The Use of Wikipedia, Wikimedia, and Open Access Content for Artificial Intelligence and Text and Data Mining

Eric Luth

ABSTRACT

The role of Wikimedia platforms and the broader Digital Commons in developing artificial intelligence (AI) models remains significant yet underexplored. Wikimedia content, licensed under Creative Commons (CC) licenses, constitutes a primary source of training data for many large language models (LLMs), with implications for both the sustainability of the Digital Commons and compliance with copyright law. This article examines the compatibility of CC licenses with AI training, particularly under the European Union's Copyright Directive on the Digital Single Market (CDSM Directive), which introduced new exceptions for text and data mining (TDM). It identifies scenarios where CC-licensed content can be legally used for AI training and discusses unresolved questions about reproduction, derivation, adaptation, attribution, and share-alike requirements under these licenses. The analysis highlights how stakeholders within the Digital Commons—Wikimedia, GLAM institutions, educational organizations, and intergovernmental organizations (IGOs)—influence the quality and ethical use of AI models. It also examines risks posed by AI usage, such as reduced visibility of source platforms, a decline in volunteer contributions, and diminished sustainability of open knowledge ecosystems. Strategies to uphold the Digital Commons include enforcing share-alike obligations, fostering collaboration among stakeholders, and engaging with AI developers to ensure compliance with CC licenses. The findings underscore the dual potential of open access to enhance AI model quality while maintaining the integrity of digital commons ecosystems. Digital Commons stakeholders must be open in a way that promotes qualitative AI development while maintaining sustainable open knowledge dissemination.

1. INTRODUCTION

The extent to which Artificial Intelligence (AI) developers use freely licensed text, imagery, and data from the Wikimedia platforms to train the models is unknown. The Wikimedia Foundation states that all large language models (LLMs) are trained on Wikipedia text,[1] and according to *The Washington Post,* Wikipedia and content from the other Wikimedia platforms is almost always the largest source of training data in the data sets for those LLMs.[2] The Pile, one common open-source dataset for large language models (LLMs), includes for example Wikipedia as a standard source of high-quality text.[3]

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. The Wikimedia platforms build on Creative Commons (CC) licences, allowing reuse under certain conditions.[4] CC licences are examples of free and open licences designed to let creators and rights holders waive the automatic assignment of certain exclusive rights under copyright law (such as the right to reproduction, commercial exploitation, and modification), to benefit the general public.[5] Meanwhile, the licences allow creators to retain certain rights to the

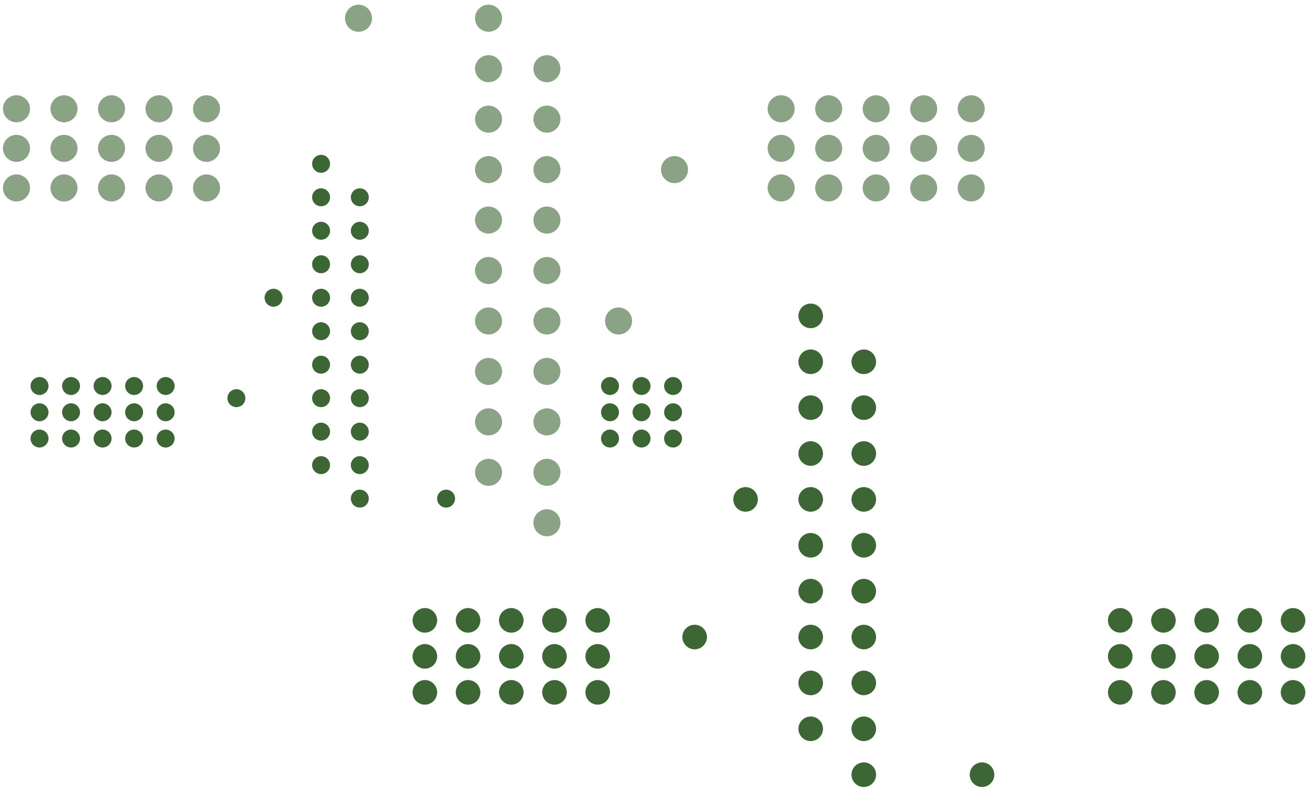
[1] Selena Deckelmann, 'Wikipedia's Value in the Age of Generative AI' (*Wikimedia Foundation*, 12 July 2023) <https://wikimediafoundation.org/news/2023/07/12/wikipedias-value-in-the-age-of-generative-ai/>, accessed 17 October 2024.

[2] 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/> accessed 17 October 2024.

[3] S Biderman, K Bicheno and L Gao, 'Datasheet for the pile' (2022), *arXiv preprint* <https://arxiv.org/abs/2201.07311> accessed 17 October 2024.

[4] For details on Wikipedia and Wikimedia copyright policies, see Editors, 'Wikipedia:Copyrights', (English Wikipedia, 31 March 2024, <'https://en.wikipedia.org/w/index.php?title=Wikipedia:Copyrights&oldid=1216438911>. See also 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> accessed 17 October 2024, for further discussion on reuse of Wikimedia content.

[5] M Dulong de Rosnay, 'Peer to Party: Occupy the Law' (2016) 21(12) *First Monday* <https://firstmonday.org/ojs/index.php/fm/article/view/7117> accessed 17 October 2024.



work including to be credited when used and any derivative work to be licensed under the same licence.

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.[6] This sum was however calculated before the emergence of General Purpose AI (GPAI) models such as GPT.[7] Wikimedia's usage of Creative Commons licences contributes to a larger pool of freely licensed content that is sometimes referred to as *the digital commons*. Melanie Dulong de Rosnay and Felix Stalder define the digital commons as "a subset of the Commons, where the resources are data, information, culture and knowledge which are created and/or maintained online", and further highlight the importance of the concept to counter legal enclosure and foster equal access to the resources.[8] While Wikipedia is a famous example of digital Commons, many other organisations contribute to it, e.g. Galleries, Libraries, Archives, and Museums (GLAM institutions), universities and educational institutions, and others actively promoting the digital dissemination of works under open licences or in the public domain (i.e. works to which copyright no longer applies, or has never been applicable).[9]

This article suggests that Open Access stakeholders, including IGOs like United Nations agencies, the African Union, and European Union institutions, should be considered part of the digital commons movement when they publish using Creative Commons (CC) licences. It also argues that stakeholders in the digital commons have played a key role in the development of GPAI models, a role that may not be fully recognised or understood. The decisions and strategies of these stakeholders—such as the Wikimedia movement, GLAM institutions, universities, and IGOs—can influence the quality of the output from GPAI models. For example, their choices when it comes to open publishing and licensing can directly affect AI models. This raises important questions about the dependence of AI models on the digital commons and the responsibilities the AI models carry toward it.

[6] 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> accessed 17 October 2024.

[7] GPAI is not to be confused with Artificial General Intelligence (AGI).

[8] M Dulong de Rosnay and F Stalder, 'Digital Commons' (2020) 9(4) *Internet Policy Review* <https://policyreview.info/concepts/digital-commons> accessed 17 October 2024.

[9] 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> accessed 17 October 2024.

2. COMPATIBILITY OF CC LICENCES AND AI MODELS

The CDSM Directive[10] introduced two new exceptions for Text and Data Mining (TDM), defined (in art. 2.2) as "any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations":

1. The first exception in Article 3 concerns TDM for scientific research, which is limited to use by research organisations and cultural heritage institutions.
2. The second exception in Article 4 is not limited to any actor but is limited in the sense that rightsholders can expressly reserve the use (a so-called *opt-out*).

If a work can be used based on an exception or a limitation, this takes precedence over the requirements stipulated in the CC licences. This means AI developers can make use of CC-licensed material from the digital commons in three ways:

1. If they are (working on behalf of) research organisations or cultural heritage institutions, they can use the material based on the CDSM Directive's Art. 3.
2. If they are commercial or non-research developers, they can use the material based on CDSM Directive's Art. 4, as long as the creators (such as Wikipedia editors or contributors to e.g. Wikimedia Commons or Flickr) have not expressly reserved the use.
3. Anyone can use the material as long as they fulfil the requirements in the CC licences.

For the TDM exceptions to be applicable, the provisions require that the beneficiary has "lawful access" to the works used, although the term "lawful access" remains largely unexplored under EU law.[11] Some clarifications are given in the recitals of the CDSM Directive. Recital 10 reiterates that exceptions and limitations to copyright are not adapted to modern technologies, especially not in the field of scientific research, and that terms of licences in subscriptions or open access licences can exclude many works from TDM. Recital 14 of the same directive states that content is lawfully accessed when it is accessed through a subscription, based on an open access policy, or freely available online (i.e. for web scraping), allowing TDM for research purposes.[12] Web scraping, such as of works in the Digital Commons, is thus permitted for cultural heritage institutions and research organisations, and for other purposes if the data was lawfully acquired and the rightsholder has not prohibited the use.[13] In the case of the Digital Commons, most works are both open access and freely available online, meaning that use for non-research purposes is limited to the extent stated in the open access licences used.

There are still many potential cases where the TDM exceptions are not applicable; this might be because the user is not a research organisation or a cultural heritage institution because the use is commercial (in most cases excluding use under art. 3),[14] or because rightsholders have expressly reserved the use under art. 4.3. Works in the Digital Commons, licensed under a CC licence, can however still be used for AI training, to the extent permitted under the conditions of the licence.

Creative Commons offers a set of different licences, with four elements:

- Attribution (BY)
- Non-commercial (NC)
- No derivative works (ND)
- Share alike (SA).[15]

These elements can be combined into six different licences, from least to most restrictive:[16]

CC BY	Attribution		
CC BY-SA	Attribution	Share-Alike	
CC BY-NC	Attribution	No commercial use	
CC BY-NC-SA	Attribution	No commercial use	Share-Alike
CC BY-ND	Attribution	No derivatives	
CC BY-NC-ND	Attribution	No commercial use	No derivatives

[10] 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.

[11] TE Synodinou, 'Who Is a Lawful User in European Copyright Law? From a Variable Geometry to a Taxonomy of Lawful Use' In: TE Synodinou, TE., P Jougleux, C Markou, T Prastitou (eds) *EU Internet Law in the Digital Era*. (Springer, Cham, 2019). https://doi.org/10.1007/978-3-030-25579-4_2.

[12] M Bottis, M Papadopoulos, C Zampakolas, and P Ganatsiou, 'Text and Data Mining in Directive 2019/790/EU Enhancing WebHarvesting and Web-Archiving in Libraries and Archives' (2018) (9) *Open Journal of Philosophy*, <https://doi.org/10.4236/ojpp.2019.93024> accessed 17 October 2024.

[13] Chiara Gallese, 'Web scraping and Generative Models training in the Directive 790/19' (2023) 16(2) *i-lex* <https://i-lex.unibo.it/article/view/18871> accessed 17 October 2024.

[14] All commercial use is not outlawed. Recitals 11 and 12 of the CDSM Directive says that if there is a commercial actor involved, such as in a public-private partnership with a research organisation, this actor should not have preferential access to the results of the research.

[15] Kim Minjeong, 'The Creative Commons and Copyright Protection in the Digital Era: Uses of Creative Commons Licenses' (2007) 13(1) *Journal of Computer-Mediated Communication* <https://doi.org/10.1111/j.1083-6101.2007.00392.x> accessed 17 October 2024.

[16] For a delineation of all Creative Commons licenses, see Creative Commons, 'CC Licenses' (*Creative Commons*) <https://creativecommons.org/share-your-work/cclicenses/> accessed 17 October 2024. Text on Wikipedia is CC BY-SA 4.0.

Creative Commons also offers a mark to waive all rights permissible under copyright law, CC0.[17]

There are several unresolved questions when it comes to using CC licences for AI development. One fundamental question is which uses fall under the restrictions and why. CC licences are broadly concerned with the sharing and adaptation of works.[18] Sharing, in the legal code of Creative Commons, is defined as:

> to provide material to the public by any means or process that requires permission under the Licensed Rights, such as reproduction, public display, public performance, distribution, dissemination, communication, or importation, and to make material available to the public including in ways that members of the public may access the material from a place and at a time individually chosen by them.[19]

The relevant question is if all acts of TDM, where text or content from a publicly available source is ingested into an AI model, constitute an act of reproduction. Recital 9 of the CDSM directive explicitly states that:

> There can also be instances of text and data mining that do not involve acts of reproduction or where the reproductions made fall under the mandatory exception for temporary acts of reproduction provided for in Article 5(1) of Directive 2001/29/EC, which should continue to apply to text and data mining techniques that do not involve the making of copies beyond the scope of that exception.

All uses of works do accordingly not fall under the licence restrictions, and if TDM is used in a way that does not constitute an act of reproduction, then usage of CC-licensed material would likely not cause an infringement.[20]

It is also not ascertained that AI models create derivative works based on the input. As Daniel Gervais argues (in an analysis of derivative works under US law), derivative works and adapted material "is situated in a zone between (and occasionally 'beyond') reproduction, on the one hand, and uses that are inspired by, but not infringing (because they are not 'based upon').[21] While this article does not aim to discuss the nature of derivation and adaptation, it is apparent from legal literature that the usage of CC material in an AI model does not necessarily amount to reproduction or adaptation. If, or in the cases, it does not, then no infringement is taking place.

On the other hand, in cases where using such content amounts to reproduction or adaptation, there are still possibilities under some of the CC licences to use the CC-licensed content for AI training.

Each element impacts the possibility of using content when not explicitly permitted by law but in different ways. The attribution requirement partly reflects the fact that many jurisdictions, especially civil law countries, see attribution as an inalienable moral right.[22] It has been noted that the legal literature on artificial intelligence and moral rights has been much less prominent than on artificial intelligence and economic rights. Moral rights are, in contrast to economic rights, not harmonised in the European Union, leaving the legal landscape fragmented, though the right to be attributed is reflected in several of the exceptions and limitations introduced through the 2001 Infosoc Directive[23] (attribution is however not a condition for articles 3 and 4 of the CDSM Directive).[24] The AI Act,[25] passed in 2024, requires providers of foundation models to make a "sufficiently detailed summary" of the content used for training of the model publicly available, in accordance with a template provided by the AI Office. It is yet to be seen how this requirement will come into effect, and if sources provided accordingly will amount to the attribution requirement of CC licences. If no attribution is given to the content used, and a connection can be identified between the output of the model and the input data, then it would likely amount to a breach of the terms of the CC licence, in turn amounting to copyright infringement. One example of when that could be the case is if a GPAI model is used to translate a work protected by copyright, creating a derivative work, and the output fails to provide attribution to the original work in question.[26] Consequently, CC BY material can be used to

[17] I Hrynaszkiewicz and MJ Cockerill, 'Open By Default: A Proposed Copyright License and Waiver Agreement for Open Access Research and Data in Peer-reviewed Journals (2012) 5(494) *BMC Re*s Notes <https://bmcresnotes.biomedcentral.com/articles/10.1186/1756-0500-5-494> accessed 17 October 2024.

[18] G Hagedorn, D Mietchen, RA Morris, D Agosti, L Penev, W Berendsohn, D Hobern, 'Creative Commons Licenses and the Non-Commercial Condition: Implications for the Re-use of Biodiversity Information' (2011) 150 *Zoo*Keys <https://zookeys.pensoft.net/articles.php?id=3036> accessed 17 October 2024.

[19] Creative Commons, 'CC BY NC 4.0 Legal Code' (Creative Commons) <https://creativecommons.org/licenses/by-nc/4.0/deed.en> accessed 17 October 2024.

[20] Till Kreutzer, '*Open content: A practical guide to using Creative Commons licences*', German Commission for UNESCO (2014) <https://irights.info/wp-content/uploads/2014/11/Open_Content_A_Practical_Guide_to_Using_Open_Content_Licences_web.pdf> accessed 17 October 2024.

[21] D Gervais, 'AI Derivatives: the Application to the Derivative Work Right to Literary and Artistic Productions of AI Machines' (2022) 53 Seton Hall Law Review <https://ssrn.com/abstract=4022665> accessed 17 October 2024.

[22] Alexandra Giannopoulou, 'The Creative Commons Licences Through Moral Rights Provisions in French Law' (2014) 28(1) International Review of Law, Computers and Technology <https://www.tandfonline.com/doi/abs/10.1080/13600869.2013.869923> accessed 17 October 2024.

[23] 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.

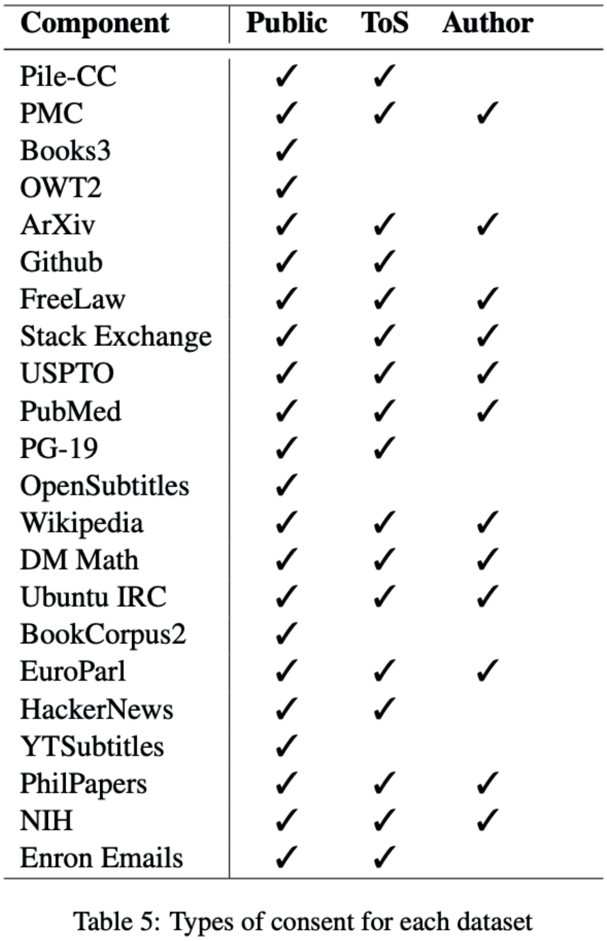
[24] M Miernicki and I Ng, 'Artificial Intelligence and Moral Rights (2021) 36 AI & Society <https://link.springer.com/article/10.1007/s00146-020-01027-6> accessed 17 October 2024.

[25] 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.

[26] D Gervais, N Shemtov, H Marmanis and C Zaller Rowland, 'The Heart of the Matter: Copyright, AI Training and LLMs' (2024) <https://papers.

train AI models if 1) it is used in a way not amounting to reproduction or adaptation or 2) the source is properly attributed, including the name and used CC licence.

One widely used data set for LLM development is The Pile, which includes Wikipedia as one of its 22 sources. Its developers claim to be aware of the complex legislative framework on copyright and TDM/AI development, but consider that their "use of copyright data is in compliance with US copyright law", not touching on compatibility with EU law.[27] The Pile includes over 800GB of copyrighted works scraped from legal or illegal sources (including 100GB of copyrighted books), in many cases without the author's knowledge and consent.



Component	Public	ToS	Author
Pile-CC	✓	✓	
PMC	✓	✓	✓
Books3	✓		
OWT2	✓		
ArXiv	✓	✓	✓
Github	✓	✓	
FreeLaw	✓	✓	✓
Stack Exchange	✓	✓	✓
USPTO	✓	✓	✓
PubMed	✓	✓	✓
PG-19	✓	✓	
OpenSubtitles	✓		
Wikipedia	✓	✓	✓
DM Math	✓	✓	✓
Ubuntu IRC	✓	✓	✓
BookCorpus2	✓		
EuroParl	✓	✓	✓
HackerNews	✓	✓	
YTSubtitles	✓		
PhilPapers	✓	✓	✓
NIH	✓	✓	✓
Enron Emails	✓	✓	

Table 5: Types of consent for each dataset

Table from Gao et. al., showing components of The Pile, and whether it is public data, allowed according to (their analysis of) the terms of use or with direct consent from the author. Gao et. al. licensed under CC BY 4.0.

The Pile is used by AI companies such as Anthropic, Nvidia, Apple, and Salesforce, and the dataset lists bare URLs as sources, potentially violating attribution and thus copyright requirements. Creators and researchers have had to use specially developed tools to search for additional metadata. It remains unclear whether such usage is legally in compliance with the attribution requirements in e.g. CC BY.

The share-alike (SA) element also opens up for AI training under certain conditions. Kacper Szkalej and Martin Senftleben provide a comprehensive overview of the SA requirement and its impact on AI training, arguing that what they call the CC community can "use copyright strategically to extend SA obligations to AI training results and AI output" by using rights reservation mechanisms, such as the opt-out system in Art. 4 of the CDSM Directive, to "subject the use of CC material in AI training to SA conditions".[28] In this way, they argue, a "tailor-made license solution" can be developed granting broad freedom for AI developers to use CC works while being forced to accept the share-alike obligations of the CC BY-SA license. In their proposal, this would be ensured via a chain of contractual obligations, where SA conditions are passed on via each step.

One challenge with such an approach would be to define who can actually make the legal case. Wikipedia text, for example, is licensed under CC BY-SA 4.0.[29] This means that all Wikipedia contributors retain the right to be attributed and it requires the text to be reused under the same license. The editors, however, remain the rightsholders. No rights are transferred to the Wikimedia Foundation. At the same time, art. 4 of the CDSM directive makes it clear that it is the rightsholder who has the right to expressly reserve the usage. A challenge for the approach proposed by Szkalej and Senftleben is to identify to what extent the community can act collaboratively to enforce the SA requirement.

A further challenge concerns the feasibility of opting out for individual files, e.g. if an individual user wants to prohibit the use of a Wikipedia article or a photo on Wikimedia Commons from AI training. Open Future puts forth some thoughts on how that could be done technically through unit-based rather than location-based identifiers, based on unique locations such as URLs.[30] For media content in the Digital Commons, a part of the solution might be the Commons Database project, a pilot project funded by the European Commission and developed by Liccium, Institute for Information Law at the University of Amsterdam, Europeana and Wikimedia Sverige. The pilot aims to build a database of unique media file identifiers alongside sourced rights information about the files,

ssrn.com/sol3/papers.cfm?abstract_id=4963711> accessed 17 October 2024.

[27] L Gao, S Biderman, S Black, L Golding, T Hoppe, C Foster, C Leahy, 'The Pile: An 800gb Dataset of Diverse Text for Language Modeling' (2020) <https://arxiv.org/abs/2101.00027> accessed 17 October 2024.

[28] Kacper Szkalej and Martin 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> accessed 17 October 2024.

[29] See Gregory Varnum, 'Licensing of content' under Terms of Use policy (Wikimedia Foundation, 30 March 2024), <https://foundation.wikimedia.org/wiki/Policy:Terms_of_Use#7._Licensing_of_Content>, accessed 21 November 2024.

[30] 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>.

including its copyright protection,[31] but the same system could potentially also be used to store information about opt-out reservations.

3. THE IMPACT OF THE DIGITAL COMMONS ON AI MODELS

Openness in AI development can refer to many things. Nick Bostrom has listed a number: "open source code, open science, open data, or to openness about safety techniques, capabilities, and organisational goals, or to a non-proprietary development regime generally."[32] All aspects, however, refer to the release into the public domain, rather than the (re)use of the public domain, which is mysteriously overlooked. The Digital Commons, including Wikipedia, is a sensitive ecosystem. The heavy traffic to Wikipedia pages has been channelled through Google's search engine, where Wikipedia pages have been prioritized compared to many other websites. This traffic has resulted in both donations and volunteers.[33] The introduction of Google Knowledge Graph, heavily relying on CC0 licensed data from Wikidata, has reduced the traffic to Wikipedia, and thereby also the understanding among web users of where the information originally comes from. McMahon et. al. warn of a 'death spiral', "in which a decrease in visitors leads to a decline in both overall edits and new editors, not to mention much-needed donations".[34] As Zachary J. McDowell and Matthew A. Vetter discuss,[35] Wikimedia projects are susceptible to large-scale commercial reuse by GPAI developers. They call the extraction, reappropriation, and commodification of Wikimedia content and data beyond the intent of its original creators a "re-alienation" of knowledge. Whereas the more permissive licences, especially the CC0 mark used by Wikidata, in their analysis limit the Commons, the share-alike requirement maintains and even enlarges the Commons. The death spiral described by McMahon et. al. could potentially lead to a negative spiral: GPAI developers use Wikipedia and other Digital Commons content to train their model, without properly attributing or compensating the source. This leads, according to the idea, to less traffic to the pages of the Digital Commons, and thereby fewer volunteers, donations, and ultimately new content. Less and less content in the Digital Commons, in turn, leads to worse and worse AI models, and a vicious cycle is born.[36]

At least two potential strategies among Digital Commons stakeholders could be envisioned to challenge this 'death spiral':

1. Stakeholders use the CC licences strategically, such as in the way described by Szkalej & Senftleben, to uphold the Commons and restrict large-scale commercial reuse by GPAI developers, to the detriment of the quality of AI models;[37]
2. Digital Commons Stakeholders increase collaboration to maintain the ecosystem of free knowledge, including on open access policies, applications for funding, and in conversations and negotiations with AI developers, to ensure the long-term sustainability of the digital Commons.

The latter strategy could involve collaborating with IGOs such as the United Nations, African Union, and European Union agencies, as well as national governments, to make sure that official documents, reports, and data feed into the digital Commons. Several UN agencies, including UNESCO, as well as the special Envoy on Technology,

[31] 'Project and Research Coordinator' (*Open Future*) <https://openfuture.eu/project-and-research-coordinator/>, accessed 22 November 2024.

[32] Nick Bostrom, 'Strategic implications of openness in AI development', in Roman V. Yampolskiy (ed), *Artificial intelligence safety and security* (Chapman and Hall/CRC 2018).

[33] ZJ McDowell, MA Vetter, 'Rethinking Artificial Intelligence: Algorithmic Bias and Ethical Issues| The Realienation of the Commons: Wikidata and the Ethics of "Free" Data.' (2023) 18 International Journal of Communication <https://ijoc.org/index.php/ijoc/article/view/20807> accessed 17 October 2024.

[34] C McMahon, I Johnson and B Hecht, 'The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies' (2017) 11(1) Proceedings of the International AAAI Conference on Web and Social Media <https://ojs.aaai.org/index.php/ICWSM/article/view/14883> accessed 17 October 2024.

[35] McDowell and Vetter (2023).

[36] The idea is similar to what Cory Doctorow has called 'enshittification'. See Cory Doctorow 'Social Quitting. Special Features' [2023] Locus 90(1).

[37] Szkalej and Senftleben (2024).

recognize the importance of open access and open source for positive digital transformation.[38] Along similar lines, Paul Keller analyzes in a blog post for Open Future the positive impact publicly available datasets developed by non-profit organizations, such as is the case with LAION (also published under open Creative Commons licenses[39]), could have on AI development. This positive impact includes allowing creators to see to what extent their works are used for AI development, to register opt-outs (per Art. 4 of the CDSM Directive), and allowing researchers to understand biases and problematic patterns in the dataset.[40]

The two named strategies can of course be combined, in the sense that a larger pool of stakeholders collaborate both to open up and disseminate more open-access content and data and to make sure that AI developers use this content in compliance with legislation or licenses. Several of these insights are also reflected in the objectives and paragraphs of the global digital compact, adopted by UN member states.[41] They also reflect an idea that was raised during two workshops with Wikimedia volunteers, namely that stakeholders in the digital Commons should work collaboratively to make sure that the conditions for reuse of the CC licenses are upheld.[42] McDowell and Vetter mention the role that Wikimedia Enterprise, a commercial service from the Wikimedia Foundation offering "'Enterprise-grade APIs Built for Search, Social, and Voice Assistants' [...] to data and information in Wikimedia's products", could play in safeguarding the ecosystem of Wikimedia platforms.[43] These examples attempt to show that combining the two strategies in order to uphold the Digital Commons will also require a plethora of means and initiatives.

4. CONCLUSION

In a response to the US Copyright Office, the Wikimedia Foundation (WMF) stated that Wikimedia projects play an important role in relation to AI since machine learning and AI technology help support the quality of the Wikimedia projects and make the work of the editors more efficient, but also since Wikimedia content "forms one of the most important bases for training generative AI programs."[44] Meanwhile, WMF infers that some AI developers are out of compliance with both the attribution and share-alike clauses, and while WMF supports the use of Wikimedia content for AI training, they encourage reuse to comply with the licenses and for reusers to release the models they develop under open licenses too.[45]

This article argues along similar lines, showing how the CC-licensed material on the Wikimedia platforms and in other Digital Commons repositories can be used for AI model development and still comply with the requirements of the licenses. It remains unclear to what extent AI developers are obliged to comply with the CC licenses, but as the analysis shows, there are cases where AI development falls outside the scope of the two new TDM provisions in EU law, and in such cases, failure to comply with the licenses could amount to copyright infringement. At the same time, the analysis shows the important role that the Digital Commons can play in combatting disinformation and misinformation through AI models, and that open access and open licensing such as through CC licenses can be an efficient way of improving the output of generative AI models. The stakeholders of the Digital Commons could collaborate between themselves and with AI developers to explore ways how to use open access strategically to promote high-quality AI models while maintaining the integrity of the CC licenses and open access.

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.

[38] See e.g. Office of the Secretary-General's Envoy on Technology, 'Open Source Digital Transformation' (UN, 9 July 2024) <https://www.un.org/techenvoy/content/open-source-digital-transformation> accessed 17 October 2024.

[39] C Schuhmann, 'LAION-400-MILLION Open Dataset', (LAION, 20 August 2021), <https://laion.ai/blog/laion-400-open-dataset/>, accessed 25 November 2024.

[40] P Keller, 'LAION vs Kneschke' (Open Future, 10 October 2024), <https://openfuture.eu/blog/laion-vs-kneschke/>, accessed 25 November 2024.

[41] UN Global Digital Compact 2024 <https://www.un.org/global-digital-compact/sites/default/files/2024-09/Global%20Digital%20Compact%20-%20English_0.pdf> accessed 17 October 2024.

[42] Insights from these workshops are to be published.

[43] McDowell & Vetter (2023).

[44] 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>.

[45] Wikimedia Foundation (2023).