Werkt de mijnwerk opt-out voor mijn werk?

Arnoud Engelfriet en Dirk Visser*

De opkomst van generatieve AI en de noodzaak deze met steeds grotere hoeveelheden data te trainen heeft de auteursrechtelijke uitzondering voor gebruik voor tekst- en datamining (TDM) een forse nieuwe impuls gegeven. Naast TDM voor wetenschappelijk onderzoek staat de Auteurswet namelijk ook TDM voor niet-wetenschappelijke doeleinden, en daarmee dus ook commerciële diensten toe, zij het met een 'opt-out' of voorbehoudregeling voor rechthebbenden. Dit artikel onderzoekt in hoeverre deze opt-out regeling praktisch toepasbaar is.

De opkomst van datamining

Sinds de opkomst van internet worstelt het auteursrecht met de grenzen van hergebruik van informatie. De onstilbare datahonger van Artificial Intelligence diensten zoals ChatGPT heeft de discussie hierover in een stroomversnelling gebracht: de huidige versie van ChatGPT is getraind op ruim een *petabyte* aan data, een één met 15 nullen oftewel grofweg de tekst van een miljard boeken.[1] Het is een voorbeeld van 'tekst- en datamining' (TDM), een begrip waar vele definities van zijn maar dat de DSM-Richtlijn definieert als *'een geautomatiseerde analysetechniek die gericht is op de ontleding van tekst en gegevens in digitale vorm om informatie te genereren zoals, maar niet uitsluitend, patronen, trends en onderlinge verbanden'*.[2]

TDM is lange tijd een ondergeschoven kindje gebleven. De opkomst van Big Data begin jaren tien bracht weliswaar meer aandacht, maar de rechtsonzekerheid over het mogen mijnen van andermans werken bleef beperkt tot analyse op eigen data of specifieke vakgebieden.[3] TDM geniet bijvoorbeeld populariteit bij biomedisch onderzoek, en het vakgebied van de natuurlijke taalverwerking (NLP) maakt ook dankbaar gebruik van taalanalyse. De explosief gegroeide populariteit van tekst- en beeldgeneratoren zoals ChatGPT, DALL-E en Midjourney – meer algemeen *generatieve AI* geheten – heeft de discussie over AI, auteursrecht en TDM op scherp gezet.

De noodzaak voor TDM bij generatieve AI is simpel: het AI model trainen op meer data levert direct een enorme sprong in kwaliteit op.[4] De eenvoudigste manier om de enorme hoeveelheid benodigde gegevens te verkrijgen is het hele internet aflopen en alles *downloaden* dat men aantreft. En het lijkt erop dat dit inmiddels ook gebeurt door veel aanbieders van generatieve AI. De zakelijke furore die nu wordt gemaakt met deze diensten en de mate waarin dit op bestaand werk lijkt te leunen, maakt dat rechthebbenden (groot én klein) zich nog eens achter de oren krabben: hoezo is dit geen auteursrechtinbreuk, en waarom krijgen wij geen vergoeding voor zulke grootschalige inzet van creatieve werken?

Datamining in het auteursrecht

Tot de introductie van de DSM Richtlijn stond het onderwerp van auteursrecht versus TDM maar beperkt in de belangstelling.[5] In 2016 besloot de Europese Commissie, onder verwijzing naar rechtsonzekerheid bij partijen die TDM toepassen, een expliciete uitzondering op te nemen in het ontwerp voor de DSM-richtlijn voor reproductiehandelingen in het kader van TDM. Deze rechtsonzekerheid betrof een ruime uitleg van het reproductiebegrip van artikel 2 van de Auteursrechtrichtlijn, die vrijwel iedere TDM handeling zou raken. Dit voorstel was destijds wel beperkt tot TDM door niet-commerciële onderzoeksorganisaties en cultureel erfgoedinstellingen; TDM voor andere (en daar-

* Mr. ir. A.P. Engelfriet is partner en Chief Knowledge Officer bij adviesbureau ICTRecht te Amsterdam. Prof. mr. D.J.G. Visser is hoogleraar IE in Leiden, advocaat in Amsterdam (Visser Schaap & Kreijger) en redacteur van dit blad.

1 Y. Zhong, J. Lian & H. Huang, 'Uncovering the Affordances of ChatGPT in Education from a Social-Ecological Perspective: A Data Mining Approach', *SSRN* 2023. Beschikbaar via https://ssrn.com/abstract=4518523.

2 Overweging 18 van Richtlijn (EU) 2019/790 van het Europees Parlement en de Raad van 17 april 2019 inzake auteursrechten en naburige rechten in de digitale eengemaakte markt en tot wijziging van Richtlijnen 96/9/EG en 2001/29/EG ('DSM-Richtlijn').

3 S. Filippov, 'Mapping Text and Data Mining in Academic and Research Communities in Europe', *The Lisbon Council* mei 2014.

4 J. Hestness e.a., 'Deep learning scaling is predictable, empirically', arXiv preprint arXiv:1712.00409, 2017.

5 J. Triaille, J. de Meeûs d'Argenteuil & A. de Francquen, 'Study on the Legal Framework of Text and Data Mining (TDM)', Europese Unie 2014.

mee dus ook commerciële) doeleinden is pas in een later stadium aan het voorstel toegevoegd. Vermoedelijk is de reden daarvoor geweest dat men zich realiseerde dat als er geen beperking ten gunste van commerciële TDM zou komen, die vorm van TDM helemaal onder het verbodsrecht zou vallen, hetgeen niet wenselijk geacht werd voor het investeringsklimaat in deze technologie.

De TDM-excepties zijn terug te vinden in artikelen 3 en 4 van de DSM Richtlijn, welke zijn omgezet in artikelen 15n en 15o Auteurswet. Artikel 15n is specifiek geschreven voor wetenschappelijk onderzoek en bevat een specifieke uitzondering op reproductiehandelingen in het kader van TDM, inclusief het recht brongegevens tijdelijk te mogen bewaren. Voor de praktijk rondom generatieve AI is artikel 15o het meest relevant. Dit artikel beperkt haar toepassingsgebied namelijk niet tot een specifieke doelgroep of tot gebruikers met een niet-commercieel oogmerk, maar bevat een uitzondering ten aanzien van reproductiehandelingen van 'rechtmatig toegankelijke werken en andere materialen' in het kader van TDM. De enige aanvullende restrictie staat in lid 3: het recht van TDM geldt niet wanneer dit 'op passende wijze uitdrukkelijk is voorbehouden'. Critici van het nieuwe stelsel verwelkomen de Europese harmonisatie maar menen ook dat er nog altijd rechtsonzekerheid is gezien de ruimte voor interpretatie die de DSM-richtlijn laat, vooral ten aanzien van wat een afdoende rechtenvoorbehoud is.[6] Voorbehouden gebruik vereist een licentie, waardoor afhankelijkheid zou ontstaan van de bereidheid van rechthebbenden die te geven, weliswaar tegen redelijke en niet-discriminerende voorwaarden.[7] De meer op 'fair use' gebaseerde benadering die in andere regio's, zoals AI-zwaargewicht de VS wordt gekozen, zou volgens deze kritiek beter recht doen aan de balans tussen auteursrecht en AI.[8] Ook wordt gewezen naar Japan waar TDM op basis van een expliciete wettelijke uitzondering zonder specifieke beperkingen is toegestaan.[9]

In de literatuur wordt inmiddels ook wel gesuggereerd een stelsel van wettelijke heffingen een oplossing zou zijn.[10] Hierbij wordt het TDM-exploitanten (onbeperkt?) toegestaan om datamining uit te voeren, in ruil voor betaling van een wettelijke heffing aan een incasso-organisatie. Het voorbehoud zou dan verdwijnen, of misschien alleen in zeer uitzonderlijke gevallen toegepast mogen worden. Eén voorstel is dat de heffing wordt betaald door de aanbieders van de AI-toepassingen die drijven op de via TDM verkregen data, terwijl de verveelvoudigingshandelingen gebeuren door de TDM-exploitanten en dat de opbrengst in een fonds voor cultuursubsidie worden ondergebracht.[11] Daarbij is er dan geen distributie van gelden op basis van gebruik of populariteit, omdat gegevens daarvoor ook ontbreken. Een dergelijke oplossing blijft hier verder buiten beschouwing.

Artikel 15o Auteurswet

In dit artikel analyseren wij nader de individuele componenten van artikel 15o. Dit artikel luidt als volgt:

'Onverminderd het bepaalde in artikel 15n wordt een reproductie in het kader van tekst- en datamining niet als **inbreuk op het auteursrecht** op een werk van letterkunde, wetenschap of kunst beschouwd mits degene die de tekst- en datamining verricht **rechtmatig toegang** heeft tot het werk en het auteursrecht door de maker of zijn rechtverkrijgenden niet **uitdrukkelijk op passende wijze is voorbehouden**, zoals door middel van **machinaal leesbare middelen** bij een online ter beschikking gesteld werk.'

De vetgedrukte passages worden hieronder een voor een behandeld.

Inbreuk op het auteursrecht

Artikel 15o is geformuleerd als een uitzondering op het auteursrechtelijk reproductierecht. Zoals hierboven aangegeven veronderstelt dit dat TDM een handeling vormt die in beginsel inbreuk op het auteursrecht zou opleveren. Op deze veronderstelling is fundamentele kritiek mogelijk. De verwerking die bij datamining plaatsvindt, is gericht op extractie van feitelijke informatie of statistische patronen, niet op reproductie, verspreiding of openbaarmaking van (delen van) het werk als zodanig.[12] Anders gezegd: niet het

6 J. Griffiths, T. Synodinou & R. Xalabarder, '*Comment of the European Copyright Society Addressing Selected Aspects of the Implementation of Articles 3 to 7 of Directive (EU)*', *GRUR* 2023 nr. 790, p. 22-36.

7 P.B. Hugenholtz, 'Artikelen 3 en 4 DSM-richtlijn: tekst- en datamining', *AMI* 2019-5, p. 167-171.

8 M. Senftleben e.a., 'Ensuring the Visibility and Accessibility of European Creative Content on the World Market: The Need for Copyright Data Improvement in the Light of New Technologies', *SSRN* 2021.

9 Article 30-4 van de Japanse Auteurswet: '*It is permissible to exploit a work, in any way and to the extent considered necessary, in any of the following cases, or in any other case in which it is not a person's purpose to personally enjoy or cause another person to enjoy the thoughts or sentiments expressed in that work; provided, however, that this does not apply if the action would unreasonably prejudice the interests of the copyright owner in light of the nature or purpose of the work or the circumstances of its exploitation: […] if it is done for use in data analysis (meaning the extraction, comparison, classification, or other statistical analysis of the constituent language, sounds, images, or other elemental data from a large number of works or a large volume of other such data*'. www.japaneselawtranslation.go.jp.

10 Zie bijv. C. Geiger & V. Iaia, 'The forgotten creator: Towards a statutory remuneration right for machine learning of generative AI', *Computer Law & Security Review*, 2024, 52, p. 105925.

11 Zie de voorstellen van Martin Senftleben in diverse publicaties, waaronder 'Generative AI and Author Remuneration', in: *IIC – International Review of Intellectual Property and Competition Law*, 2023, vol. 54, pp. 1535-1560.

12 M.W. Carroll, 'Copyright and the progress of science: Why text and data mining is lawful', *UC Davis L. Rev.*, 2019, 53, p. 893 (uitgaande van Amerikaans auteursrecht). Zie ook T. Margoni & M. Kretschmer, 'The Text and Data Mining exception in the Proposal for a Directive on Copyright in the Digital Single Market: Why it is not what EU copyright law needs', UK Copyright and Creative Economy Centre University of Glasgow Technical Report, 2018.

werk wordt gekopieerd, maar alleen de feitelijke of statistische informatie die daarin vervat is.[13]

In een uitgebreide analyse van de diplomatieke conferenties omtrent de Berner Conventie (1967) en het WIPO-Auteursrechtverdrag (1996) laat Senftleben zien dat het niet de bedoeling van de verdragsluitende staten was om een handeling onder het auteursrecht te brengen *'when, from the outset, copies are not made for the purpose of allowing an individual to perceive, reproduce or otherwise communicate the work'*.[14] Anders gezegd: hier is geen sprake van 'mededeling aan het publiek', omdat het publiek bestaat uit mensen en bij TDM nemen mensen het werk niet waar.[15]

Hoewel het auteursrechtelijk reproductierecht Europees is geharmoniseerd en (zeer) ruim wordt uitgelegd, blijft discussie mogelijk over de vraag of TDM in de vorm van generatieve AI toepassingen auteursrechtelijk relevant zou moeten worden geacht. Maar de discussie is inmiddels grotendeels verschoven naar de vraag of TDM in de vorm van generatieve AI toepassingen onder een beperking valt. Binnen het EU systeem zal bij het reproduceren in het kader van TDM ten eerste de vraag aan de orde zijn in hoeverre een beroep kan worden gedaan op de uitzondering voor tijdelijke reproducties van voorbijgaande of incidentele aard (art. 13a Auteurswet). In de *Infopaq* zaken oordeelde het HvJ EU dat het digitaal scannen van artikelen om op basis van automatische woordherkenning een selectie te maken waarbij de scans later weer worden gewist, binnen deze exceptie viel.[16] Daar staat echter tegenover dat voor de reproduceerbaarheid van onderzoek en het kunnen hertrainen van machine learning modellen het verworven corpus aan data juist langdurig moet worden bewaard, zij het nog steeds niet *'for the purpose of allowing an individual to perceive.'* Bovendien geldt bij deze exceptie de eis dat de tijdelijke reproductie geen zelfstandige economische waarde mag hebben, wat volgens *Infopaq II* aan de orde is wanneer de tijdelijke reproducties zélf worden geëxploiteerd.[17] De vraag is dan wat bij TDM de 'tijdelijke reproductie' is – de tekstanalyse, het daarvan afgeleide generatieve AI model zelf of nog iets anders?

Daarnaast is de driestappentoets van belang (art. 5, lid 5 van Richtlijn 2001/29/EG). Uitzonderingen, zoals die van artikel 13a Aw maar ook die voor TDM, moeten voldoen aan de drie daar genoemde eisen: een uitzondering mag 'slechts in bepaalde bijzondere gevallen worden toegepast mits daarbij geen afbreuk wordt gedaan aan de normale exploitatie van werken of ander materiaal en de wettige belangen van de rechthebbende niet onredelijk worden geschaad.' Nu steeds meer rechthebbenden zich realiseren dat hun gegevensverzamelingen goud waard zijn voor de bouwers van AI-systemen en AI-systemen met de exploitatie van hun werk kunnen concurreren, zal deze toets steeds vaker worden ingeroepen als tegenargument tegen de toepasselijkheid van een beperking op het auteursrecht, wat de rechtsonzekerheid voor TDM-uitvoerders juist weer vergroot.[18] Het is de vraag of dit wenselijk is. De strekking van de driestappentoets is het creëren van een balans tussen het auteursrecht en andere maatschappelijke belangen zoals innovatie.[19] Is het feit dat rechthebbenden geld zouden kunnen willen vragen voor TDM en er concurrentie van kunnen ondervinden voldoende om van 'wettige belangen' te spreken? Dit gaat wat ver: het moet gaan om een 'reële inkomstenbron' die wordt afgesneden, niet slechts de theoretische mogelijkheid dat geld te verdienen zou zijn wanneer de handeling toestemmingsplichtig zou zijn.[20]

Rechtmatige toegang tot het werk

Een beroep op artikel 15o Auteurswet vereist dat de verkrijger rechtmatige toegang tot het bronwerk had. Overweging 14 van de DSM-Richtlijn geeft hierbij aan dat '[o]nder rechtmatige toegang moet worden verstaan toegang tot content op basis van open-accessbeleid of via contractuele regelingen tussen rechthebbenden en onderzoeksorganisaties of instellingen voor cultureel erfgoed, zoals abonnementen, of op andere legale wijze.' Ook toegang tot content die vrijelijk online beschikbaar wordt gesteld levert 'rechtmatige toegang' op.

Wanneer sprake is van vrije online beschikbaarheid – dus zonder gebruiks- of licentievoorwaarden – zal meestal voldaan zijn aan dit vereiste. Maar geldt dat ook als de aanbieder van het werk handelt in strijd met gebruiks- of licentievoorwaarden? Wanneer zijn die rechtsgeldig aan hem opgelegd? Wat is de status van werk dat vrij toegankelijk op

13 M. Borghi & S. Karapapa, 'Non-display uses of copyright works: Google Books and beyond', *Queen Mary Journal of Intellectual Property*, 2011, 1(1), pp. 21-52. Zie ook overweging 9 van de DSM-Richtlijn: *'Tekst- en datamining kan ook worden verricht met betrekking tot zuivere feiten of gegevens die niet auteursrechtelijk zijn beschermd en in dergelijke gevallen is op grond van het auteursrecht geen toestemming vereist.'*

14 M. Senftleben, 'Compliance of national TDM rules with international copyright law: an overrated nonissue?', *IIC-International Review of Intellectual Property and Competition Law*, 2022, 53(10), pp. 1477-1505.

15 R. Ducato & A. Strowel, 'Ensuring Text and Data Mining: Remaining Issues With the EU Copyright Exceptions and Possible Ways Out', *CRIDES Working Paper Series* no. 1/2021; *European Intellectual Property Review*, 2021/5, p. 322-337.

16 Zie HvJ EU 16 juli 2009, C-5/08, ECLI:EU:C:2009:465, *NJ* 2011/288 (Infopaq I) en HvJ EU 17 januari 2012, C-302/10, ECLI:EU:C:2012:16, *AMI* 2012-2, nr. 7 (*Infopaq II*). Zie ook Hugenholtz *supra* noot 7.

17 In *Infopaq II* werden de uiteindelijke samenvattingen door mensen gemaakt en was 'tussen de partijen in het hoofdgeding in confesso dat het schrijven van een samenvatting op zich rechtmatig is en geen toestemming van de auteursrechthebbenden vereist'(punt 18). Dit roept de interessante vraag op wat de uitkomst zou zijn geweest of nu zou zijn als hier een AI-systeem voor was of wordt gebruikt.

18 D.J.G. Visser, 'Robotkunst en auteursrecht', *NJB* 2023/454.

19 C. Geiger, J. Griffiths & R. Hilty, 'Towards a Balanced Interpretation of the 'Three-Step Test' in Copyright Law', *EIPR*, 2008, vol. 4, pp. 489-496.

20 Vgl. P. B. Hugenholtz & R. Okediji, 'Conceiving an International Instrument on Limitations and Exceptions to Copyright', *Amsterdam Law School Legal Studies Research Paper* No. 2012-43, 6 maart 2012, p. 3.

internet staat, maar alleen achter een betaalmuur gepubliceerd had mogen worden. Veel content staat immers op tal van plaatsen op internet gratis en zonder gebruiksvoorwaarden online, zonder toestemming van de rechthebbenden. Denk aan de vele plaatjessites of het bij wetenschappers bekende Sci-Hub dat zich ten doel stelt wetenschappelijk werk te 'bevrijden' van de betaalmuren van uitgevers. Voor partijen die toegang willen tot deze werken is het al dan niet rechtmatige karakter niet altijd duidelijk.

Het is de vraag welk niveau van kennis of onderzoek hieromtrent gevergd mag worden van partijen die TDM wensen uit te voeren. Een vergelijking is mogelijk met de situatie van het GS *Media/Sanoma*-arrest.[21] Dit arrest bepaalt dat een commerciële partij een onderzoeksplicht heeft en vermoed moet worden kennis te hebben van het illegale karakter van materiaal achter een hyperlink. In een analoge uitleg van deze regel (het arrest betrof immers alleen het mededelingsrecht ten aanzien van hyperlinks) zou een commerciële partij – de doelgroep van artikel 150 – de bewijslast krijgen dat bronmateriaal rechtmatig aangeboden wordt.

Bronnen kunnen ook onder voorwaarden ontsloten worden. Die kunnen vermoedelijk niet eenzijdig worden opgelegd met een simpele disclaimer of popup,[22] maar bij een account of abonnement is dit uiteraard mogelijk. Een voorbeeld hiervan staat centraal in een Amerikaanse rechtszaak tussen Getty Images en de makers van afbeeldinggenerator Stable Diffusion.[23] Deze maker had via een betaald abonnement toegang tot Getty-afbeeldingen gekregen en daarop met behulp van TDM haar generator verder ontwikkeld. Getty's bezwaar komt erop neer dat alleen het duurste abonnement toestaat om afbeeldingen voor TDM-toepassingen te gebruiken, en dat daarmee het gebruik onder het gekozen – goedkoopste – abonnement niet rechtmatig zou zijn.

Vanuit het beginsel van contractsvrijheid (het gaat immers om twee zakelijke partijen) is goed verdedigbaar dat een contractuele voorwaarde genoeg kan zijn om TDM aan specifieke regels te onderwerpen of geheel te verbieden.[24] De DSM-Richtlijn biedt hier ook ruimte voor, gezien het feit dat artikel 7 juist artikel 4 (de basis van artikel 150) niet als van dwingend recht aanmerkt. Verdedigbaar is echter ook dat bij online ter beschikking gestelde werken uitsluitend een machinaal leesbaar voorbehoud gebruikt mag worden om TDM te verhinderen.[25] Een verbod in contractuele voorwaarden is dan niet bindend.

Een ander argument tegen contractuele voorwaarden om rechtmatige toegang te beperken kan worden ontleend aan het *Infopaq II*-arrest.[26] Daarin staat dat 'een gebruik als rechtmatig [wordt] beschouwd indien het door de betrokken rechthebbende is toegestaan of indien het door de toepasselijke regeling niet wordt beperkt.' De hier toepasselijke regeling legt geen beperking op aan de vorm van het gebruik, zodat een specifieke contractuele beperking in strijd zou zijn met de uitleg van deze term. Daar staat dan weer tegenover dat overweging 18 vermeldt dat de machinaal leesbare middelen ook 'de voorwaarden van een website of dienst' kunnen zijn. Omdat voorwaarden in gewonemensentaal worden opgesteld, roept dit de vraag op wat er dan 'machineleesbaar' moet zijn aan de voorwaarden. Over deze kwestie meer hieronder in de sectie 'Machine-leesbare middelen'.

Het auteursrechtelijk voorbehoud

De mogelijkheid zich bij TDM op artikel 150 te beroepen vervalt wanneer het auteursrecht 'uitdrukkelijk op passende wijze is voorbehouden, zoals door middel van machinaal leesbare middelen bij een online ter beschikking gesteld werk'. Dergelijke voorbehouden of opt-outs zijn zeldzaam in het auteursrecht: de hoofdregel is dat verveelvoudiging of openbaarmaking simpelweg niet mag, tenzij met toestemming of binnen de grenzen van een van de uitzonderingen.

Een auteursrechtelijke uitzondering met de mogelijkheid van voorbehoud komt op twee plaatsen in de Auteurswet voor. De persexceptie (artikel 15) maakt het overnemen van actueel nieuws uit de pers door de pers mogelijk, mits 'het auteursrecht niet uitdrukkelijk is voorbehouden' (lid 4). Dit voorbehoud kan eenvoudig worden gemaakt, zoals in het colofon van de bronuitgave of centraal op de website[27] en het gebeurt in de praktijk tegenwoordig bij alle nieuwsmedia. Deze mogelijkheid staat ook van meet af aan in de Berner Conventie.[28]

Bij overheidsuitgaven (artikel 15b) is verdere uitgave of verspreiding door derden toegestaan, tenzij het auteursrecht 'uitdrukkelijk is voorbehouden'. Dit voorbehoud moet 'blijkens mededeling op het werk zelf of bij de openbaarmaking daarvan' zijn gedaan of in een wet, besluit of verordening zijn vermeld. Voorbeelden zijn de regelingen voor euromunten[29] en de huisstijl van de politie.[30]

21 HvJ EU 8 september 2016, C-160/15, ECLI:EU:C:2016:644 (*GS Media BV/Sanoma Media Netherlands BV en anderen*).

22 Hof Den Haag 23 januari 2018, ECLI:NL:GHDHA:2018:61 (in cassatie in stand gebleven ECLI:NL:HR:2019:1445).

23 Z.Ü. Kahveci, 'Attribution problem of generative AI: a view from US copyright law', *Journal of Intellectual Property Law and Practice*, 2023, jpad076.Top of Form

24 De kwestie wanneer een overeenkomst wordt gesloten en in hoeverre een eenzijdig vermelde voorwaarde op een online kanaal bindend is, blijft hier buiten beschouwing. Zie over deze kwestie in ieder geval Hof Den Haag 23 januari 2018, ECLI:NL:GHDHA:2018:61.

25 P.B. Hugenholtz, 'The New Copyright Directive: Text and Data Mining (Articles 3 and 4)', *Kluwer Copyright Blog*, 2019. Beschikbaar via http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/.

26 HvJ EU 17 januari 2012, C-302/10, ECLI:EU:C:2012:16, *AMI* 2012-2, nr. 7 (*Infopaq II*).

27 D.J.G. Visser, commentaar op art. 15 Aw, in: *Tekst & Commentaar Intellectuele eigendom*, Deventer: Wolters Kluwer, 2017.

28 Aanvankelijk in artikel 7, inmiddels in artikel 10bis van de Berner Conventie.

29 Zie de 'Regelingen voorbehoud auteursrecht'(*Stcrt.* 1998, 232, gewijzigd *Stcrt.* 2008, 149), artikel 1. Dit gebeurt ook in aparte besluiten ten aanzien van allerlei euro-herdenkingsmunten. Zie bijvoorbeeld artikel 2 van het 'Besluit vaststelling bestanddelen beeldenaar munten van vijf en tien euro ter herdenking van Johan Cruijff', *Stcrt.* 2017, 57430.

30 *Stcrt.* 1993, 131, artikel 1.

Het voorbehoud voor TDM kent als formulering 'uitdrukkelijk op passende wijze is voorbehouden' en sluit daarmee taalkundig aan bij het voorbehoud van overheidsuitgaven, wat erop wijst dat enkel een zinsnede in een colofon of gebruiksvoorwaarden niet genoeg is. Dit blijkt ook uit de daaropvolgende verduidelijkende bijzin: 'zoals door middel van machinaal leesbare middelen bij een online ter beschikking gesteld werk.' Uiteraard laat dat onverlet dat de TDM-exceptie gebaseerd is op artikel 4 van de DSM-richtlijn, zodat een vergelijkbare formulering elders in het nationaal recht niet primair bepalend kan zijn bij de uitleg.
Een mogelijke verklaring voor de keuze voor het instrument van het voorbehoud zou kunnen zijn dat rechthebbenden die zelf actief hun eigen content via TDM willen exploiteren de mogelijkheid willen hebben dit exclusief te doen. Het voorbehoud kan echter door eenieder en zonder bijzondere reden worden gemaakt. Inmiddels lijkt het erop dat rechthebbenden massaal dit voorbehoud willen en zullen maken, omdat zij een vergoeding wensen en de concurrentie vrezen van de output van TDM in de vorm van generatieve AI. Onder die omstandigheden is een systeem met een voorbehoud een (te) ingewikkelde oplossing. Een eenvoudig verbod had tot dezelfde praktijksituatie geleid, zonder onduidelijkheid over hoe en waar een voorbehoud te vinden, interpreteren en toe te passen.
Uit de tekst van de AI Act zoals die eind december 2023 bekend is, blijkt echter dat de EU blijft inzetten op de TDM-exceptie met het bijbehorende voorbehoud ter regulering van deze materie.[31] Ook hier is echter geen duidelijkheid of richting gegeven over (on)wenselijkheid van inzet van het voorbehoud.

Machine-leesbare middelen

'Het voorbehouden van die rechten [moet] enkel als passend worden beschouwd indien hierbij machinaal leesbare middelen worden gebruikt', zo bepaalt overweging 18 van de Richtlijn. Deze eis van machineleesbaarheid is begrijpelijk als men bedenkt dat TDM een volstrekt geautomatiseerd proces is, wat impliceert dat ook het voorbehoud geautomatiseerd getoetst of herkend moet kunnen worden. Het bekendste precedent hierbij is het protocol 'robots.txt', dat sinds de opkomst van zoekmachines gebruikt wordt om aan te geven welke pagina's of categorieën niet door robots of webspiders (ook wel crawlers) mogen worden overgenomen voor gebruik in zoekresultaten.[32] Dit protocol, ontstaan uit zelfregulering in de internetsector, is wijd en zijd bekend en wordt door vrijwel alle zoekmachines ondersteund.
Een forse beperking aan het gebruik van robots.txt als voorbehoud voor TDM is dat men slechts twee keuzes heeft: óf enkel de bij naam genoemde robots buitensluiten, óf alle robots, dus ook die van traditionele zoekmachines. Dat laatste zal niemand willen, maar het uitsluiten op naam is te beperkend: daarmee worden toekomstige, nu nog onbekende robots immers niet tegengehouden. Gezien de achtergrond van robots.txt (het buitensluiten van bij naam bekende webspiders die niet goed met de site kunnen omgaan en het voor alle spiders afschermen van niet voor zoekresultaten bestemde pagina's) is deze beperking begrijpelijk, maar zij maakt dit protocol als oplossing voor TDM weinig nuttig. Het is onredelijk te verlangen dat een uitgever die geen TDM-hergebruik wil, zichzelf tevens moet uitsluiten van zoekmachines zoals Google of Bing. Overigens kondigde Google in 2023 'Google-Extended' aan, waarmee de optie geboden wordt specifieke AI-applicaties wel, maar zoekmachines niet uit te sluiten.[33] Dit geldt dan wel enkel voor de Google-webcrawler, en niet voor overige generatieve AI-dataverzamelaars.
De brede bekendheid van robots.txt maakt dat velen toch proberen dit protocol in te zetten als machinaal leesbaar middel voor het TDM-voorbehoud. Enkele uitgevers (zoals *Die Zeit, Le Monde* en *The Guardian*) hebben in hun robots.txt bijvoorbeeld teksten opgenomen die een voorbehoud voor data mining uitdrukken. Zij doen dit technisch gezien in de vorm van een commentaar (een regel die begint met het '#'-symbool), een tekst die naar zijn aard nu net *niet* bedoeld is om door machines gelezen te worden.[34] De juridische houdbaarheid hiervan lijkt dan ook niet sterk. Het machinaalleesbaarheidsaspect van robots.txt en andere documenten (zoals de Creative Commons auteursrechtlicenties) zit hem erin dat vooraf afspraken zijn gemaakt over vorm en betekenis van de mogelijke symbolen in zo'n document. Daarmee zijn die symbolen geautomatiseerd te lezen en te verwerken. De te verwachten informatie is dan afgebakend en bekend is welk symbool welke betekenis moet krijgen. Zonder dergelijke afspraken zijn de symbolen in zo'n document betekenisloos voor een computer die ze moet lezen. Standaardisatie – het vastleggen in afspraken – is bij dergelijke protocollen dan ook een absoluut vereiste.
Men zou vandaag de dag kunnen stellen dat de opkomst van AI-taalmodellen het mogelijk moet maken dat documenten in gewoon Nederlands ook door machines gelezen worden. Experimenten bij het automatisch lezen van websitevoorwaarden en privacyverklaringen ondersteunen deze gedachte.[35] Echter, hoewel vele websites op enigerlei wijze een verbod opnemen betreffende TDM, loopt de

31 'Providers of general-purpose AI models shall: [...] put in place a policy to respect Union copyright law in particular to identify and respect, including through state of the art technologies where applicable, the reservations of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790' (AI Act bepaling uit tekst gepubliceerd op copyrightblog.kluweriplaw.com op 11 december 2023 door Paul Keller).

32 M. Boonk, 'No robots clauses: zijn ze effectief? Een analyse van robots.txt files en no robots clauses metatags', *Computerrecht* 2009/2.

33 D. Zhang e.a. 'Tag Your Fish in the Broken Net: A Responsible Web Framework for Protecting Online Privacy and Copyright', *arXiv preprint arXiv:2310.07915*, 2023.

34 D. Crocker (red.), 'Augmented BNF for Syntax Specifications', *Internet rfc* 5234, januari 2008, sectie 3.9.

35 R. Ducato & A. Strowel, 'Limitations to text and data mining and consumer empowerment: making the case for a right to 'machine legibility'', *IIC-International Review of Intellectual Property and Competition Law* 50(6), 2019, pp. 649-684.

gebruikte terminologie zeer uiteen en ontbreken definities.[36] Geautomatiseerde interpretatie van zulke teksten is dus foutgevoelig, wat op gespannen voet staat met de strekking van machine-leesbaar: direct te herkennen en eenduidig op te volgen. Uiteindelijk zal het Hof van Justitie dan ook over de scope van deze term en het benodigde niveau van standaardisatie moeten oordelen.[37]

Het eenvoudigst zou natuurlijk zijn als er een nieuw protocol zou komen dat door alle betrokkenen gedragen zou worden. De precedenten zijn echter weinig geruststellend. Eind jaren nul werd het fijnmazige Automated Content Access Protocol geïntroduceerd door uitgevers als alternatief voor robots.txt. Het werd door vrijwel niemand overgenomen: internetstandaarden komen er alleen als er een breed gedragen consensus is, en die ontbrak simpelweg.[38] Er zijn diverse nieuwe standaarden voorgesteld, zoals 'ai.txt' met specifieke uitsluitingsmogelijkheden voor AI-gerichte TDM[39], maar deze leiden een kwijnend bestaan. Een recent voorbeeld is het in december 2022 in W3C-verband voorgestelde TDM Reservation Protocol.[40] Het protocol heeft naar verluidt (eind september 2023) enige tractie gekregen onder Europese uitgevers,[41] maar geen enkele TDM-uitvoerder heeft meegedaan aan de discussie of enige interesse uitgesproken in het ondersteunen van dit protocol.

Andere voorstellen gaan uit van een voorbehoud gekoppeld aan het bronbestand zelf, zoals de foto of tekst die in een TDM proces wordt opgehaald. Een simpel voorbeeld is het 'NoImageAI' label dat afbeeldingsdienst DeviantArt promoot. Bij deze dienst kan eenieder zelfgemaakte afbeeldingen publiceren en voorzien van allerlei labels (tags), waarmee afbeeldingen gericht gezocht kunnen worden. Het toevoegen van dit label bij de publicatie zou dan het voorbehoud opleveren. Echter, wie het werk downloadt, krijgt de labels er niet bij. Bij een herpublicatie van het werk ontbreekt deze dan, zodat de TDM-uitvoerder vrijelijk gebruik zou kunnen maken van deze kopie.

Een variant hierop zijn de 'content credentials' van softwareuitgeverij Adobe.[42] Hierbij wordt metadata, maar ook eventuele voorbehouden, op een machineleesbare en gestandaardiseerde manier opgeslagen, waarna robots gegeven een bronbestand deze geautomatiseerd erbij kunnen vinden. Dit zou ook moeten werken bij aanpassing of kopiëren van het bronbestand. Dit veronderstelt natuurlijk wel dat iedere uitgever voor ieder van zijn beschermde werken een dergelijke credential aanmaakt. Ook hiervan is nog vrij weinig te merken in de markt.

Dit gebrek aan adoptie van een werkbaar protocol is frustrerend voor rechthebbenden. De wet gaat immers uit van de mogelijkheid van een door de uitgever en dus eenzijdig te maken *voorbehoud*. Daarmee lijkt toch lastig te rijmen dat TDM-uitvoerenden toepassing van die exceptie zouden kunnen frustreren door niet mee te werken aan standaarden voor een TDM-voorbehoud. Dat zou wel de consequentie zijn als een voorbehoud pas geldig is wanneer het voldoet aan een standaard waarover consensus bij zowel uitgevers als TDM-uitvoerders bestaat. Daarmee zouden die TDM-uitvoerders en AI-diensten die vanwege hun datahonger er weinig belang bij hebben standaarden voor het vastleggen van voorbehouden vrijwillig te volgen, zich aan het voorbehoud kunnen onttrekken. Tegelijkertijd hebben ook marktpartijen behoefte aan rust en duidelijkheid in de markt en is het dus zeker niet ondenkbaar dat zij meer initiatieven zullen ontplooien op dit gebied. De recent voorgestelde AI Act bevat hierbij een nadere prikkel: deze zal vereisen dat aanbieders van algemene AI modellen beleid moeten maken en publiceren omtrent het respecteren van machineleesbare voorbehouden.

Er zijn nu al enige bewegingen vanuit de markt merkbaar. Zo gaf marktleider OpenAI recent aan robots.txt te zullen respecteren, gevolgd door concurrenten Microsoft/Bing en Google.[43] Een mogelijke verklaring voor deze draai is dat deze grote gevestigde diensten een comfortabele uitgangspositie hebben met de reeds verzamelde data, en de financiële middelen hebben om licenties te kopen. Hun recent gestarte concurrenten hebben dit alles niet, en zullen alomtegenwoordige voorbehouden dan ook als een zware hindernis ervaren. Dit roept vragen op over het effect van zo'n standaard op innovatieve diensten. Wellicht is een regeling nodig zoals gekozen bij artikel 17 van de DSM-Richtlijn, waarbij uiteindelijk expliciet verschillende auteursrechtelijke regimes voor grote en kleine aanbieders zijn opgenomen.[44]

36 T. Margoni & L. Schirru, 'Arts 3 and 4 of the CDSM Directive as regulatory interfaces: Shaping contractual practices in the Commercial Scientific Publishing and Stock Images sectors', *Kluwer Copyright Blog*, 2023. Beschikbaar via https://copyrightblog.kluweriplaw.com/2023/08/22/arts-3-and-4-of-the-cdsm-directive-as-regulatory-interfaces-shaping-contractual-practices-in-the-commercial-scientific-publishing-and-stock-images-sectors/.

37 S. Havlikova, 'Web Scraping and Text and Data Mining Exception: Could the CDSM Directive, Designed to Support the Reuse of Publicly Available Data, Have Had the Opposite Effect?', SSRN (2023). Beschikbaar via https://ssrn.com/abstract=4605551.

38 D. Ippolito & Y. W. Yu, 'DONOTTRAIN: A Metadata Standard for Indicating Consent for Machine Learning', in: Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA, PMLR 202, 2023.

39 'What is ai.txt?', https://site.spawning.ai/spawning-ai-txt.

40 L. le Meur, 'TDM Reservation Protocol (TDMRep)', Final Community Group Report 2022, beschikbaar via https://www.w3.org/2022/tdmrep/.

41 Paul Keller & Zuzanne Warso, Defining best Practices for Opting Out of ML Training, Open Future Policy Brief # 5, https://openfuture.eu/wp-content/uploads/2023/09/Best-_practices_for_optout_ML_training.pdf.

42 Adobe, 'What are Content Credentials', 9 juni 2023. Beschikbaar via https://helpx.adobe.com/creative-cloud/help/content-credentials.html.

43 M. Hendrikman, 'OpenAI gaat verbod op crawling in robots.txt respecteren', Tweakers.net, 8 augustus 2023.

44 T. Spoerri, 'On upload-filters and other competitive advantages for Big Tech Companies under Article 17 of the directive on copyright in the digital single market', *J. Intell. Prop. Info. Tech. & Elec. Com. L.*, 2019, 10, p. 173.

De grote vraag bij zo'n nieuwe standaard is dan ook hoe af te dwingen dat marktpartijen zich hieraan conformeren. Alle bestaande standaardisatieprocessen gaan uit van vrijwillige deelname (zoals de vele internetstandaarden) of van wettelijke dwang (zoals bij wettelijk verplichte NEN normen). Het zou uniek zijn als een groep uitgevers een standaard voor een TDM-voorbehoud ontwikkelt en bij de rechter afdwingt dat het niet navolgen daarvan inbreuk op het auteursrecht oplevert.

Conclusie

De Europese wetgever heeft geprobeerd duidelijkheid te scheppen voor tekst- en datamining (TDM): enerzijds is TDM toegestaan, anderzijds kan een voorbehoud gelden. Maar hoe dat voorbehoud eruit moet zien en wanneer men daar rekening mee moet houden, is vooralsnog niet uitgekristalliseerd. Zo is er het probleem dat beschermde content op tal van plaatsen te vinden is, ook zonder toestemming van de rechthebbenden, die daar dan ook geen voorbehoud bij hebben kunnen plaatsen.

Daarnaast is ons inziens onvoldoende nagedacht over de vorm en uitvoerbaarheid van de opt-out variant, met name bij vrijelijk toegankelijke content waar de besproken machinale leesbaarheid van de opt-out vereist is. De ideale situatie waarin er een door zowel rechthebbenden als gebruikers breed gedragen protocol of standaard bestaat voor het wettelijk beoogde machine leesbare voorbehoud bestaat (nog?) niet, maar het lijkt toch ook moeilijk denkbaar dat de wetgever het mogelijk heeft willen maken voor gebruikers om zich eenvoudig 'blind' te houden voor machine leesbare voorbehouden omdat *hun* machines die voorbehouden niet zouden kunnen begrijpen. Een regeling die rechthebbenden een opt-out pretendeert te geven, moet die pretentie ook waar kunnen maken en niet afhankelijk zijn van de bereidheid van AI aanbieders om in de toekomst bepaalde voorbehouden te zullen respecteren. De AI Act is hierbij misschien de doorslaggevende factor.

Dit 'blind houden' is dan weer begrijpelijk vanuit de angst van dataverzamelaars dat uitgevers simpelweg op al hun werken voorbehouden zullen plaatsen, zodat van de TDM-uitzondering niets meer overblijft en zij zich gedwongen zullen zien voor iedere dataverzameling licentieovereenkomsten te sluiten met navenante vergoedingen. Dit raakt met name de mkb-ondernemingen en startups die als motor voor innovatie gezien worden, maar niet de budgetten of juridische ondersteuning hebben om dit te doen. Het meest wenselijk is dan ook dat er op korte termijn een standaard komt voor het interpreteren van voorbehouden; als de markt dit niet zelf regelt, dan bijvoorbeeld als *guidance* vanuit de Europese Commissie. Anders zal de discussie over TDM, inbreuk en voorbehouden zich verplaatsen naar de rechtszaal, met alle risico's van uiteenlopende interpretaties en elkaar tegensprekende vonnissen van dien.

Er is geen twijfel over dat door de opkomst van generatieve AI het belang van de TDM-exceptie en het bijbehorende voorbehoud veel groter is geworden dan de wetgever in 2018 voor ogen stond. Er lijkt geen reden te zijn om als rechthebbende terughoudend te zijn met een voorbehoud. En de wetgever lijkt nu met de AI Act vol in te zetten op de mogelijkheid TDM voor generatieve AI met behulp van het voorbehoud tegen te gaan. Het is de vraag of dit wenselijk is. De te verwachten uitkomst zal – nadat er gestandaardiseerd is ten aanzien van de machine-leesbaarheid – waarschijnlijk zijn dat het voorbehoud massaal zal worden gemaakt en de EU TDM-exceptie daarmee effectief wordt uitgeschakeld. Het blijft afwachten wat dat in de praktijk gaat betekenen.