TDM en de *AI-Verordening*

Prof. mr. P.B. Hugenholtz

Studiemiddag

Vereniging voor Auteursrecht

Amsterdam, 18 oktober 2024

DSM Richtlijn (2019/790): definitie van TDM

Art. 2(2):**tekst- en datamining**: “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”

TDM-beperkingen in DSM-richtlijn

[voor auteursrecht, naburige rechten en databankenrecht]

- Art. 3 DSM: beperking voor reproduceren i.h.k.v. TDM voor non-profit wetenschappelijk onderzoek [= art. 15n Aw]
- Art. 4 DSM: beperking voor reproduceren i.h.k.v. TDM voor algemene doeleinden [= art. 15o Aw]
  - Mits werken "rechtmatig toegankelijk" zijn (bijv. boeken, web)
  - Tenzij voorbehoud door rechthebbende [opt-out]
    - Voor inhoud die online beschikbaar wordt gesteld "door middel van *machinaal leesbare middelen"* (zie O.18)

artikel 5, lid 1, van Richtlijn 2001/29/EG. Om in deze situaties meer rechtszekerheid te bieden en ook in de private sector innovatie aan te moedigen, moet deze richtlijn voorzien in een uitzondering of beperking onder bepaalde voorwaarden voor reproducties en opvragingen van werken of andere materialen ten behoeve van tekst- en datamining, en in de mogelijkheid om de gemaakte kopieën zo lang te bewaren als nodig is ten behoeve van die tekst- en datamining.

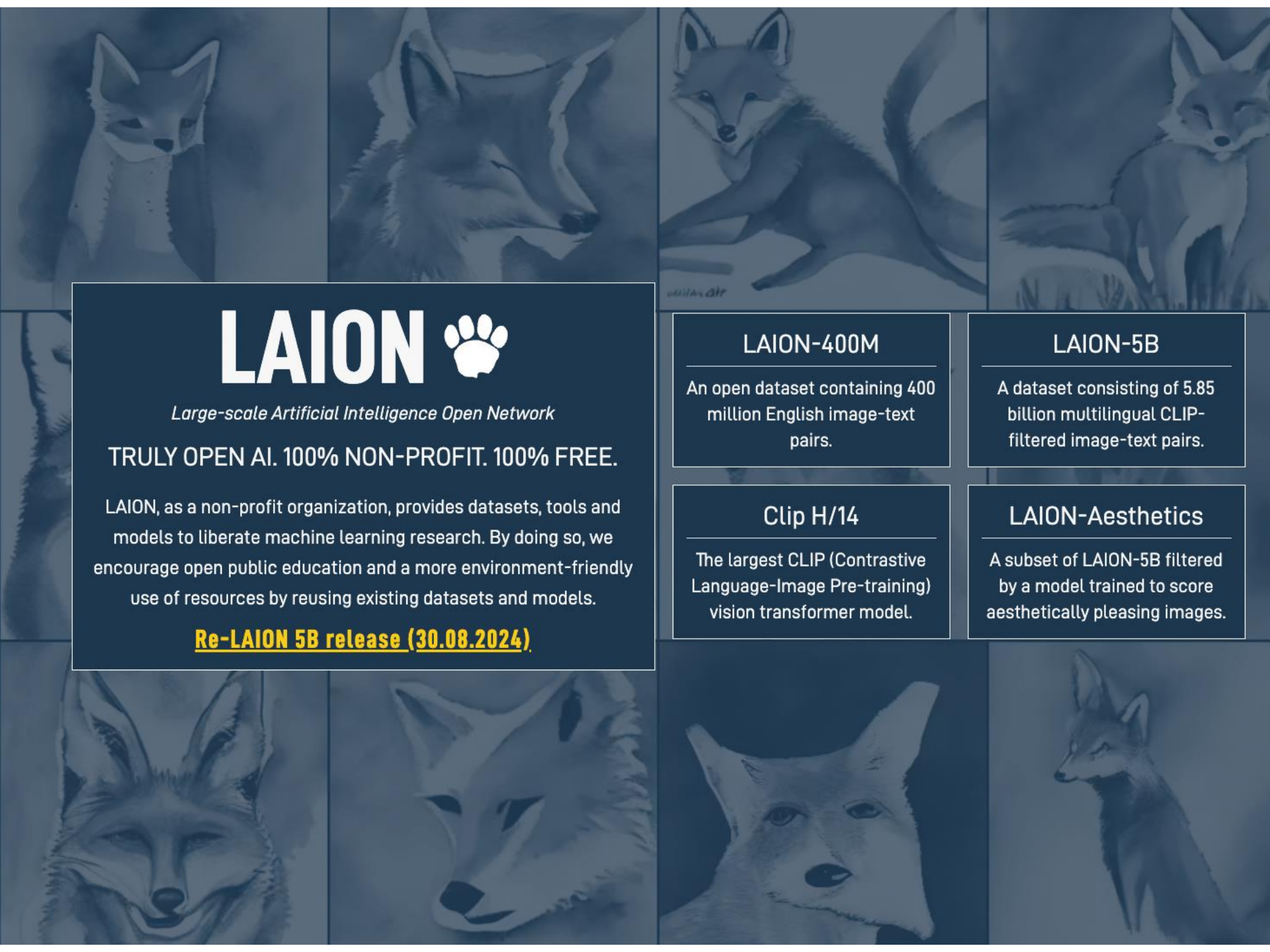
Deze uitzondering of beperking moet alleen gelden wanneer de begunstigde rechtmatig toegang heeft verkregen tot het werk of ander materiaal, onder meer wanneer het voor het publiek online beschikbaar is gesteld, en voor zover de rechthebbenden de rechten om reproducties te maken en opvragingen te verrichten ten behoeve van tekst- en datamining niet op passende wijze hebben voorbehouden. Bij content die online voor het publiek beschikbaar is gesteld, moet het voorbehouden van die rechten enkel als passend worden beschouwd indien hierbij machinaal leesbare middelen worden gebruikt, waaronder metagegevens en de voorwaarden van een website of een dienst. Het voorbehouden van rechten ten behoeve van tekst- en datamining mag geen betrekking hebben op andere vormen van gebruik. In andere gevallen kan het passend zijn om rechten voor te behouden met behulp van andere middelen, zoals contractuele overeenkomsten of een eenzijdige verklaring. Rechthebbenden moeten maatregelen kunnen nemen om te waarborgen dat indien zij de rechten hebben voorbehouden, dit ook wordt nageleefd. Deze uitzondering of beperking mag geen afbreuk doen aan de verplichte uitzondering voor tekst- en datamining voor wetenschappelijk onderzoek als vastgesteld in deze richtlijn, noch aan de bestaande uitzondering voor tijdelijke reproductiehandelingen als vastgesteld in artikel 5, lid 1, van Richtlijn 2001/29/EG.

Artikel 5, lid 3, onder a), van Richtlijn 2001/29/EG staat de lidstaten toe om uitzonderingen te maken of beperkingen te stellen op [Table of contents] tie, mededeling aan het publiek en

egal-content/NL/TXT/?uri=CELEX%3A32019L0790#


CHATGPT
OpenAI

Machine learning



LAION

Large-scale Artificial Intelligence Open Network

TRULY OPEN AI. 100% NON-PROFIT. 100% FREE.

LAION, as a non-profit organization, provides datasets, tools and models to liberate machine learning research. By doing so, we encourage open public education and a more environment-friendly use of resources by reusing existing datasets and models.

Re-LAION 5B release (30.08.2024)

LAION-400M

An open dataset containing 400 million English image-text pairs.

LAION-5B

A dataset consisting of 5.85 billion multilingual CLIP-filtered image-text pairs.

Clip H/14

The largest CLIP (Contrastive Language-Image Pre-training) vision transformer model.

LAION-Aesthetics

A subset of LAION-5B filtered by a model trained to score aesthetically pleasing images.

LG Hamburg 27-9-24 Kneschke/LAION

- Geen tijdelijke kopieen (art. 5 lid 1 InfoSoc)
- Geen inbreuk, want gedekt door art. 3 DSM:
  - Non-profit onderzoek
  - TDM omvat correlatie bestanden & metadata
- Obiter dictum:
  - TDM voor AI training valt wsch. onder art. 4 DSM
  - Scraping-verbod in gebruiksvoorwaarden op website van fotoagentschap is mogelijk geldige opt-out
    - Tekstje op website is “machine-leesbaar” (!)

TDM Reservation Protocol (TDMRep)

Final Community Group Report 02 February 2024

This version:
https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240202/

Editor:
Laurent Le Meur (EDRLab)

Previous version
TDMRep Initial version

Useful links
TDMRep Documents
TDMRep CG Home Page

Feedback:
GitHub w3c/tdm-reservation-protocol (pull requests, new issue, open issues)
public-tdmrep@w3.org with subject line `[tdmrep] … message topic …` (archives)

Copyright © 2021-2024 the Contributors to the TDM Reservation Protocol (TDMRep) Specification, published by the Text and Data Mining Reservation Protocol Community Group under the W3C Community Final Specification Agreement (FSA). A human-readable summary is available.

Abstract

This specification defines a simple and practical Web protocol, capable of expressing the reservation of rights relative to text & data mining (TDM) applied to lawfully accessible Web content, and to ease the discovery of TDM licensing policies associated with such content.

This initiative is a technical answer to the constraints set by the Article 4 of the new European Directive on

Publicatieblad
van de Europese Unie

NL
L-serie

2024/1689

12.7.2024

VERORDENING (EU) 2024/1689 VAN HET EUROPEES PARLEMENT EN DE RAAD

van 13 juni 2024

tot vaststelling van geharmoniseerde regels betreffende artificiële intelligentie en tot wijziging van de Verordeningen (EG) nr. 300/2008, (EU) nr. 167/2013, (EU) nr. 168/2013, (EU) 2018/858, (EU) 2018/1139 en (EU) 2019/2144, en de Richtlijnen 2014/90/EU, (EU) 2016/797 en (EU) 2020/1828 (verordening artificiële intelligentie)

(Voor de EER relevante tekst)

HET EUROPEES PARLEMENT EN DE RAAD VAN DE EUROPESE UNIE,

Gezien het Verdrag betreffende de werking van de Europese Unie, en met name de artikelen 16 en 114,

Gezien het voorstel van de Europese Commissie,

Na toezending van het ontwerp van wetgevingshandeling aan de nationale parlementen,

Gezien het advies van het Europees Economisch en Sociaal Comité ([1]),

Gezien het advies van de Europese Centrale Bank ([2]),

Gezien het advies van het Comité van de Regio's ([3]),

Handelend volgens de gewone wetgevingsprocedure ([4]),

Overwegende hetgeen volgt:

(1) Deze verordening heeft ten doel de werking van de interne markt te verbeteren door een uniform rechtskader vast te stellen, met name voor de ontwikkeling, het in de handel brengen, het in gebruik stellen, en het gebruik van artificiële-intelligentiesystemen (AI-systemen) in de Unie, in overeenstemming met de waarden van de Unie, de introductie van mensgerichte en betrouwbare artificiële intelligentie (AI) te bevorderen en te zorgen voor een hoge mate van bescherming van de gezondheid, de veiligheid en de grondrechten zoals vastgelegd in het Handvest van de grondrechten van de Europese Unie (het "Handvest"), met inbegrip van de democratie, de rechtsstaat en de bescherming van het milieu, te beschermen tegen de schadelijke effecten van AI-systemen in de Unie, alsook innovatie te ondersteunen. Deze verordening waarborgt het vrije verkeer van op AI gebaseerde goederen en diensten over de grenzen heen, zodat de lidstaten geen beperkingen kunnen opleggen aan de ontwikkeling, het in de handel brengen en het gebruik van AI-systemen, tenzij dat door deze verordening uitdrukkelijk wordt toegestaan.

Auteursrechtbepalingen in AI-Verordening (2024/1698)

- Aanbieders van GPAI-modellen moeten:
  - “een beleid opstellen ter naleving van het Unierecht inzake auteursrechten en naburige rechten en dan met name ter vaststelling en naleving, onder meer door middel van geavanceerde technologieën, van een op grond van artikel 4, lid 3, van Richtlijn (EU) 2019/790 tot uitdrukking gebracht voorbehoud van rechten.” (art. 53 lid 1 (c) AIV)

beperkingen ingevoerd die onder bepaalde voorwaarden toestaan dat werken of andere materialen voor doeleinden van tekst- en datamining worden gereproduceerd of geëxtraheerd. Op grond van deze regels kunnen rechthebbenden ervoor kiezen hun rechten op hun werken of andere materialen voor te behouden teneinde tekst- en datamining te voorkomen, tenzij dit gebeurt voor doeleinden van wetenschappelijk onderzoek. Indien opt-outrechten uitdrukkelijk en op passende wijze zijn voorbehouden, dient een aanbieder van AI-modellen voor algemene doeleinden indien hij de werken voor tekst- en datamining wil gebruiken, toestemming aan de rechthebbenden te vragen.

(106) Aanbieders die AI-modellen voor algemene doeleinden in de Unie in de handel brengen, zijn gehouden de naleving van de desbetreffende verplichtingen in deze verordening te waarborgen. Zij moeten daartoe een beleid invoeren ter naleving van het Unierecht inzake auteursrechten en naburige rechten, met name om kennis te hebben van het door rechthebbenden geuite voorbehoud van rechten op grond van artikel 4, lid 3, van Richtlijn (EU) 2019/790 en dit na te leven. Aanbieders die in de Unie een AI-model voor algemene doeleinden in de handel brengen, zijn hiertoe verplicht, ongeacht de jurisdictie waarin de auteursrechtelijke relevante handelingen plaatsvinden die helpen die AI-modellen voor algemene doeleinden te trainen. Alleen op deze manier kan gezorgd worden voor een gelijk speelveld voor aanbieders van AI-modellen voor algemene doeleinden, waar geen enkele aanbieder zich een concurrentievoordeel op de Uniemarkt mag kunnen verschaffen met lagere auteursrechtelijke normen dan de in de Unie toepasselijke normen.

(107) Voor een grotere transparantie ten aanzien van de bij de pre-training en training van AI-modellen voor algemene doeleinden gebruikte data, met inbegrip van door het auteursrecht beschermde tekst en data, is het passend dat aanbieders van dergelijke modellen een voldoende gedetailleerde samenvatting maken en publiceren van de voor de training van het AI-model voor algemene doeleinden gebruikte content. Terdege rekening houdend met de noodzaak tot bescherming van bedrijfsgeheimen en vertrouwelijke bedrijfsinformatie moet deze samenvatting breed van karakter zijn in plaats van technisch gedetailleerd, teneinde partijen met legitieme belangen, waaronder houders van auteursrechten, in staat te stellen hun rechten uit hoofde van het Unierecht uit te

Auteursrechtbepalingen in AI-Verordening (2024/1698)

- Aanbieders van GPAI-modellen moeten:
  - “een voldoende gedetailleerde samenvatting opstellen en openbaar maken over de voor het trainen van het AI-model voor algemene doeleinden gebruikte content, volgens een door het AI-bureau verstrekt sjabloon.” (art. 53 lid 1 (d) AIV)

die in de Unie een AI-model voor algemene doeleinden in de handel brengen, zijn hiertoe verplicht, ongeacht de jurisdictie waarin de auteursrechtelijke relevante handelingen plaatsvinden die helpen die AI-modellen voor algemene doeleinden te trainen. Alleen op deze manier kan gezorgd worden voor een gelijk speelveld voor aanbieders van AI-modellen voor algemene doeleinden, waar geen enkele aanbieder zich een concurrentievoordeel op de Uniemarkt mag kunnen verschaffen met lagere auteursrechtelijke normen dan de in de Unie toepasselijke normen.

(107) Voor een grotere transparantie ten aanzien van de bij de pre-training en training van AI-modellen voor algemene doeleinden gebruikte data, met inbegrip van door het auteursrecht beschermde tekst en data, is het passend dat aanbieders van dergelijke modellen een voldoende gedetailleerde samenvatting maken en publiceren van de voor de training van het AI-model voor algemene doeleinden gebruikte content. Terdege rekening houdend met de noodzaak tot bescherming van bedrijfsgeheimen en vertrouwelijke bedrijfsinformatie moet deze samenvatting breed van karakter zijn in plaats van technisch gedetailleerd, teneinde partijen met legitieme belangen, waaronder houders van auteursrechten, in staat te stellen hun rechten uit hoofde van het Unierecht uit te oefenen en te handhaven. Zo kan er bijvoorbeeld een opsomming worden gegeven van de belangrijkste gegevensverzamelingen of -reeksen waarmee het model is getraind, zoals grote particuliere of openbare databanken of gegevensarchieven, en kan een uitleg in de vorm van een relaas worden gegeven over andere gebruikte gegevensbronnen. Het is passend dat het AI-bureau een model voor die samenvatting verstrekt, dat een eenvoudig en doeltreffend model dient te zijn waarmee de aanbieder de vereiste samenvatting in de vorm van een relaas verstrekken kan.

(108) Wat betreft de verplichtingen van aanbieders van AI-modellen voor algemene doeleinden om een beleid in te voeren om aan het recht van de Unie inzake auteursrecht te voldoen en om een samenvatting van de voor de training gebruikte content te publiceren, moet het AI-bureau controleren of de aanbieder aan die verplichtingen voldoet, zonder evenwel geval voor geval de trainingsdata te controleren of te beoordelen qua naleving van het auteursrecht. Deze verordening doet geen afbreuk aan de handhaving van de auteursrechtregels krachtens het Unierecht.

Generatieve AI: training data

The following tables provides a summarization of included training data.

Dataset	Description	Percentage	Total Size (Tokenized)
Web Crawls	Large web scrape corpora (e.g. Common Crawl) containing various styles and sources	71%	2,77TB
Books	Fiction and non-fiction literature providing well-structured and coherent text on various topics	20%	0,79TB
Political and Legal Sources	Data provided by the EU parliament, legislation and speeches	5%	0,18TB
Wikipedia	Wikipedia provides well-structured and mostly factual information	2%	0,07TB
News	News articles from various journals	2%	0,06TB
	Collection of smaller, more		

Auteursrechtbepalingen in AIV: Inwerkingtreding en handhaving

- AIV aangenomen op 13/6/2024, i.w. 2/8/2024
- Regeling GPAI-modellen (Hfdst V) per 2/8/2025
  - GPAI-modellen al op de markt: per 2/8/2027 (!)
- Handhaving: Europese Commissie/AI Office
  - “zonder evenwel geval voor geval de trainingsdata te controleren of te beoordelen qua naleving van het auteursrecht” (O. 108)
- *GPAI Code of Practice* gereed per 2/5/2025

List of Chairs & Vice-Chairs

Working Group 1: Transparency and copyright-related rules

Co-Chair (Transparency): Nuria Oliver (Spain)

Nuria Oliver is the Director of the ELLIS Alicante Foundation and holds a PhD in AI from MIT. She has 25 years of research experience in human-centric AI, spanning academia, industry, and NGOs. Nuria is an independent board member of the Spanish Supervisory Agency of AI, a member of the International Expert Advisory Panel to the Scientific Report on the Safety of Advanced AI, and a Fellow of IEEE, ACM, EurAI, and ELLIS. She is also the co-founder and vice-president of ELLIS.

Co-Chair (Copyright): Alexander Peukert (Germany)

Alexander Peukert is a Professor of Civil, Commercial, and Information Law at Goethe University Frankfurt am Main. With over 25 years of experience, he is a leading expert on European and international copyright law, focusing recently on the intersection of copyright and artificial intelligence. He has been a member of the Expert Committee on Copyright of the German Association for the Protection of Intellectual Property (GRUR) since 2004 and is a founding member of the European Copyright Society, which he chaired in 2023/2024.

Vice Chair (Transparency): Rishi Bommasani (US)