text because two out of every five news portals worldwide (40.7 %) feature the same house rules.[4] This time, the file contained two additional lines of text,[5] but being prepended by hashtag characters (#), our spider knew they were meant to be read by humans and incomprehensible to machines.

8 Next, the little spider requested the landing page from the server. This would usually be called index.html or something to that effect; here, it was simply "/Home". The server knew what to deliver, and sent our spider a file that it devoured eagerly. Some eighty lines at the start of this file were written in machine language, opening with:[6]

[1] The following is adapted from a German long-form article from which this paper derives: *Hanjo Hamann*, 'Nutzungsvorbehalte für KI-Training in der Rechtsgeschäftslehre der Maschinenkommunikation' (2024) 16 ZGE/IPJ 113.

[2] HTTP Response Header of <beck-online.beck.de/robots.txt> (accessed 4 Mar 2024). See *infra*, section C.IV.

[3] File contents of <beck-online.beck.de/robots.txt> (accessed 4 Mar 2024) See *infra*, section C.II.

[4] Data and sources *infra* (n. 98).

[5] Literally: „# Legal notice: Verlag C.H.BECK oHG expressly reserves the right to use its content for commercial text and data mining (§ 44b Urheberrechtsgesetz). – # The use of robots or other automated means to access our websites or collect or mine data without the express permission of Verlag C.H.BECK oHG is strictly prohibited."

[6] File contents of <beck-online.beck.de/Home> (accessed 4 Mar 2024). See *infra*, section C.V.

9
```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01
Transitional//EN" "http://www.w3.org/TR/html4/loose.
dtd">
<html lang="de" class="">
<head>[…]
   <title>Homepage - beck-online</title>
   <meta name="format-detection" content="telephone=no"
/> […]
   <meta http-equiv="content-type" content="text/
html;charset=utf-8" />
   <meta http-equiv="Content-Style-Type" content="text/
css" />
   <meta name="tdm-reservation" content="1">
   <meta name="tdm-policy" content="https://rsw.beck.de/
beck-online-service/tdm-vorbehalt">
   <meta name="robots" content="noai, noimageai">
```

10 Our spider copied the contents of this file into her database and proceeded to follow each of the file's hyperlinks. One of them was labelled "AGB" and pointed to a file on a different subdomain. The spider requested to read it. This file, too, began in machine language, but continued as a garbled mix of human- and machine-readable text. For instance, the spider found this string of characters:[7]

11
```
<div >[…]<h4>9. Schutzrechte</h4><p>[…]<br /></font>9.2
Der Verlag beh&auml;lt sich gem&auml;&szlig; &sect; 44b
Abs. 3 UrhG&nbsp;das Recht vor, Vervielf&auml;ltigungen
[…] zum Zwecke des Text und Data Mining vorzunehmen.<br
/><br />
```

12 The spider could not make sense of this, as it did not speak human language, let alone German.[8] All it could do was to use the interspersed bits of machine language to display a well-formatted text for humans to read. But there was no human wanting to read it, so the spider, following its protocol, saved the file's contents, and proceeded to visit the next hyperlink. This one was labelled "These General Terms and Conditions in English (PDF)", and it pointed to a binary-encoded file[9] rather than plain text that the spider might have saved. Another file that the spider did save that day (called "/Impressum") contained a string of characters not unlike the one cited above:[10]

13
```
<p><b>[…]&nbsp;Text and Data Mining according to &sect;
44b UrhG<br></b>[…]<br>The publisher reserves the right
to reproduce for text and data mining according to
&sect; 44b UrhG.</p>
```

14 Little did the spider know that this was in a different language than the one in the previous quote – it was still human language. The most the spider could have determined, based on a statistical comparison of both strings and their overlapping use of bigrams like "data mining" and "44b UrhG", was that both files were surely dealing with similar issues. But no one had told (or taught) the spider to do this, so it continued to visit the next batch of hyperlinks. Most of them pointed at files of about 10 kilobytes in size, which for a human would have looked something like this:

> **15** "You can access the requested file only if you are logged in. If you do not have personal login data, you can subscribe to one of the database modules mentioned above."[11]

16 Our spider diligently saved each of these error messages, and continued to visit many other websites that day. All of them were saved in the same manner: File by file, link by link. Soon the spider had gathered billions of texts in its database. And since robot-spiders never die, it continued to crawl and save the web happily ever after.

17 What is the moral of our story? Did the spying spider violate European copyright law?

B. Copyright Reservations against AI Web-Scraping

18 Legal debate about artificial intelligence is ubiquitous. So, too, in copyright law. Yet, although much has been written and discussed about protecting the output of AI (i.e., the "downstream" of digital value-creation), this paper is concerned with its inputs, i.e., "the upstream side, which might be slightly less aesthetic, but from a practical point of view [.] far more pressing. Surprisingly, to date these questions have attracted little academic attention."[12]

[7] Quote from <rsw.beck.de/beck-online-service/agb-beck-online> (accessed 4 Mar 2024). See *infra*, section C.I.

[8] Or else it would have read, "9. Protected rights, 9.2 The publisher reserves the right under Sec. 44b(3) German Copyright Code to reproduce contents for purposes of text and data mining."

[9] Namely <rsw.beck.de/docs/librariesprovider138/kam-support-dokumente/general_terms_and_conditions_beck_online_2023_08_23.pdf> (accessed 4 Mar 2024). See *infra*, section C.I.

[10] Quote from <beck-online.beck.de/Impressum> (accessed 4 Mar 2024). See *infra*, section C.I.

[11] Quote translated from German ("Sie können das gewünschte Dokument […] nur aufrufen, wenn Sie eingeloggt sind. […] Besitzen Sie kein persönliches Login […], dann können Sie eines der oben genannten Module abonnieren") taken from <beck-online.beck.de/vpath=bibdata%2Fkomm%2FDieNotKosBer%2Ehtm> (accessed 4 Mar 2024). See *infra*, section C.VII.

[12] *Daniel Schönberger*, 'Deep Copyright: Up- and Downstream Questions Related to Artificial Intelligence (AI) and Machine Learning (ML)' (2018) 10 ZGE/IPJ 35, 47.

19 The view that questions of input regulation appear "less aesthetic" seems to result, at least in part, from their technicality. As we will see throughout this paper, effective regulation of AI inputs requires diving deep into technical specifications. This lies beyond the comfort zone of most lawyers. What this paper will also show, however, is that lawyers need to get comfortable interpreting technical standards just as they have been interpreting legal jargon. Otherwise, any attempt at governing the digital realm by way of half-understood terms of art (such as "machine-readability") will merely turn the law into a dysfunctional barrier against innovation. Before we turn to such technical aspects, let us first consider the currently applicable laws and their doctrinal structure.

I. AI Web-Scraping as a Use Case of Text and Data Mining (TDM)

20 In order to train algorithms such as large language models ("LLMs"), AI developers require large amounts of textual data. In obtaining such training data, they commonly send spiders to scrape the web and download available online contents. Each download involves copying a file, which infringes upon rightsholders' reproduction right under Article 2 of Directive 2001/29/EC on Copyright and Related Rights in the Information Society ("InfoSocD"),[13] unless AI developers can invoke a copyright exception. Such an exception may be found in the Directive (EU) 2019/790 on Copyright in the Digital Single Market ("CDSMD"), which requires member states to introduce an exception for general-purpose *text and data mining* ("TDM"). This is defined as

> **21** "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" – article 2(2) CDSMD

22 In the past, there was considerable uncertainty whether web-scraping for AI training falls under the purview of this definition. Nowhere did the CDSMD refer specifically to artificial intelligence, so "there is no provision in the Directive that expressly deals with the training of AI",[14] which some say "has obviously been overlooked".[15] The Directive merely acknowledged vaguely that "text and data mining technologies are prevalent across the digital economy" (recital 8 CDSMD),[16] and sought to "provide for more legal certainty in such cases and to encourage innovation also in the private sector" (recital 18 subpar. 1 CDSMD).

23 While AI certainly exemplifies innovation in the private sector, there are reasonable doubts whether today's transformer architectures – as black box processes that even AI developers cannot understand or explain intelligibly – are really *aimed at analysing in order to generate information* in the sense of article 2(2) CDSMD. Many authors find it "not without a degree of uncertainty",[17] or outright "unclear whether the exceptions also cover" reproductions "for the development, training, and testing of AI systems".[18] Such reasonable doubts notwithstanding, most copyright scholars agree that "classical TDM and machine learning [...] use the same key algorithms to discover patterns in data",[19] so that the TDM exception "could be invoked, *a priori*, within the framework of any ML project".[20] Some have even

[13] This is a simplification. There is more than meets the eye to the question whether AI trainers actually "use copyright protected subject matter" in a legal sense. See *Mezei*, 'A saviour or a dead end? Reservation of rights in the age of generative AI' (2024) 46 Eur. IP Rev. 461, 463.

[14] *Jan Bernd Nordemann, Jonathan Pukas*, 'Copyright exceptions for AI training data – will there be an international level playing field?' (2022) 17 J. of IP Law & Pract. 973, 974.

[15] *Christophe Geiger*, 'When the Robots (Try to) Take Over: Of Artificial Intelligence, Authors, Creativity and Copyright Protection' in Thouvenin/Peukert/Jaeger/Geiger (eds.), 'Kreation Innovation Märkte – Creation Innovation Markets: Festschrift Reto M. Hilty' (2024), 67, 77, reasoning that the TDM exception was "not designed to cover machine learning by generative AI systems".

[16] This seems to be what *Mezei* (n. 13), 465 at fn. 47 refers to as "developments of AI".

[17] *Nordemann/Pukas* (n. 14), 974; earlier doubts by *Schönberger* (n. 12), 56: "a relationship might be seen [...] although ML is much further down the line than TDM"; most recently, *Mezei* (n. 13), 465: "even if the TDM exceptions were designed in light of the developments of AI, they were not drafted in light of GenAI."

[18] *Peter Georg Picht, Florent Thouvenin*, 'AI and IP: Theory to Policy and Back Again – Policy and Research Recommendations at the Intersection of Artificial Intelligence and Intellectual Property' (2023) 54 IIC 916, 928; similarly undecided *Andres Guadamuz*, 'A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs' (2024) 73 GRUR Int. 111, 120: CDSMD exceptions "should work to allow some machine learning operations to take place legally, but there will be some room for interpretation depending on the particulars of each situation."

[19] *Eleonora Rosati*, 'Copyright in the Digital Single Market' (2021), 72, concluding that "TDM plays a significant role in the advancement of AI applications."; similarly, *Séverine Dusollier*, 'The 2019 Directive on Copyright in the digital single market: Some progress, a few bad choices, and an overall failed ambition' (2020) 57 CMLR 979, 984: "artificial intelligence, based on machine-learning, is also deeply reliant on data mining".

[20] *Theodoros Chiou*, 'Copyright lessons on Machine Learning: what impact on algorithmic art?' (2019) 10 JIPITEC 398, 409

criticized the TDM exception as being "overly broad" exactly because its definition was construed to encompass "a vast field that includes most forms of modern artificial intelligence applications".[21]

24 The final nail in the coffin[22] of this controversy came, arguably, with the Artificial Intelligence Act recently adopted as Legislative Resolution 2024/138 by the European Parliament ("AI Act"). Recital 105 of the AI Act clearly states that "text and data mining techniques may be used extensively" in the context of "large generative models" for the "retrieval and analysis of such content, which may be protected by copyright and related rights." While one might argue that recitals are not themselves legal acts but merely the "reasons on which they are based" in the sense of article 296(2) TFEU, the proper text of the AI Act also mentions data mining as one of the "procedures for data management [...] performed before and for the purpose of [...] high-risk AI systems" (article 17(1) f AI Act). This makes it abundantly clear that the European legislator has decided to apply the TDM exception in cases of reproduction for purposes of AI web-scraping.[23]

II. Rationales of the TDM Exception: Justifying An Exception-Exception

25 There are at least two rationales for the legislator to let AI developers invoke the TDM exception when reproducing works for inclusion in training datasets. Both conversely justify a critical carve-out to the exception.

26 One rationale is rights-based. Speaking in terms of Charter 2012/C 326/02 of Fundamental Rights of the European Union ("EUCFR"), the right of AI developers to mine text and data is protected by the more general freedoms of scientific research (article 13 EUCFR) and the freedom to conduct a business (article 16 EUCFR). Indirectly, it also protects downstream AI end users' freedom of expression and information (article 11 EUCFR) and freedom of the arts (again, article 13 EUCFR). Conversely, however, the right of AI developers to mine text and data encroaches upon authors' and creators' rights of expression and information (again, article 11 par. 1 EUCFR), and their right to intellectual property (article 17 par. 2 EUCFR). Given this head-on collision of fundamental rights, one objective of (copyright in general and particularly) the CDSM Directive is "to achieve a fair balance between the rights and interests of authors and other rightsholders, on the one hand, and of users on the other" (recital 6 CDSMD). To that end, article 7(2) CDSMD incorporates the three-step test from article 5(5) of the InfoSocD, based on article 9(2) of the Berne Convention.

27 The other rationale is incentive-based. The CDSMD in particular (and copyright in general) seeks to "stimulate innovation, creativity, investment and production of new content" (recital 2 CDSMD). While the TDM exception is meant "to encourage innovation also in the private sector" through incentivizing AI developers, it simultaneously needs to incentivize rightholders by enabling them to "license the uses of their works or other subject matter" (recital 18 CDSMD).

28 Both rationales interlock, and demand a counterbalance for the TDM exception in order to protect and incentivize rightsholders affected by it. This would usually take the form of monetary compensation.[24] The Directive does not prohibit this solution, but does not recommend it either.[25] Instead,

(marginal 22); *Jonathan Griffiths, Tatiana Synodinou, 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 Int. 22, 25 at fn. 42; *Martin Senftleben*, 'Generative AI and Author Remuneration' (2023) 54 IIC 1535, 1542 at fn. 33; *Juha Vesala*, 'Developing Artificial Intelligence-Based Content Creation: Are EU Copyright and Antitrust Law Fit for Purpose?' (2023) 54 IIC 351, 355; *Katharina de la Durantaye*, 'Garbage In, Garbage Out. Regulating Generative AI Through Copyright Law', translation of a German journal article (ZUM 2023, 645) available through SSRN as of 13 Oct 2023 <doi.org/10.2139/ssrn.4572952>.

[21] *Thomas Margoni, Martin Kretschmer*, 'A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology' (2022) 71 GRUR Int. 685, 686 – see also ibid. 688: "under the misleading label of TDM, what has been regulated at the EU level in Arts. 3 and 4 goes far beyond a mere copyright exception. In fact, it should be reclassified as [...] a property-right approach to the regulation of AI."

[22] *Alexander Peukert,* 'Copyright in the Artificial Intelligence Act – A Primer' (2024) 73 GRUR Int. 497, 503 after fn. 88.

[23] *Peukert* (n. 22) 503 at fn. 90: "EU legislator confirmed this prevailing view qua lex posterior"; on the other hand, see *Geiger* (n. 15), 77: "the discussion is not over"; *Guadamuz* (n. 18), 111: "growing debate".

[24] For example, see the proposal by *Geiger* (n. 15), 78–81.

[25] Recital 17 CDSMD justifies to "not provide for compensation for rightholders" only insofar as "potential harm created to rightholders through this exception would be minimal" because "of the nature and scope of the exception, which is limited to entities carrying out scientific research". This does not apply to commercial TDM, which is justified in Recital 18 CDSMD without reference to compensation at all.

the legislator designed an opt-out process (an *exception-exception* of sorts) whereby rightsholders can unilaterally declare a "reservation" to suspend the TDM exception in particular cases. This mechanism applies to any TDM use including the use for AI training, as the AI Act clarifies:

29 "rightsholders may choose to reserve their rights over their works or other subject matter to prevent text and data mining [...] providers of general-purpose AI models need to obtain an authorisation from rightsholders if they want to carry out text and data mining over such works." (recital 105 AI Act)

30 Despite what the first part of this quote suggests, the reservation instrument is not really designed to "prevent" TDM. Plausible though as this might seem as a means of protecting authors' moral rights (by allowing them to oppose AI training as a matter of principle),[26] the Regulation intends instead – as the second part of the quote shows – to nudge parties into bargaining, thereby instrumentalizing unilateral reservations as a conduit to create a (demand-driven) market for TDM licenses. Such market-creation is the ultimate objective of counterbalancing the TDM exception. Hence its exception-exception (*Rückausnahme*) reads:

31 "The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders [...]" – article 4(3) CDSMD

III. Who's Afraid of Article 4(3) Reservations?

32 If rightsholders can opt out of the TDM exception, some fear that this makes the law ineffective. But, which law? Two camps have expressed diametrically opposing fears:

33 For one camp, "the law" is the TDM exception, and the reservation of rights "a provision that may very well frustrate its efficacy"[27] and "will most likely leave the practice of commercial text and data mining for non-research purposes uncertain".[28] This camp expects that "all relevant providers of content will make such reservations" so that TDM would "become practically impossible" and the "purposes of the exception would get turned on their head".[29] Some authors have even advocated for abolishing article 4(3) to improve effectiveness and economic efficiency of the TDM exception.[30]

34 For another camp, "the law" is the rights reservation, which they fear might be "extremely time-consuming and consequently expensive", hence inoperable in practice.[31] As a case in point, German journalists[32] have expressed concerns that "utilising this option in any given case" will be "difficult in practice" because "very few authors have the requisite skills and knowledge to draft a reservation [...] or to monitor compliance."[33] In addition, "it can also be unclear whether reservations have been made by rightsholders themselves or at their behest, or only by a service provider (in which case they would not prevent mining)."[34] The reservation mechanism may therefore turn out to be have no practical effect at

[26] See, e.g., *de la Durantaye* (n. 20), 9 at fn. 57: "Many authors are not exclusively guided by economic interests. Quite a few of them are principally opposed to their works being used for training generative AI."

[27] *Margoni/Kretschmer* (n. 21), 695; *Picht/Thouvenin* (n. 18), 928: "The scope of these exceptions is therefore limited."; *Dusollier* (n. 19), 987: "The exception [...] is thus rather precarious"; *Geiger* (n. 15), 76: "usefulness of this provision might be rather limited [...] can make the provision rather ineffective"; *Mezei* (n. 13), 464: "We cannot but agree with the reviewers' frustration with the substance and the practical functionality of these rules.".

[28] *Christophe Geiger, Elena Izyumenko*, 'Towards a european "Fair Use" grounded in Freedom of Expression' (2019) 35 Am. U. Int. L. Rev. 1, 18–19.

[29] *Matthias Hartmann, Jonas Jacobsen*, '„Maschinenlesbarkeit" des Rechtevorbehalts im neuen § 44b UrhG' [2021] MMR-Aktuell #441332, sub I.: „praktisch unmöglich machen und damit die Ziele der Schranke in ihr Gegenteil verkehren [...] dass alle relevanten Anbieter von Inhalten einen entsprechenden Vorbehalt anbringen".

[30] In German, see *Brockmeyer*, 'Text und Data Mining: Eine rechtsökonomische Analyse der neuen Schranken im Urheberrecht' (2022), 166–170; similarly, *Emre Bayamlıoğlu*, 'Machine Learning and the Relevance of IP Rights: An Account of Transparency Requirements for AI' (2023) 31 Eur. Rev. of Priv. Law 329, 346 perceived the reservation mechanism a "major shortcoming of the provision which is likely to render it inefficient"; more cautiously, *Mezei* (n. 13), 468: "whether the CDSM Directive shall be amended, is far from being certain. [...] In general, Article 4(3) CDSM Directive shall be revisited to provide for more certainty [...] With the end of the von der Leyen Commission's tenure in 2024, this time is not 'ideal' for any such updates."

[31] *Mezei* (n. 13), 465: "how such a reservation shall operate in real life is far from clear [...] it is a doctrinal and practical minefield."

[32] For other voices from the German discussion, see *Hamann* (n. 1), 135–137 (C.IV.).

[33] Deutscher Journalisten-Verband, Legislative Amicus Brief of 6 Nov 2020 <t1p.de/1qfzk>, p. 8: „In der Praxis wird es schwierig, von dieser Option im Einzelfall Gebrauch zu machen. Die wenigsten Urheber:innen verfügen über die nötigen Fähigkeiten und Kenntnisse, einen solchen Vorbehalt in einer maschinenlesbaren Form zu verfassen und dessen Einhaltung zu kontrollieren".

[34] *Vesala* (n. 20), 357.

all.

35 Both camps' concerns are serious in view of the rationales sketched out earlier (B.II.). Inefficacy of the TDM exception might jeopardize fundamental rights of AI developers and diminish their incentives for innovation – leaving them to train their models on antiquated content in the public domain. Inefficacy of the reservation mechanism might be equally as problematic, potentially jeopardizing fundamental rights of content creators and diminishing their financial incentives for creation. As one author put it,

36 "Article 4(3) CDSM Directive cannot serve the purpose it was designed for – neither for the benefit of authors (who were the targeted beneficiaries of this provision), nor for the AI industry (whose contribution to humankind's development is unquestionable)."[35]

37 We cannot know, of course, which of the two fears is actually warranted unless we first clarify the doctrinal requirements for an effective reservation (in the next two sections) and compare them with the real potentials of current technologies (infra C.).

IV. Opt-Out Reservations in International Copyright Law

38 In order to clarify the doctrinal requirements of the reservation instrument, we need to first understand its context and prefigurations. For instance, some have criticized the opt-out model in general terms as a back-handed way to "subordinate the legislative exception to private will".[36] Yet, this exception/reservation mechanism is hardly unique in copyright law, so earlier models may provide guidance on how to construe its newest instantiation. Consider a long-established provision from the 2001 Directive upon which the CDSMD built:

39 "Member States may provide for exceptions or limitations [... for] reproduction by the press [...] of published articles on current economic, political or religious topics [...] in cases where such use is not expressly reserved" – article 5(3)c InfoSocD

40 This exception had been equally "subordinated" to "private will", allowing the press to protect "current" contents from getting reproduced, by means of reserving such use. This was itself an almost verbatim copy of a much older article in the Berne Convention, which allowed signatories to create such exceptions for "articles published in newspapers or periodicals on current economic, political or religious topics", but limited to cases in which such use was "not expressly reserved." The exact wording of this carve-out had a long and varied history since the Convention first passed in 1886:[37]

41 1886, article 7(1)1: "... unless the authors or publishers have expressly forbidden it."

42 1896, article 7(2)1 amended: "... when the authors or editors shall have expressly declared ... that reproduction is forbidden"

43 1908, article 9(2)1: "... unless the reproduction thereof is expressly forbidden."

44 1928/1948, article 9(2)1: "... unless the reproduction thereof is expressly reserved"

45 1967/1971, article 10bis(1)1: "... in cases in which [... use] is not expressly reserved"

46 As this synopsis shows, the instrument that was later implemented in article 5(3)c InfoSocD started out as a prohibition ("forbidding" users to reproduce contents) but ended up becoming a "reservation" from 1928 onwards. This semantic reorientation is meaningful in view of the purposes of the reservation instrument, and it might help to justify why nowadays, in TDM cases, the *droit moral* tends to take a back seat to market-creating incentive rationales.[38]

47 Another significant parallel with today's TDM exception is that the press exception covered materials that were once "widely believed not to be copyrightable in the first place."[39] Hence the exception could be construed as creating a new penumbra of protection, rather than dutifully protecting natural *a priori* rights. This would mean that no moral standards kept the exception from being "subserv[i]ent to its prohibition by rightholders", as is now the case for the TDM exception.[40]

[35] *Mezei* (n. 13), 462.

[36] *Rossana Ducato, Alain Strowel*, 'Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility"' (2019) 50 IIC 649, 666.

[37] Sources documented as online appendix to *Ricketson/Ginsburg*, 'International Copyright and Neighbouring Rights: The Berne Convention and Beyond', 2nd ed. 2005 <global.oup.com/booksites/content/9780198259466>.

[38] See *supra* marginal 30.

[39] *Jane C Ginsburg*, 'Berne-Forbidden Formalities and Mass Digitization' (2016) 96 Boston U. L. R. 745, 759–760 (citing to pp. 249–254 of the *travaux*, the Records of the 1908 Revision Conference).

[40] *Dusollier* (n. 19), 987.

48 Insofar as the doctrine on the reservation of press rights can actually inform the reservation of TDM rights, it is still open. While some German interest groups had proposed to directly model the transposition of article 4 CDSMD on the older reservation of press rights,[41] others have argued that

49 "the drafting history of the Berne Convention indicates that art. 10bis(1) is a 'lex specialis,' a sui generis provision that [...] does not create a basis for generalization into a technique for instituting declaratory measures."[42]

50 As we will discuss later in section V., there are some questions regarding the reservation of TDM rights on which the doctrine regarding the reservation of press rights might, arguably, be brought to bear. On the other hand, the new reservation may provide unprecedented challenges, especially regarding its territorial reach. That is because the recently passed European AI Act requires all "providers of general purpose AI models" in the European Union – no matter how liberal the jurisdiction in which they trained their models[43] – to

51 "put in place a policy to respect Union copyright law in particular to identify and respect, including through state of the art technologies, the reservations of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790" – article 53(1)c AI Act (with recital 106)

52 Not only does this obligation enforce a Brussels effect on copyright law and revisit the principle of territoriality once more.[44] It also raises the question of what "state of the art technologies" are, and how rights reservations might be made intelligible to them.[45] This question will be the focus of the latter half of this paper (C.).

V. On Standards of Expressivity and Machine-Readability

53 The legal requirements for an effective reservation of TDM rights have been described as "an aspect of the commercial TDM exception or limitation that did not spark enough discussion in the EU so far".[46] In fact, there are two standards that article 4 CDSMD requires to be fulfilled cumulatively:

54 First, as cited previously, article 4(3) CDSMD requires rightsholders to "expressly" reserve TDM uses. This element seems to create some discomfort as authors tip-toe around a clear definition,[47] and lawmakers in member states such as Germany transposed article 4(3) CDSMD through "omission of the 'express' element", despite causing "linguistic divergences in its transposition".[48] So what should "expressly" mean, if one took the requirement seriously?

55 The term does not appear elsewhere in the Directive. Yet, a recital in another context uses the adjective "explicit",[49] which most language versions of the Directive equate with "express".[50] This suggests that an "express" reservation needs to be *expressis verbis*, i.e., "explicit" rather than implicit – which excludes some technological measures that we will encounter later (C.VII.). In addition, the Directive requires that "other uses should not be affected by the reservation" (recital 18 subpar. 2 CDSMD), meaning that it needs to be *use-specific*. A third requirement can be derived from the doctrine on the reservation of press rights under the Berne Convention introduced earlier (B.IV.). During its continual reformulation,[51] article 10bis was temporarily extended by a sentence saying:

56 "In the case of periodicals it shall be sufficient if such prohibition is indicated in general terms at the beginning of each number." – article 7(2)1 Berne Convention 1896–1908

57 This sentence was dropped from later versions of the Convention, suggesting that "express" should no longer include wholesale reservations in a central location. This is well-founded in the objective of having rightsholders decide in view of specific contents whether their use should be reserved

[41] BDZV/VDZ/VDL, Legislative Amicus Brief of 31 Jan 2020 <t1p.de/ahzbb>, p. 10 („Diese Vorgabe kann durch eine Formulierung erreicht werden, die dem Rechtevorbehalt in § 49 UrhG [German transposition of article 5(3)c InfoSocD] nachgebildet ist.")

[42] See *Ginsburg* (n. 39), 759 around fn. 58.

[43] See the country survey by *Sean M.Fiil-Flynn et al.*, 'Legal reform to enhance global text and data mining research' (2022) 378 Science 951.

[44] See already Madiega (European Parliamentary Research Service), 'EU copyright reform: Revisiting the principle of territoriality', Briefing of Sep 2015 <europarl.europa.eu/RegData/etudes/BRIE/2015/568348/EPRS_BRI(2015)568348_EN.pdf>.

[45] Likewise skeptical, *Mezei* (n. 13), 469: "It is [...] far from being clear how the EU has imagined the respect of opt-out privileges via a 'policy'."

[46] *Mezei* (n. 13), 465.

[47] For instance, *Mezei* (n. 13), 465 defines "expressly" by saying that "rightholders shall openly and expressly claim ...", which is circular.

[48] *Margoni/Kretschmer* (n. 21), 695.

[49] Recital 69 CDSMD.

[50] In the French version, both "express" and "explicit" get translated to « expressément », in the German version to „ausdrücklich", in the Italian version to « espressamente ».

[51] See *supra* margin als 41-45.

or not.[52] Otherwise they could not reassess their stance vis-à-vis TDM reservations later, rendering themselves unable "to decide whether they want to include the new contents in their earlier reservations or not."[53] To sum up, the three dimensions of the "express" element preclude reservations that

> 58 "are complex, nested [or fully implied, HH] or cannot be accessed on the specific page of the content, as well as those that do not expressly refer to text and data mining".[54]

59 The second requirement of article 4(3) CDSMD is that reservations need to be made "in an appropriate manner, such as machine-readable means in the case of content made publicly available online."[55] For online content (which is most relevant for AI training), the "appropriate manner" requirement is slightly ambiguous: Due to its exemplification through "such as", machine-readability might be construed as one case of an *appropriate manner in the case of content available online*. If this reading was correct, then other (non-machine-readable) manners could be equally as appropriate. This is not, however, what the Directive intended. Its recital clarifies in most language versions[56] that

> 60 "[i]n the case of content that has been made publicly available online, it should *only* be considered appropriate to reserve those rights by the use of machine-readable means […]" – recital 18 subpar. 2 CDSMD

61 This means that the provision is correctly construed by reading *machine-readable means in the case of content available online* as an example of the "appropriate manner". In our context, therefore, the second requirement is not appropriateness in general, but machine-readability. However, as with "express", the Directive neither defines "machine-readable" nor uses it in other contexts. Very few scholars have devoted significant attention specifically to the meaning of "machine-readable",[57] despite its being a cornerstone of article 4(3) CDSMD. It also requires the most guidance due to incorporating a strictly technological concept.

62 There is a wide range of potential interpretations of "machine-readable". It could be construed conservatively or liberally. The most conservative reading would only include *native machine code*, i.e., binary-encoded commands on the base layer of CPU language. The most liberal reading might include "any digitally provided information" that can "be 'read' into a computer's working memory".[58] The range of these potential interpretations has caused great uncertainty in the transposition of article 4(3) CDSMD.[59] While there is no doubt that "machine-readable means do not exclude human-readability of the reservation",[60] powerful interest groups such as the US Motion Picture Association have lobbied for the converse: They tried to convince legislators that "any reservation that a human could read is equally as machine-readable".[61] This would mean that "machine-readable" is really just synonymous with "readable", turning the "machine" limiter into inconsequential jargon. Opposing interest groups such as the Association of European Research Libraries have correctly highlighted the "theoretical" absurdity of such a boundless conception, stressing that

> 63 "[i]t is vitally important that it is clear this relates to widely used machine readable 'standards' […] If this is not the case then anything is machine readable, and the wording is tantamount to requiring all terms and conditions on a website having to be read and interpreted by a human one by one."[62]

64 In this quote, "standards" cannot refer to mere linguistic conventions, despite what some authors suggested by proposing to exclude "lay-person phrasing in reservations" in favor of well-defined boilerplate text such as "Text und Data Mining

[52] This is also the general understanding of the respective German provision, see *Hamann* (n. 1), 149, 154 (near the end of E.I. and E.II. respectively).

[53] *Mezei* (n. 13), 468 and further: "rightholders might indeed change their mind and want to allow certain TDM activities for third parties."

[54] *Hartmann/Jacobsen* (n. 29) sub II.3: „komplexe, verschachtelte oder nicht auf der konkreten Seite der Inhalte abrufbare Vorbehalte oder solche, die nicht ausdrücklich auf das Text und Data Mining abstellen".

[55] I omitted this adverbial phrase earlier when citing article 4(3) CDSMD; it takes the place of the ellipsis at the end of section B.II.

[56] See IBM Intellectual Property Law, Legislative Amicus Brief of 6 Sep 2019 <t1p.de/u5umi>, p. 3: "In the German translation of recital 18, this understanding is unfortunately not so clear".

[57] Namely, *Hartmann/Jacobsen* (n. 29); *Lisa Löbling, Christian Handschigl, Kai Hofmann, Jan Schwedhelm*, 'Navigating the Legal Landscape: Technical Implementation of Copyright Reservations for Text and Data Mining in the Era of AI Language Models' (2023) 14 JIPITEC 499; 505–509; *Mezei* (n. 13).

[58] *Hartmann/Jacobsen* (n. 29) sub II.2.a), II.2.c): „jede digital hinterlegte Information […], denn solche Daten können in den Arbeitsspeicher eines Rechners ‚gelesen' werden."

[59] See *Hamann* (n. 1), 128–133 (D.II.).

[60] Mezei (n. 13), 466 after fn. 49.

[61] MPA, Legislative Amicus Brief of 31 Jan 2020 <t1p.de/m1c3c>, p. 2.

[62] LIBER, Legislative Amicus Brief of 31 Jan 2020 <t1p.de/hb30r>, p. 2 (no. 8).

vorbehalten".[63] Presenting this proposal *verbatim* to an international audience instantly highlights its most obvious flaw: Not quite every AI developer on the planet speaks German fluently. Even some German authors acknowledge this by advising to "reserve rights in English language (lingua franca) just in case".[64] Yet, as our introductory example shows,[65] spiders do not speak English either. The question which human language should be the "lingua franca" of TDM reservations is therefore moot. None should. Natural language, as will soon be illustrated (C.I.), is simply not amenable to sufficient standardisation. The only "machine-readable" languages can thus be artificial ones, created by well-defined technical standards.

65 This interpretation is backed by Directive (EU) 2019/1024 on Open Data and the Re-Use of Public Sector Information (PSI2D)[66] which defines "machine-readable format" as

> **66** "a file format structured so that software applications can easily identify, recognise and extract specific data, including individual statements of fact, and their internal structure" – article 2(13) PSI2D

67 While this definition, which originated in 2013,[67] directly applies only to "documents held by public sector bodies" (article 1 no. 1 PSI2D),[68] there are good reasons to use it in construing the machine-readability requirement of article 4(3) CDSMD as well.[69] After all, both instances of machine-readability serve the same purpose of automated processability. In that sense, the reservation of TDM rights is another instance of "Code is Law",[70] where a Code determines what *can* be expressed so that the Code definition becomes the authoritative interpretation of what *is* expressed.

68 However, instead of specifying a well-defined code interface, article 4(3) CDSMD "is lacking such a specification of the interface".[71] Despite the quasi-legal effect of the interface used, the legislator eschewed standard-setting and left "the number of different opt-out models" to "grow exponentially".[72] As syntax standards proliferate, we need to review them one by one to determine which ones qualify as "machine-readable" under article 4(3) CDSMD. This is the objective of the next chapter.

C. Human-to-Machine Communication of Copyright Reservations

69 Now that the legal requirements for an effective reservation of rights have been clarified, it is time to discuss the available standards. Scholars complain that "as of now, a specific technical standard is lacking",[73] and one renowned software developer's IP department emphasised the practical need for such a standard:

> **70** "It is important that technical hurdles in any transposition of Article 4(3) are kept to a minimum, because [...] any technical hurdle/limitation will quickly have ramifications on speed-to-market and progress of AI solutions. Any transposition must be kept broad and flexible enough to accommodate improvements in the advancement of the technology of defacto or standard practices."[74]

71 This quote highlights a pronounced ambiguity: On the one hand, industry needs a precisely defined standard to enhance legal certainty as a means of reducing hurdles and increasing speed-to-market. On the other, industry needs its improvements in standard practices, even *de facto* ones, to be accommodated by the law. This ambiguity, one might argue, was bound to paralyse the lawmaker and prohibit them from precisely specifying any standard of machine-readability in article 4(3) CDSMD.

72 Unsurprisingly, then, legal scholars have found "the substance and the functioning of rights reservation" to be nothing short of "a mystery".[75] A more sober

[63] *Hartmann/Jacobsen* (n. 29) sub II.2.c). This translates to "text and data mining reserved".

[64] *David Bomhard*, 'KI-Training mit fremden Daten – IP-Rechtliche Herausforderungen rund um § 44b UrhG' (2023) 14 DSRI-Tagungsband 255, 266: „sicherheitshalber immer auch in englischer Sprache (lingua franca)".

[65] See *supra* after n. 10.

[66] See *Griffiths/Synodinou/Xalabarder* (n. 20), 29 with reference to "other pieces of EU legislation".

[67] Article 1(2) and recital 21 of the Directive 2013/37/EU of 26 Jun 2013 amending Directive 2003/98/EC.

[68] Likewise, recital 1 of Directive 2013/37/EU (n. 67): "documents produced by public sector bodies of the Member States".

[69] *Griffiths/Synodinou/Xalabarder* (n. 20), 29: "article 4 DSMD should be interpreted in combination with the PSI-II Directive".

[70] *Lessig*, 'Code and Other Laws of Cyberspace' (1999), 6 (lessig.org/images/resources/1999-Code.pdf), with end note 7 (p. 241) citing, foremost, *Mitchell*, 'City of Bits: Space, Place, and the Infobahn' (1995), 111.

[71] *Hartmann/Jacobsen* (n. 29) sub II.2.b): „an einer solchen Spezifikation der Schnittstelle fehlt es".

[72] *Mezei* (n. 13), 465.

[73] *de la Durantaye* (n. 20), 10 after fn. 58.

[74] IBM (n. 56), p. 2.

[75] *Mezei* (n. 13), 465; more cautiously, *Ducato/Strowel* (n. 36), 666: "questions remain as to what the reservations in a machine-

policy brief that reviewed some (not all) potential technologies concluded dryly:

73 "There are currently no generally recognized standards or protocols for the machine-readable expression of the reservation of rights provided for in Article 4 of the Directive."[76]

74 Despite (or because of?) this perception of failed standardisation, few scholars even try to systematically review the available protocols for reservations under article 4(3).[77] Some authors do refer to some technologies, but mostly without explaining their specific functioning. Conversely, technical experts propose protocols that cannot fulfil the legal requirements set out above.

75 The question remains, which technologies are machine-readable *in the sense of article 4(3) CDSMD*. Answering it requires an interdisciplinary perspective that integrates technological process knowledge and normative reasoning. In order to illustrate this process, our introductory example (supra A.) will illustrate most of the technologies discussed hereinafter. The paper thus comes full-circle by returning to our little spider's journey through the web: Has it violated copyright? Which of the reservations that it encountered but ignored, were actually valid under the CDSM Directive?

I. Terms and Conditions

76 The CDSMD recital clarifying that "only" machine-readable reservations should be appropriate for online content was cited partially earlier.[78] In place of the quote's closing ellipsis, the recital actually reads "including [...] terms and conditions of a website or a service." (recital 18 subpar. 2 CDSMD). Some authors read this to say that terms and conditions are one example given by the Directive of machine-readable means. If this reading was correct, it would follow that "AI trainers must take into account [...] terms and conditions of websites and online services"[79] because the "language in their terms of use" might "constitute an effective reservation".[80] Indeed, the online service in our introductory example did actually include such language in its T&Cs.[81]

77 As literally apt as this reading of "machine-readable means, including terms and conditions" might seem, it would upend the entire purpose of machine-readability that we discussed earlier (B.V.). Consider the variety of potential wordings that terms and conditions might take.[82] Even in our introductory example, the T&C's current language is very different (and located in a different provision) from the previous version of the same document just months earlier.[83] This explains why IT experts assume that identifying or parsing a reservation expressed in natural language would be "difficult to near impossible" without the use of "the most sophisticated technology".[84] From one experiment on TDM reservations across 100 websites, researchers have similarly concluded that "effective opt-out management would require advanced NLP methods".[85] Yet, advanced natural language processing (NLP) is itself a case of text and data mining (TDM). It may have to rely on a corpus of reproduced website contents, which could not be in turn justified under any copyright exception. In other words, one cannot simply use TDM to find out whether using TDM is permissible.

78 In addition, neither the location nor the file format of T&C documents are standardised in any way. Some websites include reservation language in the imprint,[86] and even the T&C document for the website being scraped in our introductory example (beck-online.beck.de) was found in another domain scope (rsw.beck.de) with an English version available only as a pdf file.[87] While many websites provide T&Cs in pdf format for best printability,

readable format are, and how they could be implemented".

[76] *Keller/Warso*, 'Defining Best Practices for Opting Out of ML Training' (29 Sep 2023), OpenFuture Policy Brief #5 <openfuture.eu/wp-content/uploads/2023/09/Best-_practices_for_optout_ML_training.pdf>.

[77] Without engaging technical specifications in detail, see *Löbling et al.* (n. 57), 505–509. After finishing the first draft of this paper, I learned of a draft version of *Mezei* (n. 13), who likewise notes that "research papers either omit or struggle with these problems" (465), then reviews technologies on an issue-by-issue basis rather than explaining or even discussing each of them.

[78] *Supra* marginal 60.

[79] *Senftleben* (n. 20), 1544.

[80] *Vesala* (n. 20), 357 ("e.g. banning reverse engineering or similar methods, or the storing of available content"); likewise, *Mezei* (n. 13), 465: "There is a risk that expressed terms of end-user licence agreements can exclude the lawfulness of TDM".

[81] See *supra* at n. 7, translated in n. 8.

[82] Review of TDM terms on 21 platforms in *Ducato/Strowel* (n. 36), 669–673.

[83] See no. 10.9 of the T&Cs of 9 Mar 2022, archived on 15 Jan 2024 at <web.archive.org/20240115214414/rsw.beck.de/docs/librariesprovider138/default-document-library/general_terms_and_conditions_beck_online_2022_03_09.pdf>.

[84] IBM (n. 56), p. 2.

[85] *Löbling et al.* (n. 57), 504.

[86] See *supra* marginal 13. This may be a German *sonderweg* because the German legislator equated "metadata" (recital 18 subpar. 2 CDSMD) with "imprint", see *Hamann* (n. 1), 146–149 (E.I.).

[87] See *supra* n. 9.

this format is notoriously ill-standardised, so that even advanced algorithms cannot reliably parse it. T&C documents were simply not made to be read by machines. Experts hence argue that if "a PDF, terms and conditions etc" were considered machine-readable, then "anything on a computer screen is".[88] This would revive the lobby position rejected earlier that "readable" and "machine-readable" are synonymous (marginal 62 at n. 61).

79 Given these challenges, the Directive's recital needs to be corrected by inserting the missing preposition "in": The correct construal of recital 18 has rightsholders "reserve those rights by the use of machine-readable means, including *in* terms and conditions". Consequently, AI developers may ignore any "reservations not expressed in code", which includes (but is not limited to) "when TDM restrictions are found in website terms and conditions in PDFs, images or as website text".[89]

80 Even if courts came to view this question differently and accepted at least some T&C documents written in natural language as "machine-readable", then the additional requirement of "express" reservation still limits its effect to the document within which it is found (i.e., the terms document itself). As discussed earlier for "wholesale reservations in a central location",[90] reservation statements cannot affect multiple contents because each of them needs to be reserved "expressly", i.e., content-specifically.

II. Robots Exclusion Protocol (robots.txt)

81 Apart from its terms and conditions, the website in our introductory example also reserved TDM rights in a file called robots.txt.[91] This file has aptly been called "the text file that runs the internet" because even five years ago, it was used on half a billion websites according to 2019 estimates by Google.[92] Each of these text files instantiates "an exclusion protocol that content providers can insert into the root directory to prevent crawling or indexing activities".[93] The protocol was proposed in 1994 by Dutch search engine pioneer *Martijn Koster* and became a *de facto* "established standard"[94] for repelling search engine spiders. Its formal canonization is rather recent, as the Internet Engineering Task Force (IETF) formalized this "Robots Exclusion Protocol" (REP) as an official standard in 2022.[95]

82 Given its widespread use and its machine-readability (except for comments in natural language, see example supra n. 5), the Robots Exclusion Protocol was quickly proposed – both by special interest groups[96] and academics[97] – as a suitable standard for reservations under article 4(3) CDSMD. Indeed, an ongoing empirical survey of 886 US-American and 273 other news portals from 31 countries shows that currently two-fifths of them (40.7 %) deny access in their robots.txt to the same spiders as the website in our introductory example (at marginal 6), while more than half of them (54.3 %) deny access to at least one of the spiders from the introductory example, or that of Google AI.[98]

83 It is important to note that by its very definition the Robots Exclusion Protocol is "not a form of access authorization" (rule 1 subpar. 4 REP), but a collection of "rules [...] that crawlers are requested to honor" (rule 1 subpar. 3 REP). It therefore does not really prevent spiders from entering a website,[99] but simply requests them to stay out. The REP is therefore best understood as a form of "Private Ordering Through Opt-Outs".[100] Some large crawlers openly defy the

[88] LIBER (n. 62), p. 2 (no. 8).

[89] *Griffiths/Synodinou/Xalabarder* (n. 20), 30; *Löbling et al.* (n. 57), 502.

[90] See *supra* marginal 57.

[91] See *supra* marginals 5 and 6.

[92] *Pierce*, 'The text file that runs the internet', The Verge, 14 Feb 2024 <theverge.com/24067997/robots-txt-ai-text-file-web-crawlers-spiders>.

[93] *Ducato/Strowel* (n. 36), 674; IBM (n. 56), 2: "a protocol/format that is used widely by web crawlers and web robots today".

[94] *Hartmann/Jacobsen* (n. 29) sub II.3): „So ist ein Standard etabliert, Anweisungen an Suchmaschinen in einer spezifischen Datei abzulegen (‚robots.txt').“

[95] *Koster/Illyes/Zeller/Sassman*, 'Standard RFC 9309: Robots Exclusion Protocol', as of Sep 2022, documented at <rfc-editor.org/rfc/rfc9309.html>.

[96] As two of just many, see IBM (n. 56), 2, and LIBER (n. 62), 2 (no. 8).

[97] *Ducato/Strowel* (n. 36), 674; *Dusollier* (n. 19), 987: "machine-readable means as robots.txt files"; *Tan/Lee* (n. 104), 1039: "owners may even adopt a Robots Exclusion Protocol"; *Senftleben* (n. 20), 1544: "AI trainers must take into account metadata, such as robots.txt files"; *Löbling et al.* (n. 57), 502: "setting up a robots.txt file can express an opt-out" (similarly ibid., 506); *Griffiths/Synodinou/Xalabarder* (n. 20), 25 after fn. 42: "machine-readable means, including [...] robot.txt type metadata"; *Mezei* (n. 13), 467: "inclusion of relevant computer-readable language in the robots.txt file".

[98] Own analysis of data by *Welsh*, 'Who blocks OpenAI, Google AI and CC?', palewire, accessed on 2 Apr 2024 <palewi.re/docs/news-homepages/openai-gptbot-robotstxt.html>: 629 of 1.159 news publishers disallow either Google AI („Google Extended“), OpenAI („GPTBot“, „ChatGPT-User“) or Common Crawl („CCBot“). 472 disallow only the latter two, 421 disallow all three.

[99] See *infra* at marginal 128.

[100] *Matthew Sag*, 'Copyright and Copy-Reliant Technology' (2009) 103 Nw. U. L. Rev. 1607, 1666–1668.

Robots Exclusion Protocol,[101] and there are good reasons not to rely on it for communicating TDM reservations either:

84 First, a spider's name (so-called *product token*) cannot uniquely identify it because under the Robots Exclusion Protocol, "crawlers set their own name" (rule 2.2.1 REP). This is why our introductory example said that "our spider might have called itself...". Some spiders do not identify themselves at all,[102] and "many others attempt to operate in relative secrecy"[103] or to "maliciously bypass REPs".[104]

85 Second, the list of product tokens at <robotstxt.org/db.html> has not been updated since 2011, which means that even identifying today's AI spiders requires a lot of traffic analysis.[105] Major AI developers reacted to this issue promptly by officially announcing their crawlers' tokens[106] – probably not least in hopes of evading more effective regulation by supporting the dated Robots Exclusion Protocol.

86 Third, the Robots Exclusion Protocol cannot communicate reservations for large amounts of content. The protocol allows crawlers to adopt a "parsing limit to protect their systems" (rule 2.5 REP), whereby they need not process more than 512,000 characters of a given robots.txt file ("parsing limit must be at least 500 kibibytes"). If a website of just a few hundred content files sought to communicate TDM reservations for each of those files to each known crawler, it would quickly exceed the parsing limit and fail its purpose.[107] If, instead, the TDM reservation was couched in general terms (as in our introductory example[108]) it could no longer be content-specific and would fall short of the expressivity standard, as discussed previously.

87 Fourth, the Robots Exclusion Protocol defines only two potential declarations to begin with: "Allow" to designate contents that are free to crawl, and "Disallow" for others (rule 2.2.2 REP). Additional declarations could be made,[109] but they are not standardised. So, what does a "Disallow" declaration mean? The protocol does not precisely define its purpose other than stating that "it may be inconvenient for service owners if crawlers visit the entirety of their URI space." (rule 1 subpar. 3 REP) This harkens back to the early days of the Internet when search engine crawlers caused so much traffic that "all it took was a few robots overzealously downloading your pages for things to break and the phone bill to spike."[110] This purpose is no longer relevant, so a different rationale has taken its place:

88 "It's been a while since 'overloaded servers' were a real concern for most people. 'Nowadays, it's usually less about the resources that are used on the website and more about personal preferences,' says John Mueller, a search advocate at Google."[111]

89 Yet, since the Robots Exclusion Protocol was never meant to communicate sophisticated preferences and their subtle distinctions, its binary syntax ("geared toward search engine crawlers") does "not necessarily serve" other purposes.[112] In particular, it cannot communicate conditional permissions, as would be needed to reserve TDM content for automatable commercial licensing.[113] The REP cannot even distinguish between different crawling purposes, so that bots serving multiple purposes (e.g., search engine indexing *and* AI data collection) cannot be rejected for the latter reason without also engendering the former.[114] This is exactly what the Directive's "express" requirement should avoid.[115]

[101] *Pierce* (n. 92): "The Internet Archive, for example, simply announced in 2017 that it was no longer abiding by the rules of robots.txt. [...] And that was that."

[102] See *Wiese*, 'Robots.txt is not the answer', Search Engine Land, 18 Jul 2023 <searchengineland.com/robots-txt-new-meta-tag-llm-ai-429510>.

[103] *Pierce* (n. 92), and further: "finding a sneaky crawler is needle-in-haystack stuff".

[104] *David Tan, Thomas Lee Chee Seng*, 'Copying Right in Copyright Law - Fair Use, Computational Data Analysis and the Personal Data Protection Act' (2021) 33 Sing. Acad. Law J. 1032, 1070: "a key scenario is when web robots maliciously bypass REPs".

[105] *Waldvogel*, 'How to block AI crawlers with robots.txt', netfuture.ch of 9 Jul / 31 Dec 2023 <netfuture.ch/2023/07/blocking-ai-crawlers-robots-txt-chatgpt>.

[106] OpenAI christened its „GPTBot" on 8 Aug 2023 (platform.openai.com/docs/gptbot), Google introduced the product token "Google-Extended" on 28 Sep 2023 (developers.google.com/search/docs/crawling-indexing/overview-google-crawlers).

[107] *Wiese* (n. 102).

[108] See *supra* marginal 6: "Disallow: /", where the forward slash denotes all contents of the website.

[109] According to rule 2.2.4 REP, "crawlers MAY interpret other records that are not part of the robots.txt protocol – for example, 'Sitemaps'".

[110] *Pierce* (n. 92).

[111] *Pierce* (n. 92).

[112] *Graham* cited in *Pierce* (n. 92).

[113] See *infra* marginal 114.

[114] *de la Durantaye* (n. 20), 10 at fn. 60: "robots.txt files do not allow for differentiation: If you communicate that you do not wish your website to be scraped for training purposes, it will not appear in search engines either. De facto, then, your work will cease to exist online."; similarly, *Löbling et al.* (n. 57), 505 who thus propose a reform of the REP standard (507–509) but do not address any of the other aforementioned concerns.

[115] See recital 18 subpar. 2 CDSMD, cited *supra* marginal 60.

III. Spawning Protocol (ai.txt)

90 Given these limitations of the Robots Exclusion Protocol, a newer standard has been proposed to "keep yourself searchable, while restricting AI training".[116] Or so runs the sales pitch of Minneapolis-based startup "Spawning" founded by musician *Holly Herndon*.[117] This startup set out on a mission to develop "data governance for generative AI", and more broadly to "build the consent layer for AI" by collaborating with major actors on both sides: AI developers such as Hugging Face and Stability AI as well as repertoire owners such as Shutterstock and ArtStation.[118]

91 One of the first Spawning products is a protocol presented on 30 May 2023 under the moniker ai.txt, which caught the attention of only a few legal scholars.[119] It strongly resembles the robots.txt discussed in the previous section (with which it shares a similar syntax placed as a text file in the root folder and voluntarily respected by crawlers), but a thorough comparison is hindered by a lack of public documentation.

92 From what Spawning's website reveals, its protocol seems to be an improved version of the REP in at least two dimensions of expressivity: Regarding use-specificity, TDM reservations in ai.txt are stored separately and apart from search index permissions in robots.txt. Regarding content-specificity, ai.txt is designed to be checked whenever a file is accessed through the proprietary "Spawning API" (a programming interface sold to AI developers), whereas robots.txt gets accessed only once upon entering a website through the landing page ("front door") and never laterally by direct hyperlink.[120] However, he extent to which Spawning has addressed other shortcomings of the REP (parsing limit, lack of conditional permissions, etc.) remains unclear.

IV. HTTP Response Header (tdm-reservation, X-Robots-Tag)

93 Another technology has rarely ever been discussed in relation to article 4(3) CDSMD,[121] namely Response Headers in the HyperText Transfer Protocol (HTTP). What this means is simply the machine-readable reply of a server to a file request sent by a user, as illustrated in our introductory example (marginal 4).

94 This reply starts with a status code (in our example, "200" for "OK") and delivers additional "meta" data (from Greek μετά for "after, behind; among, between"[122] in the sense of "appended" data that accompany, describe or categorize the datarequested). By virtue of this meta-communication, the Hypertext Transfer Protocol allows content-specific communication in relation to concrete files, which better fulfils the expressiveness requirement than any general reservation in a centrally located text file. As a large tech company's IP department explained,

> **95** "the most feasible method for checking reservation of rights for online content is by using common metadata. Using metadata would overcome the issue of readability as tools to parse metadata can be implemented fairly trivially and economically."[123]

96 In fact, even the Directive itself suggested "metadata" as a potential location for machine-readable reservations (recital 18 subpar. 2 CDSMD). This has been taken up by a community group of the World Wide Web Consortium (W3C), who recently proposed the Hypertext Transfer Protocol as one of three standards for implementing TDM reservations.[124] Unfortunately, their multi-pronged *TDM Reservation Protocol* ("TDM ReP") has received little attention in legal literature thus far.[125]

97 The core of this proposal is to insert into a server's response a meta declaration "tdm-reservation" with value 1 and a meta declaration "tdm-policy" containing the URL for a file containing contractual details (rule 6.2 TDM ReP) – as has been done in our introductory example.[126] In our example, the TDM policy file contained no contractual details, but merely the same proviso as the website's imprint:

[116] Spawning ai.txt, accessed 7 Mar 2024 <spawning.ai/ai-txt> and <site.spawning.ai/spawning-ai-txt>.

[117] See *Dredge*, 'Holly Herndon reveals plans for her AI-focused startup Spawning', music:)ally of 16 Nov 2023 <musically.com/2023/11/16/h>.

[118] About Spawning, accessed 7 Mar 2024 <spawning.ai/about>.

[119] See *Keller/Warso* (n. 76), 8–9; *Mezei* (n. 13), 467-468.

[120] *Miller*, 'ai.txt: A new way for websites to set permissions for AI', Spawning Blog on 30 May 2023 <spawning.substack.com/p/aitxt-a-new-way-for-websites-to-set>.

[121] Only *Mezei* (n. 13), 467 casually mentions "declaring a choice in an HTTP response".

[122] See <etymonline.com/word/meta->.

[123] IBM (n. 56), p. 2.

[124] W3C TDMRep Final Community Group Report of 2 Feb 2024 (w3c.github.io/cg-reports/tdmrep/CG-FINAL-tdmrep-20240202).

[125] See only *Keller/Warso* (n. 76), 7–8; *Löbling et al.* (n. 57), 507; *de la Durantaye* (n. 20), 10 in fn. 60; *Mezei* (n. 13), 467 at fn. 60.

[126] See *supra* marginal 4.

98 “Text and Data Mining according to § 44b UrhG: The publisher reserves the right to reproduce for text and data mining according to § 44b UrhG.”[127]

99 Since this “policy” is akin to T&Cs, it is equally as non-machine-readable.[128] If it were to become machine-readable, the policy file could not be written in HTTP syntax, because as a transfer protocol it is limited to short, transfer-related responses. Another language protocol would be required in addition, and we will later encounter examples (including another proposal by the W3C community group) of how such policies might be encoded machine-readably (infra C.VI.).

100 As an additional limitation, it is worth noting that unlike the Robots Exclusion Protocol, the TDM Reservation Protocol is not without alternatives. There have been at least two other proposals for reservation standards based on the Hypertext Transfer Protocol. Both repurpose the meta declaration “X-Robots-Tag”, which (like robots.txt) had once been developed to control search engine indexing:

101
```
X-Robots-Tag: noai, noindex[129]
X-Robots-Tag: usage-rights: CC-BY, noindex[130]
```

102 While these proposals are unlikely to outcompete the TDM Reservation Protocol with its authoritative backing (W3C) and well-crafted, open documentation, the race has not been run yet and it is too early to tell which variant will be adopted more widely.

V. HyperText Markup Language (<meta>, data-notdm)

103 Another type of metadata appears in our introductory example at marginal 9. These are the “meta elements of an HTML-conformant website”, which some legal scholars have considered a suitable medium for TDM reservations.[131]

104 HTML (HyperText Markup Language) is a so-called markup language, i.e., a human-readable text format that allows to encode both *semantic* content and *syntactic* information. Just like natural language structures text through syntax elements (such as these brackets, which separate parenthetical comments and illustrations from the main text), the Hypertext Markup Language spins structuring information off into so-called “tags” using less-than- and greater-than-signs to stand in for <angled brackets>. For example, in the text quoted earlier (marginal 13) both occurences of the <br> tag would have been rendered by any browser as an on-screen line break.

105 Despite sharing the moniker “metadata” with hypertext transfer metadata, hypertext markup metadata are not “appended” to a file, but to its content instead. Using an analogy from the physical world, one could say that HTTP metadata are like the packing slip of a book, while HTML metadata are its imprint. The latter is placed within the book but nonetheless appended to its actual content. The analogy shows that metadata in the Hypertext Transfer Protocol and in the Hypertext Markup Language serve very different purposes, even though some information may be contained in both (like the book title or year of publication in our metaphor) while others only make sense in one of the two places (like the date of delivery in a packing slip and the names of illustrators in an imprint).

106 Returning to the introductory example, tagged metadata make up most of the “eighty lines […] in machine language” mentioned in marginal 8. Hence, rightsholders might consider “using tags” as “a predefined format/syntax” for their TDM reservations.[132] Indeed, the TDM Reservation Protocol[133] refers to HTML tags of the class <meta …> as its second prong for communicating TDM reservations (rule 6.3 TDM ReP). This would use the same attributes as in the HTTP Response Header, namely “tdm-reservation” and “tdm-policy” with the values of 1 and the policy URL, respectively.

[127] Quote from <rsw.beck.de/beck-online-service/tdm-vorbehalt>, accessed 7 Mar 2023. For the corresponding imprint language, see *supra* marginal 13.

[128] Rule 5.2 TDM ReP: “A TDM Policy is considered human readable if its content-type is text/html. It is considered machine-readable if its content-type is either application/json or application/ld+json.”; *Löbling et al.* (n. 57), 507: “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”.

[129] *Emanuel Maiberg*, ‘An AI Scraping Tool Is Overwhelming Websites With Traffic’, VICE, 25 Apr 2023 <vice.com/en/article/dy3vmx/a> on “Romain Beaumont, the creator of the image scraping tool img2dataset” who designed it “to scrape images from any site unless site owners add https headers like ‘X-Robots-Tag: noai,’ and ‘X-Robots-Tag: noindex.’”

[130] *Wiese* (n. 102), explaining this reservation as “the page should not be used for search results but can be used for commercial LLMs as long credit is given to the source”, but without clarifying how a general prohibition against TDM should be communicated (or whether it be included in “noindex”).

[131] *Hartmann/Jacobsen* (n. 29) sub II.3); *Löbling et al.* (n. 57), 506: “meta tags could serve as suitable machine-readable methods to accurately convey opt-outs for TDM”.

[132] IBM (n. 56), p. 2.

[133] See *supra* marginal 96.

107 Since a hypertext markup file can contain multiple <meta …> tags, this would even let rightsholders distinguish between different contents of the same file, enabling them to set highly granular permissions. On the other hand, it only works in HTML-conformant files; the sole other format covered by the TDM Reservation Protocol are e-books in .epub format (rule 6.4 TDM ReP).

108 The standard envisioned by the TDM Reservation Protocol gets jeopardized by a considerable proliferation of HTML-based standards. Including the TDM ReP, at least five different <meta> tags have been proposed since 2012 to reserve TDM rights:

109
```
<meta name="CCBot" content="nofollow">[134]
<meta name="robots" content="noai, noimageai">[135]
<meta name="usage-rights" content="CC-BY-SA" />[136]
<meta name="generative-ai" content="notraining">[137]
<meta name="tdm-reservation" content="1"> <meta
name="tdm-policy" content="…">[138]
```

110 Even the website of a major legal publisher known to be highly rights-sensitive uses just two of these five declaration standards.[139] Not to speak of other proposals that rely not even on <meta> tags, but on newly minted HTML attributes such as “data-notdm”.[140]

VI. JavaScript Object Notation (tdmrep.json, Reich’s ai.txt, C2PA)

111 The third and final protocol utilized by the World Wide Web Consortium’s community group was JavaScript Object Notation (JSON), a language specified since 1997 in two standards (RFC 8259 und ECMA-404). The website in our introductory example does not seem to use this language yet, which is unsurprising given JSON’s powerful-yet-demanding scripting syntax.

112 According to rule 6.1 of the TDM Reservation Protocol, reservations can be declared by placing a text file with the filename *tdmrep.json* in the root directory, wherein information get encoded as pairs of attribute (e.g., „vcard:hasEmail“) and value (e.g., “mailto:contact@provider.com“). These can be grouped and nested – as is common in many machine languages – through brackets and indentation. This enables rightsholders to even encode legal obligations by implementing, for example, the “Open Digital Rights Language” (ODRL).[141] One such sample declaration might read:[142]

113
```
"permission": [{
      "action": "tdm:mine",
      "duty": [{
        "action": "compensate"
        }]
}]
```

114 This code snippet defines a “permission”, wherein the permissible “action” (of text and data mining) is coupled with a “duty”, which itself is an “action” (of compensating). In other words, the code contains a contractual offer for a paid TDM license.[143] This syntax for what is essentially an automatable “smart contract” transcends any simplistic Allow/Disallow dichotomy and empowers users to create more complex obligations which actually serve the Directive’s objective of market creation (see *supra* marginal 30 at the end). Insofar, this component of the TDM Reservation Protocol is truly visionary. At the same time, it is by far the most demanding (and, consequently, error-prone) coding language yet proposed in the TDM reservation context.

115 It is not unique either. In a “Guide for Preparing Website Content for Large Language Models” published online on 18 May 2023,[144] AI entrepreneur *Robert Reich* proposed another standard, meant to be “more akin to RSS feeds than robots.txt”,[145] which is seemingly JSON based. In addition to lacking a public documentation, it shares another point in common with the Spawning protocol discussed earlier: It is meant to be published in a file called ai.txt. This

[134] Common Crawl FAQ since 6 Dec 2012 <commoncrawl.org/faq>.

[135] DeviantArt, ‘UPDATE All Deviations Are Opted Out of AI Datasets’, 11 Nov 2022 <deviantart.com/team/journal/UPDATE-All-Deviations-Are-Opted-Out-of-AI-Datasets-934500371>, using yet another “robots” attribute originally designed for search engines.

[136] *Wiese* (n. 102) without clarifying how a prohibition against TDM should be communicated.

[137] *Bustos*, ‘Generative AI in web development. A new AI meta tag?’, LinkedIn on 29 Jul 2023 <linkedin.com/pulse/generative-ai-web-development-new-meta-tag-eduardo-bustos>, proposed less in view of article 4(3) CDSMD, but in view of excluding AI *output* from future training in order to avoid “feedback loop[s] result[ing] in a degradation of the model”.

[138] Rule 6.3 TDM ReP, see marginal 106.

[139] See *supra* marginal 9.

[140] Notably, *Löbling et al.* (n. 57), 509.

[141] See ODRL Information Model 2.2 (W3C Recommendation) of 15 Feb 2018 <w3.org/TR/odrl-model>.

[142] From example 14 in rule 7.1.5.3 TDM ReP.

[143] The snippet does not define the price (as an *essentiale negotii*) but it could be specified using the “payment” element and its attributes, see example 21 in the ODRL Information Model (n. 141).

[144] User *menro*, ai.txt, accessed 7 Mar 2024 <github.com/menro/ai.txt>.

[145] User *menro* (n. 144).