To Mine or Not to Mine: Knowledge Custodians Managing Access to Information in the Age of AI

Ana Lazarova and Eric Luth

ABSTRACT

The article addresses the legal challenges surrounding the computationally-driven reuse of digital cultural heritage collections for the purpose of training large AI models. It examines the role of knowledge custodians, such as public sector actors like cultural heritage institutions, but also non-governmental commons-based projects such as Wikimedia Commons and Flickr Commons and intergovernmental organisations such as UN agencies, in managing access to these materials. Focusing on the EU's text and data mining (TDM) regime, this contribution considers the impact of copyright and related rights on AI training. It further highlights the complexities faced by knowledge custodians in navigating access rights and copyright management, particularly in exercising rightsholder reservations under Article 4 of Directive (EU) 2019/790, with respect both to content that remains under copyright and such that has entered the public domain.

1. INTRODUCTION

Recently, the expanding use of Artificial Intelligence (AI) in the creation of diverse artistic works, along with the increasing availability of sophisticated generative AI models to the general public, has drawn the creative industries into active discussions about the implications of the technology. This heightened engagement has brought significant attention to the challenges that the development and deployment of such systems pose to the copyright and related rights legal frameworks. This contribution focuses on specific issues around the legal status and regulation of materials used to train large foundation models (so-called input issues), which have sparked new tensions between copyright maximalists and advocates of open access to knowledge.[1]

Given that AI training requires the processing of vast quantities of content, including content sourced from knowledge institutions, these institutions have recently assumed the role of, sometimes reluctant, go-betweens for content providers and a new generation of content users—AI system developers and deployers. Such public sector actors include educational, research, and cultural heritage institutions (CHIs), but also intergovernmental organisations such as UN agencies, as well as platforms that serve as repositories of content from CHIs, such as Europeana. In a broader spectrum of knowledge custodians, commons-based projects such as Wikimedia Commons and Flickr Commons, open-source software development and sharing platforms and other repositories hosting different types of content also play a significant role in making available such vast quantities of data needed for the training of AI models.

The growing importance of such institutions and custodians, in the wake of the emerging AI models, means that their decisions and strategies can influence the quality of the output of the AI models. Although traditionally advocates of knowledge-sharing, the rapid development of AI systems, especially General-Purpose AI (GPAI) models trained on such content, has posed new questions and issues around how all these actors govern access to the resources they manage and has also recontextualised their activity and public interest mission.

2. TEXT AND DATA MINING AND AI TRAINING

The recent advancements in language models are largely due to the use of vast, diverse datasets for training, including pretraining corpora, fine-tuning datasets curated by academics, synthetic data, and data aggregated from various platforms. Currently, over 30 lawsuits have been filed

1 According to the Report of July 2024 on digital replicas released by the US Copyright Office, "AI raises fundamental questions for copyright law and policy, which many see as existential." See United States Copyright Office, 'Copyright and Artificial Intelligence Part 1: Digital Replicas. A Report of the Register of Copyrights' (2024) <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-1-Digital-Replicas-Report.pdf>. See also *inter alia* A Guadamuz, 'A short guide to the Copyright Wars' (*Technollama*, 2024) <www.technollama.co.uk/a-short-guide-to-the-copyright-wars>.

against OpenAI and other generative AI companies in the United States, the majority of which involve allegations of copyright infringement.[2] At the heart of many of these legal battles is whether the large-scale scraping of content and subsequent use in training GPAI models qualifies as 'fair use.'

In contrast, Europe has partly solved this issue. The basis of AI training is a process called 'text and data mining' (TDM), which, according to EU law, refers to 'any automated analytical technique aimed at analysing text and data in digital form in order to generate information such as patterns, trends, and correlations' – paragraph 2 of Article 2 of Directive (EU) 2019/790 (the CDSM Directive).[3] Furthermore, under Article 3 of the same Directive, a mandatory exception permits research organisations and cultural heritage institutions to make reproductions and extractions for scientific TDM purposes, provided they have lawful access to the materials mined. This exception cannot be overridden by contracts or technical protection measures (TPMs). Article 4 introduces a broader exception applicable to both commercial and non-commercial users, which can be overridden unilaterally by rightsholders if they explicitly reserve their rights. Thus, according to EU law, AI training is a form of use covered by copyright exceptions, from which rightsholders can formally opt out in certain cases.

Even though the EU appears to have established a clearer regulatory framework on AI training than the US, it has not set itself entirely apart from the ongoing legal uncertainty concerning the use of creative content for the training of generative AI models. One ongoing debate concerns whether TDM applies to AI training at all. While this is not widely recognised as an issue in academia or among policymakers, many rightsholders argue that AI training falls outside the scope of TDM exceptions.[4] Others concede that training large foundation models technically constitutes a form of TDM, but argue that including AI training within the scope of the exceptions was not the policymakers' intention. This is incorrect. As an example of the EU legislator's intent, Article 53(1)(c) of the recently adopted AI Act[5] states that '*Providers of general-purpose AI models* shall [...] put in place a policy to comply with Union law on copyright and related rights, and in particular to *identify and comply with*, including through state-of-the-art technologies, *a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790*'. Furthermore, the inclusion of AI training within the scope of TDM was affirmed in a high-profile case before the Hamburg Regional Court—the first of its kind in Germany, and likely in the EU.[6] The case concerned a stock photographer's claims against the Large-scale Artificial Intelligence Open Network (LAION), a non-profit providing machine learning resources for the public. The court ruled that LAION's activity in relation to the LAION-5B image-text dataset for training large AI models constituted text and data mining (TDM) under EU law, and applied Article 3 of the CDSM Directive and Section 60d of the German Copyright Act.[7]

Another issue concerning the practical implementation of the TDM exceptions is the notion of lawful access. The content and scope of the term for the purposes of Articles 3 and 4 of the CDSM Directive are yet to be thoroughly interpreted by the judiciary. It should be taken into account that the associated concepts of "lawful use" and "lawful source" in the EU *acquis* are complicated.[8] They require, for the use under an exception to be lawful, that the subject matter was made available with the consent of the rightsholder. The unclear scope of the notion of the rightsholder's consent may, in the future, attach to this requirement a potentially detrimental effect on legal certainty concerning the use of licensed materials. Nevertheless, in the decision of the German court in the LAION case, the file(s) was found to be 'lawfully accessible' on the stock photo website.[9]

The foremost challenge, however, lies in the practical implementation of the aforementioned exceptions, compounded by significant confusion regarding who is entitled to opt out of the mechanisms of Article 4 and how the rightsholder reservation should be made. This outcome was hardly surprising to copyright experts, as the general TDM exception in the CDSM Directive (and, for that matter, – the fall-back exception as per paragraph 2 of Article 8 thereof) is not the first EU-level exception to include an opt-out mechanism, nor is it the first whose implementation has posed challenges for national courts. The so-called 'press review' exception, set out in the first part of Article 5(3)(c) of Directive 2001/29/EC (the InfoSoc Directive),[10] concerns reproduction by the press, communication to the public, or making available of published

2 At the time of the submission of this contribution, there are 33 lawsuits filed against OpenAI, Microsoft, Meta, Midjourney & other GPAI companies. See 'Master List of lawsuits v. AI' (*ChatGPT is Eating the World*, 27 August 2024) <https://chatgptiseatingtheworld.com/2024/08/27/master-list-of-lawsuits-v-ai-chatgpt-openai-microsoft-meta-midjourney-other-ai-cos/>.

3 Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.

4 See *inter alia* Diskurs, 'Study Reveals AI Training is Copyright Infringement' (*Urheber*, 5 September 2024) <https://urheber.info/diskurs/ai-training-is-copyright-infringement>.

5 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

6 Landgericht Hamburg, Urteil vom 27.09.2024, Az. 310 O 227/23.

7 Ibid.

8 According to Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, recital 33, 'A use should be considered lawful where it is authorised by the rightholder or not restricted by law.' See also Case C–527-15 *Stichting Brein v. Jack Frederik Wullems (Filmspieler)* [2017] ECLI:EU:C:2017:300, paras 65 et seq., and Case C–435/12 *ACI Adam BV et al. v. Stichting de Thuiskopie*, *Stichting Onderhandelingen Thuiskopie vergoeding* [2014] ECLI:EU:C:2014:254, para 38.

9 (Landgericht Hamburg, n 6).

10 Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society *OJ L 167, 22.6.2001*.

articles on current economic, political or religious topics, or of broadcast works or other subject matter of the same character, 'in cases where *such use is not expressly reserved*'. Specific requirements for the opt-out mechanism have been established through case law in many Member States.[11] In Bulgaria, for instance, the courts have in the past demonstrated great inconsistency regarding precisely who is entitled to opt out of the press review exception and the manner in which such an opt-out may be exercised. One particularly problematic interpretation in a judicial decision asserts that a rightsholder may retroactively express their objection to the free use of their article merely by filing a copyright infringement claim.[12]

3. THE RIGHTSHOLDER RESERVATION CONUNDRUM

According to Article 4(3) of the CDSM Directive, the exception '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* in an *appropriate manner*, such as by *machine-readable means* in the case of content made publicly available online.' Furthermore, paragraph 2 of Recital 18 explains that '[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*, including metadata and terms and conditions of a website or a service. [...] In other cases, it can be appropriate to reserve the rights *by other means*, such as contractual agreements or a unilateral declaration.'

Currently, both EU institutions and civil society are exploring technical solutions to address the need for a standardised machine-readable rights reservation under the general TDM exception.[13] Although it is recognised that no one-size-fits-all opt-out technical solution exists, in terms of crawling and data retrieval by search engines, the industry standard involves using a *robots.txt* file, placed in the website's root directory, to block crawlers from accessing and indexing specific parts of the site. Additionally, individual pages can use a *robots* meta tag in their header to control whether they are allowed to be indexed or cached, effectively creating an opt-out mechanism for those pages. Some authors are even arguing that the lack of such standardised automatic reservation constitutes an opt-out implied licence.[14]

On the other end of the spectrum, there are views that other forms of expressed will, including dissemination under standard public licences, and even the use of non-machine-readable, notices can constitute valid opt-outs under Article 4 of the CDSM Directive. Creative Commons was compelled to issue a formal opinion on whether the licences the organisation manages, particularly the non-free/open ones,[15] impose partial restrictions on the use of the relevant material and whether the NoDerivatives (ND) and NonCommercial (NC) clauses constitute an exercise of the opt-out option under Article 4 of the CDSM Directive. In a statement of November 2021, the organisation said that CC licences could not be perceived or interpreted as a reservation of rights within the context of Article 4 of the CDSM Directive or any relevant national provisions, as they could not, in principle, serve as a waiver of exceptions or limitations to copyright. A fundamental aspect of Creative Commons,[16] and most open licences, including the GPL,[17] is the explicit assertion that use is covered by the licence only if applicable law restricts that use, and therefore, any interpretation suggesting that they prohibit use within the context of Article 4 would be contrary to their overall design and purpose.

Commentators have recently also studied the effect of ShareAlike (SA) obligations and copyleft licensing on machine learning, AI training, and AI-generated content.[18] This particular issue seems to be pertinent, given that, according to a recent multi-disciplinary study mapping the AI data supply chain and looking at the empirical licence use for natural language processing datasets, the most common licence in a popular sample of the major supervised NLP datasets is CC-BY-SA 4.0 (15.7%), while 33% of the licences contain a ShareAlike clause (such as CC-BY-NC-SA 4.0, CC-BY-SA 3.0 and GPL v.3).[19] In gen-

11 For a detailed analysis of the divergent national implementations of the two informatory exceptions as per art 5.3.c of the InfoSoc Directive see A Lazarova, 'Re-use the news: between the EU press publishers' right's addressees and the informatory exceptions' beneficiaries' (2021) 16(3) JIPLP 236.

12 Decision No 193, Commercial Appeal Case No 3149/2015, Sofia Court of Appeal.

13 See European Commission, AI Act: Participate in the drawing-up of the first General-Purpose AI Code of Practice (2024). <https://digital-strategy.ec.europa.eu/en/news/ai-act-participate-drawing-first-general-purpose-ai-code-practice#:~:text=The%20Code%20of%20Practice%20will,of%20Practice%20to%20demonstrate%20compliance>. See also P Keller, 'Open Future Policy Brief' (*Open Future*, 24 May 2024) <https://openfuture.eu/wp-content/uploads/2024/05/240516considerations_of_opt-out_compliance_policies.pdf>.

14 H Zhang and Y Li, 'Opt-Out Implied Licenses in Copyright Law: From Search Engines to GPAI Models', (2024) 73(9) GRUR International, <https://doi.org/10.1093/grurint/ikae088>.

15 The difference between free and non-free licences is the scope of rights that are granted by the licensee. Creative Commons manages six standard licences, of which, with respect to the criteria set by the 1991 Free Software Foundation definition and the 1998 Open Source Initiative definition, two are free/open (CC-BY and CC-BY-SA) and four are not (CC-BY-NC, CC-BY-ND, CC-BY-NC-ND and CC- BY-NC-SA).

16 See for instance the legal code of CC-BY 4.0, Section 2(a)(2): "For the avoidance of doubt, where Exceptions and Limitations apply to Your use, this Public License does not apply, and You do not need to comply with its terms and conditions." Exceptions and limitations are defined in sec. 1(d) as "Exceptions and Limitations means fair use, fair dealing, and/or any other exception or limitation to Copyright and Similar Rights that applies to Your use of the Licensed Material." <https://creativecommons.org/licenses/by/4.0/legalcode>.

17 According to Section 2 of the GNU General Public License, Version 3, 29 June 2007, 'This License acknowledges your rights of fair use or other equivalent, as provided by copyright law.'

18 K Szkalej and M Senftleben, 'Generative AI and Creative Commons Licences: The Application of Share Alike Obligations to Trained Models, Curated Datasets and AI Output' (2024) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4872366>.

19 S Longpre, R Mahari and A Chen, 'A large-scale audit of dataset licensing and attribution in AI' (2024) 6 Nat Mach Intell <https://doi.org/10.1038/s42256-024-00878-8>. The study is based on an audit of AI

eral, commentators think that at present, copyleft clauses do not impede mining. However, while some believe that it may be advisable to abandon the traditional precedence of copyright exceptions in favour of an opt-out protocol that allows a more fine-grained TDM permission that includes SA obligations,[20] others argue that such licences have a direct propagating effect on the whole model, or even on its output.[21] Finally, it should be acknowledged that irrespective of doctrinal interpretations, a recent dataset audit by the Data Provenance Initiative found that more than 70% of licences for popular datasets on GitHub and Hugging Face were 'unspecified', while licences that were attached to datasets uploaded to dataset sharing platforms were often inconsistent with the licence ascribed by the original author of the dataset and often labelled as more permissive than the author's original licence.[22] The study highlighted a crisis in licence laundering and informed usage of popular datasets, with systemic problems in sparse, ambiguous or incorrect licence documentation.[23] Thus, even if public licensing of materials used for AI training could have been considered a legitimate way to opt-out of text and data mining for the purposes of the application of the general TDM exception, it seems that it is not at present a reliable opt-out tool.

The issues around the form and the effect of the rightsholder reservation under Article 4 of the CDSM Directive and the implementing provision of Section 44b of the German Copyright Act, were commented on the Hamburg Regional Court in obiter dictum (non-binding). According to the LAION decision, the photographer's opt-out clause in the website's terms and conditions might have been enforceable against commercial mining. Although the opt-out was in natural language, rather than a formal protocol (e.g. *robots.txt*), the court suggested it could still be valid, assuming available technologies could interpret such reservations.[24] In theory, and according to the first available decision on the matter, the natural language opt-out can be machine-readable. In practice, such an opt-out would most likely be 'read' by the machine after processing the data scraped from a website in its entirety, which would make the opt-out somewhat redundant. For this reason, in its CDSM implementation proposal, the Bulgarian government resorted to requiring opt-outs to be done by technical means 'immediately recognisable by the software performing the automated analysis.'[25] This part of the proposal provision was removed at the last minute on the insistence of representatives of the music industry with the motive of following the text of the Directive as strictly as possible.[26]

4. COMPUTATIONALLY-DRIVEN REUSE OF DIGITAL HERITAGE AND THE ROLE OF KNOWLEDGE CUSTODIANS

Knowledge institutions such as research organisations and memory institutions utilise AI in multiple capacities. Certain AI applications prove particularly valuable in enhancing the analysis and accessibility of knowledge and cultural heritage, achieving results that would be unattainable or excessively time-consuming without such technological assistance. There are numerous beneficial applications of TDM that align with the mission and objectives of these public sector actors as users. For example, an AI model from the Swedish National Archives can interpret historical handwriting from the 17th, 18th and 19th centuries with a prediction rate of 95%.[27] In this regard, the Swedish Government Report SOU 2024:4 proposed the introduction of a new exception in URL § 16 para 4, that would enable cultural heritage institutions to make digital reproductions for the purpose of internal management and organisation, e.g. for better metadata, explicitly stating that TDM can be a suitable method for this end. Similar exceptions already exist in Finland and Norway.[28]

Using digital heritage[29] for text mining, machine learning, computer vision etc. is not an entirely new concept. For instance, the 'Collections as Data' movement has encouraged the development of 'cultural heritage collections that support computationally-driven research and teaching' since 2016.[30] It can be argued that digital cultural

data provenance, tracing the lineage of more than 1,800 text datasets, their licences, conditions and sources.

20 (Szkalej & Senftleben, n 22).

21 Y Benhamou, 'Open Source AI: Does the Copyleft Clause Propagate to Proprietary AI Models? Revisiting the Definition of Derivatives in the AI-context' (2024) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4859623> accessed.

22 (Longpre et al. n 23).

23 Ibid.

24 (Landgericht Hamburg, n 6).

25 Bill for the Amendment and Supplement of the Law on Copyright and Related Rights, Signature 49-302-01-21, submitted in Parliament on 13 April 2023 <https://www.parliament.bg/bg/bills/ID/164728>.

26 According to the rightsholders' position, 'The letter and meaning of Article 4 of Directive 2019/790 should not be altered or supplemented, as it neither requires that the prohibition by rightsholders must occur 'before' the protected objects are accessed, nor does it stipulate the condition for the technical means to be 'immediately' recognizable by the software performing the automated analysis. Such proposals, which supplement the text of Article 4, paragraph 3 of Directive 2019/790, introduce additional and restrictive conditions that are neither based on nor provided for by the Directive's provisions.' Opinion of the Bulgarian Association of Music Producers (BAMP) regarding the Bill for the Amendment and Supplement of the Law on Copyright and Related Rights (Amendment of the Law on Copyright and Related Rights), signature 49-032-01-21, submitted by the Council of Ministers on 13 April 2023 <www.parliament.bg/bg/parliamentarycommittees/3219/standpoint/16872>.

27 O Karsvall, 'Ny banbrytande AI-modell för svenska historiska texter' (Riksarkivet, 7 February 2024) <https://riksarkivet.se/Nyhetsarkiv?item=120354>.

28 Betänkande av Utredningen om upphovsrättens inskränkningar SOU 2024:4.

29 For a definition of the term, see UNESCO, 'UNESCO Charter on the Preservation of the Digital Heritage – UNESCO Digital Library' (UNESCO, 2003) <https://unesdoc.unesco.org/ark:/48223/pf0000229034.locale=en>.

30 See T. Padilla, L Allen, H Frost, S Potvin, ER Roke and S Varner, 'Always Already Computational: Collections as Data' (2020) <https://doi.org/10.17605/OSF.IO/MX6UK>. According to Padilla et al., 'We are seeing an increasing number of requests for machine-actionable data at NYU Libraries, whether in the form of full-text collections, bibliographic metadata, or both, from data researchers seeking corpora to perform

heritage datasets are generally of high quality. They are usually carefully curated and documented and are substantial in size and diversity.[31] The collections of libraries, for instance, may include content, i) from different times, reflecting changes in language and tonality, ii) of different registers, reflecting different ways of expressing language, as well as iii) of different genres, which is crucial to provide output reflecting different kinds of prompts. A national library in a country with a legal deposit system[32] might, for example, have novels and poetry from many different centuries, political protocols and annals, newspapers, local publications on dialect, and even commercials and historical propaganda. National libraries in some EU countries are crawling the web, to store it for future generations for research purposes. The National Library of Sweden has e.g. crawled the .se domain since the mid-1990s, collecting more than 500 million web pages.[33]

Much of the content of such institutions might be out of copyright, whereas other parts are still covered by copyright. Older material is needed, as well as more modern content, in the training of the AI system. TDM on national library content has been carried out on radio broadcasts, and newspaper editorials,[34] to name two examples. TDM on book reviews was made possible through large-scale digitisation of the Swedish literary press, and has resulted both in quantitative analyses of Swedish literary criticism as well as an AI that can recognise book reviews among other texts.[35] However, the debate, being nowadays dominated by large tech companies and generative AI, as well as AI systems needing vast quantities of content from diverse sources to be able to provide qualitative output, has put the position of 'donors' of minable data of these public sector actors in a new light for both ethical and practical reasons.

Furthermore, many CHIs use third-party repositories to make their content available to the general public. This involves portals such as Europeana, or Digitalt museum in Sweden, where staff contribute content to be used by the public; it may also involve repositories and platforms such as Wikimedia platforms and Flickr, webpages for user-generated content where both individual users and staff at cultural heritage institutions contribute content. UNESCO Archives, as one IGO, have made many thousands images from its archives available via Wikimedia Commons. Such repositories and platforms, together with the institutions and users supplying content to them, enrich the wealth of publicly shared knowledge known as the Digital Commons, defined as 'a subset of the Commons, where the resources are data, information, culture and knowledge which are created and/or maintained online'.[36] All of these actors might not be defined as cultural heritage institutions, but they all play a crucial role in actively promoting the digital dissemination of works under open licences or in the public domain.[37] In doing so, they serve a pivotal function in supplying AI training data.

Wikipedia is one of several websites created by the Wikimedia movement whose mission is to make the sum of human knowledge freely available to all, building on Creative Commons Licences and allowing reuse under

topic modeling, network modeling, machine learning, and other natural language processing tests.'

31 Although according to some authors these datasets also have their limitations for the purposes of data mining, as they are marked by specific characteristics, such as being the product of multiple layers of selection, being created for different purposes than establishing a statistical sample according to a specific research question, hanging over time and being heterogeneous. See H Alkemade, S Claeyssens, G Colavizza, N Freire, J Lehmann, C Neudecker, G Osti and D van Strien, D, 'Datasheets for Digital Cultural Heritage Datasets' (2023) 9(1) Journal of Open Humanities Data, <https://doi.org/10.5334/johd.124>.

32 See e.g. Kungliga biblioteket, 'Legal deposit' (9 January 2024), <https://www.kb.se/in-english/about-us/how-we-collect-material/legal-deposit.html>, accessed 18 October 2024.

33 kulturarw3, 'Svenska webbsidor från mitten av 1990-talet och framåt', (*Kungliga biblioteket*, 2024) <https://www.kb.se/hitta-och-bestall/hitta-i-samlingarna/kulturarw3.html>.

34 M Hurtado Bodell, M Magnusson and S Mützel, 'From Documents to Data: A Framework for Total Corpus Quality' (2022) 8 Socius <https://doi.org/10.1177/23780231221135523>.

35 J Ingvarsson, D Brodén, L Samuelsson, N Zechner and V Wåhlstrand Skärström, 'The New Order of Criticism. Explorations of Book Reviews Between the Interpretative and Algorithmic' (2022) DHNB The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022) <https://ceur-ws.org/Vol-3232/paper20.pdf>. accessed 18 October 2024.

36 M Dulong de Rosnay and F Stalder, 'Digital Commons' (2020) 9(4) *Internet Policy Review* <https://policyreview.info/concepts/digital-commons>.

37 Contributions to the digital commons include: Free Culture, Free / Open Source software, Open Access, Open Data, Open Design, Open Education, Open GLAM/Open Culture, Open Government, Open Hardware, Open Internet / Open Web and Open Science. See A Tarkowski, P Keller, Z Warso, K Goliński and J Koźniewski, 'Fields of Open. Mapping the Open Movement' (*Open Future*, 6 July 2023) <https://openfuture.pubpub.org/pub/fields-of-open>.

certain conditions.[38] The extent to which AI developers use freely licensed text, imagery, and data from the Wikimedia platforms to train the models is unknown. The Wikimedia Foundation states that literally every large language model (LLM) is trained on Wikipedia text,[39] and according to *The Washington Post*, Wikipedia and content from the other Wikimedia platforms are almost always the largest source of training data in their data sets for those LLMs.[40] The Pile, a common open-source dataset for large language models (LLMs), includes for example Wikipedia as a standard source of high-quality text.[41] The educational, research, and estimated monetary value of the content on the Wikimedia platforms has grown over time; research indicates that the downstream usage of images from Wikimedia Commons produces a value of USD 28.9 billion over the lifetime of the project.[42] This sum was, however, calculated before the emergence of General Purpose AI (GPAI) models such as GPT.[43]

5. THE CUSTODIAN'S OPT-OUT

It was clarified previously, that because of their unique role within the EU TDM legal regime, public sector actors among knowledge custodians, such as CHIs in general and public and academic libraries in particular, find themselves in a pivotal position where commercial AI training is concerned. By extension, the discussions and decisions of CHIs and custodians of the commons might have a significant impact on the future development of AI tools. In addition, public sector knowledge custodians also face considerable pressure from rightsholders and information providers regarding how these institutions manage access to their collections.

In terms of eligibility to opt out of mining, knowledge custodians have an unclear standing. CHI ownership management, based on acquisition, inheritance, or first publication, is increasingly complex, especially in a digital setting.[44] That being said, in the typical scenario, copyright is not transferred to the CHIs. Thus, knowledge custodians are usually not rightsholders over the materials in their collections. Pursuant to the requirements of paragraph 3 of Article 4 of the CDSM Directive, 'The exception [...] 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*'. Thus, knowledge custodians may not be entitled to 'reserving' rights that they do not carry, on their own behalf, or on behalf of rightsholders they do not represent. In the context of Article 4, that means that the right to opt out is also not transferred to the CHI – unless the CHI, according to recital 18, is involved in 'contractual agreements or a unilateral declaration' of materials, accessible offline. The recent litigation against LAION in Germany has revealed that an opt-out can be considered valid when executed by a third party, provided there is a contractual agreement in place between the plaintiff and that third party.[45]

Currently, however, many rightsholders seem to be contractually obliging knowledge custodians as users of content for public interest purposes, to exercise tighter control on re-use than strictly required by the current EU legislation. On the one hand, there seems to be a clear trend for publishers and other information vendors to try and contract out of TDM under the research exception as per Article 3 of the CDSM Directive. A recently published study analysed 100 licensing contracts between scientific publishers, data vendors, public libraries, and research institutions and revealed that more than half of these agreements, concluded after 2019, sought to restrict even non-commercial TDM.[46] Many contracts prohibited mining by institutional users, either explicitly or implicitly – through the express prohibition of the use of robots, spiders, crawlers, or other automated downloading programmes, or on the continuous and/or automatic search or indexing of the licensed materials or databases, etc. Others limited or failed to address TDM rights altogether.[47] This trend creates legal uncertainty and a potential chilling effect on the overall use of the TDM exceptions.

Another visible trend is for collective management organisations (CMOs) to impose on CHIs an obligation to opt out of TDM on out-of-commerce collections. This is the situation in the Netherlands, where in the recent agreement on periodicals between the National Library (and affiliated CHIs) and CMOs Pictoright and LIRA, the institutional users were obliged to 'make it known by means of an appropriate machine-readable rights reservation that the Periodicals may not be used for text and

38 E Kelly, 'Reuse of Wikimedia Commons Cultural Heritage Images on the Wider Web' (2019) 14(3) *Evidence Based Library and Information Practice* <https://journals.library.ualberta.ca/eblip/index.php/EBLIP/article/view/29575>.

39 S Deckelmann, 'Wikipedia's Value in the Age of Generative AI' (*Wikimedia Foundation*, 12 July 2023) <https://wikimediafoundation.org/news/2023/07/12/wikipedias-value-in-the-age-of-generative-ai/>.

40 K Schaul, S Y Chen and N Tiku, 'Inside the Secret List of Websites That Make AI like ChatGPT Sound Smart' *Washington Post* (Washington, D. C., 19 April 2023) <https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/>.

41 S Biderman, K Bicheno and L Gao, 'Datasheet for the pile' (2022), *arXiv preprint* <https://arxiv.org/abs/2201.07311>.

42 K Erickson, F Rodriquez Perez and J Rodriguez Perez, 'What is the Commons Worth? Estimating the Value of Wikimedia Imagery by Observing Downstream Use' (2018) *Proceedings of the 14th International Symposium on Open Collaboration* <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3206188>.

43 GPAI is not to be confused with Artificial General Intelligence (AGI).

44 An example of the challenges encountered by the cultural heritage sector in relation to rights clearance is the case study of the Polish History Museum's implementation of a copyright-management strategy. See Pluszyńska, A. (2021). Copyright Management in Museums: Expediency or Necessity? Museum International, 73(3–4), 132–143 <https://doi.org/10.1080/13500775.2021.2016281>.

45 The court, in an obiter dictum (non-binding), addressed the 'general' TDM exception under Article 4 of the CDSM Directive and Section 44b of the German Copyright Act. It noted that the photographer's opt-out clause in the website's terms and conditions could potentially be enforceable against commercial data mining. (Landgericht Hamburg, n 6).

46 See A Lazarova, 'Libraries, Licences, Limitations: Assessing Licensing Provisions Between Publishers and Knowledge Institutions' (2024) <www.knowledgerights21.org/reports/ the-100-contracts-report/>.

47 Ibid.

data mining with a commercial purpose within the meaning of Article 150 of the Copyright Act and Article 4 of the DSM Directive, including use for AI training purposes'.[48] The OOCW regime, also introduced with the CDSM Directive, allows CHIs to share online materials that are no longer in commercial circulation but are still under copyright. The goal of the legal regime was to alleviate the often-insurmountable task of clearing copyright for vast collections. This is primarily done through extended collective licences (ECL), meaning that the CMO's mandate extends to all authors within a particular sector, whether or not they have explicitly signed a contract with the organisation. Thus, it is questionable whether a CMO under an ECL – which covers all authors in a certain sector regardless of the presence of a contractual relationship with the CMO or not – has the authority to enforce an opt-out.[49]

Even more problematic, this approach can transfer to out-of-copyright material, even though, in theory, this should not be possible given Article 14 of the CDSM Directive, an article sometimes referred to as the 'safeguarding to the public domain', stipulating that new copyright cannot be claimed on a reproduction of a work for which copyright no longer applies. In this regard, digitisation may create a subset of problems concerning the ownership and management of content that can translate into challenges regarding access to knowledge institutions' collections and databases. For instance, as the digitisation of cultural heritage is inherently costly and demanding not only substantial financial investment but also the dedication of expert resources of institutions tasked with preservation, there is often a certain contradiction in the motivation of the staff involved in libraries, archives and museums. Moreover, in cases where rights have expired or certain materials were not eligible for copyright protection, there can be some resistance to 'recognising' a public domain status for content concerned. Museums have for example, based on the Article 4 of the Copyright Term Directive, tried to claim the 25 years protection 'equivalent to the economic rights of the author' for the first publication or communication to the public of a previously unpublished work.[50] Other institutions take the opposite stance. The National Archives and National Museum of Sweden have both adopted policies stating that no new copyright arises on digital reproductions, and that the content produced by their staff is openly licensed.[51]

Nevertheless, some institutions may seek to control access and usage to mitigate the risk of infringement or to recoup the resources expended in digitisation. All of these factors, paired with a general trend of techno-pessimism and distrust of 'big tech', are contributing to another trend in collection management by knowledge custodians: some are routinely and indiscriminately 'closing' the entire content in their custodianship to outside automatic processing. For example, in 2023, the National Library of the Netherlands (KB) excluded bots from mining their online collections, including both copyrighted and public domain works, via *robots.txt*.[52]

In this context, the discussion around the management of access to collections and their use for AI training is also pertinent to commons-based projects. According to some commentators, Wikipedia Share-Alike licences would propagate to all the output of ChatGPT.[53] Then again, according to a statement from the Wikimedia Foundation, even though Wikimedia generally supports the use of Wikipedia content – which is freely accessible and valuable for training – for model AI development, "some model developers may be out of compliance with the attribution clause of the CC BY-SA license", since many large language models fail to disclose the sources of their training data. Compliance, however, according to Wikimedia, hinges on whether courts determine that using such data for training qualifies as fair use.[54] Accordingly, in EU context, licence conditions would only apply to AI training, if the latter is done outside of non-commercial research and there has been a formal reservation by the respective rightsholder first.

In conclusion to this part, regarding the opt-out exercised by knowledge custodians, the available legal framework at the EU, level as well as the first case law around TDM, indicate that custodians of data are unlikely to be entitled to routinely exercise reservations under Article 4 of the CDSM Directive without explicit consent from the rightsholders. This means that, above all, these actors have no legal grounds based in copyright law for limiting access to public domain materials. Even where works in their collections are in copyright, custodians are not entitled to limit user rights on their own behalf and by their own initiative. Furthermore, while contractual arrangements with rightsholders can form a basis for establishing valid opt-outs, agreements with CMOs operating under extended licensing may not constitute a valid expression

48 M Zeinstra, 'Werken die niet langer in de handel zijn' (KVAN, 2024). <https://www.kvan.nl/themas/auteursrecht-werken-die-niet-langer-in-de-handel-zijn/>.

49 A Matas, 'AI "opt-outs": should cultural heritage institutions (dis)allow the mining of cultural heritage data?' (Europeana, 2024) <https://pro.europeana.eu/post/ai-opt-outs-should-cultural-heritage-institutions-dis-allow-the-mining-of-cultural-heritage-data>.

50 See, for example the case about the so called Nebra Sky Disk, *Kosturik v. Land Sachsen Anhalt* [2010] S 216/09 Deutsches Patent- und Markenamt Dienststelle Jena, <https://www.rechtsanwaltmoebius.de/urteile/DPMA_30507066_Marke_Himmelsscheibe-von-Nebra.pdf>.

51 See e.g. Riksarkivet, 'Hantering och användning av fotografier och bildkonstverk som finns hos Riksarkivet', 1 May 2016, <https://riksarkivet.se/Media/pdf-filer/UPPHOVSR%C3%84TT%20FOTO%20160501.pdf>.

52 M Kleppe, 'Statement on Commercial Generative AI (KB – National Library of the Netherlands)' (KB, 9 January 2024) <https://www.kb.nl/en/ai-statement>. Although here again there are examples of good practices. See e.g. the Berlin State Library – CrossAsia, 'From people reading to machines learning – how Gaia-x enables digital cultural heritage' (2023) <https://blog.crossasia.org/from-people-reading-to-machines-learning-how-gaia-x-enables-digital-cultural-heritage/?lang=en>.

53 (Benhamou, n 25).

54 Wikimedia Foundation, 'Wikimedia Foundation's Responses to the United States Copyright Office Request for Comments on Artificial Intelligence and Copyright Docket No. 2023-6' (30 October 2023) <https://upload.wikimedia.org/wikipedia/commons/f/f7/Wikimedia_Foundation%E2%80%99s_Responses_to_the_US_Copyright_Office_Request_for_Comments_on_AI_and_Copyright%2C_2023.pdf>.

of will from rightsholders, since CMOs' authority over certain authors is solely based on the extended mandate and not on individual contractual agreements. Finally, this rule would also apply to collections bearing Creative Commons or other open licences, as, according to the current state of the art, these standard public licences do not in any way imply a unilateral rightsholder opt-out from a copyright exception.

6. CONCLUSION

Despite the ongoing legal and ethical challenges surrounding AI training, knowledge custodians continue to play a critical role in the digital age. Many cultural heritage institutions have, however, a traditionally cautious approach to risk, combined with a need for recognition of their work in digitising and managing collections. This approach often results in a desire to control their curated content, a conservative stance that can clash with the mission to make content publicly accessible. In addition, internal and external pressure may sometimes lead to restrictions on access to materials that the knowledge custodians may not be entitled to control, and that lack commercial value for rightsholders (such as out-of-commerce works), or that are even out of copyright.

Nonetheless, the challenges posed by AI training on digital cultural heritage, including legal considerations related not only to copyright but also to privacy and other concerns, must be carefully addressed. Knowledge custodians should not be left to navigate these issues alone. The EU has made an initial move towards establishing legal certainty by offering a multi-tiered approach to TDM, thereby addressing the training of AI models. Future efforts and resources should be dedicated to further developing technical standards and tools that would empower rightsholders to directly exercise their rights within the established legal framework. These solutions must enable effective opt-outs that meet the needs of both rightsholders and AI model developers, but also allow knowledge custodians to operate in legal certainty.

Ana Lazarova

Ana Lazarova is a lawyer specialising in IP and IT law and a senior assistant professor at the Department of European Studies at Sofia University "St. Kliment Ohridski". She is Chair of the Bulgarian digital rights association Digital Republic, Chapter Lead of Creative Commons Bulgaria, member of the Europeana Copyright Community Steering Group and National Coordinator for Bulgaria of the Knowledge Rights 21 programme.

Eric Luth

Eric Luth holds an M.A. in Comparative Literature and is currently the Project Manager for Involvement and Advocacy at Wikimedia Sverige. He is the National Coordinator for the Knowledge Rights 21 Programme, a European program funded by the Arcadia Fund to promote access to culture, learning and research, and was an expert in the public inquiry reviewing exceptions and limitations in Swedish copyright law.