Tekst **Maarten Zeinstra** | coördinator auteursrecht bij KVAN

Tekst- en datamining: tussen wet en praktijk

De poortwachter van data

Meestal zeg ik dat erfgoedinstellingen hoeders van erfgoed zijn en geen poortwachters. Maar we moeten ons verhouden tot de realiteit: AI-organisaties 'minen' onze digitale collecties op grote schaal. Dit minen ligt maatschappelijk en juridisch gevoelig. Hierdoor worden we onvermijdelijk poortwachters van onze data. De technologie om dit te reguleren heeft een technische en juridische basis; achter een simpel 'stopbord' schuilt een wereld aan nuances.

Het 'minen' van digitaal materiaal op het open internet is voor AI-training de standaard. AI-bots, ook wel *crawlers* of robots, bezoeken in groten getale websites van erfgoedinstellingen. Bij een bezoek aan een website wordt er altijd een kopie van de pagina en links van media naar de bezoeker gestuurd. Deze bezoeker maakt lokaal een kopie van die tekst en downloadt de media om deze vervolgens op het scherm te presenteren. Zo werken browsers. Crawlers doen niet veel anders. Sommige zijn alleen geïnteresseerd in tekst en zullen geen media downloaden, anderen willen alles bekijken. Ook zij maken lokale kopieën van de inhoud van webpagina's. In tegenstelling tot menselijke bezoekers vragen robots soms vele pagina's en mediabestanden per seconde op. Het is soms

> **Voor de robot is het vrijwel onmogelijk om het onderscheid tussen deze twee blokkades te maken**

alsof een jaar aan bezoekers opeens op één dag bij je digitale portaal staat. Dit kan technische en juridische problemen opleveren. Auteursrechthebbenden van objecten in erfgoedcollecties mogen dit soort AI-training verbieden. En als minen door de rechthebbende wordt verboden, dan moeten de instellingen er ook voor zorgen dat deze robots een dergelijk verbod kunnen lezen, dat bepaalde bestanden, webpagina's of teksten niet gemined mogen worden. Hiervoor gebruiken we vaak de robots.txt-standaard.

Blokkeren in twee smaken

Het verbieden van deze robots om bepaalde werken te kopiëren wordt ook wel een tekst- en dataminingvoorbehoud genoemd. En van deze blokkades bestaan er twee smaken: om je infrastructuur te beschermen en om te voldoen aan de wens van de rechthebbende.

Een erfgoedinstelling mag altijd beperkingen opleggen voor wie op haar servers komt. Bijvoorbeeld door bepaalde pagina's achter een login te plaatsen, door een andere website terug te geven voor mobiele gebruikers dan voor desktopgebruikers, door in te stellen dat iedere gebruiker maar één keer per seconde een pagina mag opvragen, et cetera. Er bestaan ook technieken die aan robots aangeven dat ze specifieke pagina's niet mogen bezoeken. Ze kunnen daarmee aan deze crawlers communiceren dat ze bepaalde bronnen niet mogen opvragen.

Dat noem ik een materieel voorbehoud. Je verbiedt de robot om bij bepaalde materie te komen. Dit kun je ook in de algemene voorwaarden opnemen. Dat is niet gebaseerd op het auteursrecht, maar op het contractrecht. De instelling treedt hier op als eigenaar van de infrastructuur. Het doel is pragmatisch: de stabiliteit van de systemen waarborgen.

De publieksdienstverlening mag immers niet bezwijken onder het geweld van duizenden robots die tegelijkertijd hoge resolutiescans ophalen.

Immaterieel voorbehoud

Er is ook een tweede verbod mogelijk. Dat noem ik een immaterieel voorbehoud. Je communiceert hier dan niet mee dat het fysieke bestand niet gekopieerd mag worden, maar geeft aan dat het auteursrechtelijke werk niet gemined mag worden. En dat zijn dan alle kopieën van dat werk. Dit volgt rechtstreeks uit de Auteurswet (artikel 15n) en richt zich op het intellectuele werk zelf (in vakjargon ook wel het *corpus mysticum*). Een tekst- en dataminingvoorbehoud op bijvoorbeeld een artikel uit een krant geldt dan op zowel het fysieke artikel als elke (digitale) kopie die beschikbaar wordt gesteld. Een immaterieel voorbehoud heeft dan ook een veel grotere impact.

Dit verbod mag dan ook alleen door de rechthebbende opgelegd worden. Het gaat dan om de wens van de rechthebbende – de fotograaf, schrijver of hun nazaten of rechtverkrijgenden – die niet willen dat hun creatieve arbeid zonder toestemming wordt gebruikt voor het trainen van commerciële AI-modellen. En let op: het zijn alleen de rechthebbenden die deze verregaande maatregel mogen inzetten. Heb je dit verzoek niet gekregen van de rechthebbende, dan is deze ook niet toepasbaar.

Dit laatste verbod is trouwens niet van toepassing op onderzoeksorganisaties en culturele erfgoedinstellingen. Deze organisaties hoeven zich niets aan te trekken van het juridische stopbord van het immaterieel voorbehoud.

Eén stopbord, twee betekenissen

Hoewel de juridische grondslag voor deze twee voorbehouden totaal verschillend is, maken ze in de praktijk gebruik van exact dezelfde technologie en technieken: robots.txt, TDM Reservation Protocol (TDMRep) en algemene voorwaarden. Het signaal naar de buitenwereld is hetzelfde: ‘Toegang geweigerd voor TDM’.

Voor de robot is het vrijwel onmogelijk om het onderscheid tussen deze twee blokkades te maken. Zij zien een stopbord, maar weten niet of dit er staat om een server tegen overbelasting te beschermen of om de eisen van een specifieke rechthebbende uit te voeren. Tegelijkertijd zetten we de onderzoeksorganisaties en culturele erfgoedinstellingen op een eigen pad. Met deze voorbehouden worden de gereedschappen (de bots) verboden, en niet het type gebruik (commercieel of wetenschappelijk onderzoek).

De overheid als bijzondere rechthebbende

Een specifiek dilemma doet zich voor bij overheidsinformatie. Hoewel de overheid in veel gevallen de auteursrechthebbende is van haar eigen publicaties en rapporten, heeft zij niet dezelfde vrijheid als een private partij om een immaterieel voorbehoud te plaatsen.

Onder de Wet hergebruik overheidsinformatie (Who) en de Wet open overheid (Woo) is het uitgangspunt dat overheidsinformatie zo drempelloos mogelijk hergebruikt moet kunnen worden, zonder discriminerende voorwaarden. We mogen

Als sector moeten we transparant zijn over de keuzes die we maken

ChatGPT niet anders behandelen dan Openarchieven.nl, het genealogisch platform van Bob Coret. Een TDM-opt-out vanuit een immaterieel voorbehoud door een overheidsinstelling moet daarom zwaar worden gerechtvaardigd door een doel van algemeen belang, zoals de continuïteit van de publieke taak. In de praktijk betekent dit dat overheden vaak geen immaterieel voorbehoud kunnen maken op hun eigen data, tenzij er heel specifieke redenen zijn. Een materieel voorbehoud (ter bescherming van de server) blijft echter altijd een legitieme route. Let er dan wel op dat alle bots geblokkeerd moeten worden.

Als sector moeten we transparant zijn over de keuzes die we maken. Door op onze websites helder te communiceren waarom we bepaalde beperkingen opleggen, helpen we niet alleen de rechthebbenden, maar bieden we ook duidelijkheid aan de gebruikers van onze data. Het technische stopbord is slechts het begin: het is aan ons om de juridische context daarachter uit te leggen. |

Dit artikel is gebaseerd op het KVAN-kennisdocument ‘Tekst- en Datamining in de Auteurswet’ (april 2026). De volledige publicatie is te vinden op de website van KVAN.