Journal of Open Humanities Data

A Position Paper on AI and Copyrights in Cultural Heritage and Research (EU and UK)

DISCUSSION PAPER

JÖRG LEHMANN
ANNA-MARIA SICHANI
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

The long-term, highly divisive discussion between the entertainment industry and the AI sector around the use of copyrighted material for the training of machine learning applications has gained traction with the recent advent of generative AI systems. However, in this debate, the voice of cultural heritage institutions and research libraries is hardly considered. Major publishing houses have begun to reserve the right to perform text and data mining (TDM) on digital assets licensed by them, which is relevant for the development of machine learning applications. Such reservations impair the access of cultural heritage institutions (CHIs) and research libraries' patrons, such as start-ups, small and medium-sized enterprises and companies belonging to the cultural sector. In this discussion paper, we propose a set of differentiated solutions which potentially open up ways in which digital assets which are under copyright may be used for TDM. These solutions consist in the development of new licences, the use of a technical protocol, and in the provision of infrastructures which allow for performing TDM on copyrighted digital assets differentiated according to data providers, data consumers and to the tasks for which machine learning applications are being trained. These proposals aim at furthering the debate on open data – keeping them as open as possible while respecting the interests of right holders and data consumers alike.

CORRESPONDING AUTHOR:
Jörg Lehmann
Staatsbibliothek zu Berlin – Berlin State Library, Berlin, Germany
joerg.lehmann@sbb.spk-berlin.de

KEYWORDS:
copyright; open data; generative artificial intelligence; machine learning; cultural heritage institutions; research libraries

TO CITE THIS ARTICLE:
Lehmann, J., & Sichani, A.-M. (2025). A Position Paper on AI and Copyrights in Cultural Heritage and Research (EU and UK). *Journal of Open Humanities Data*, 11: 25, pp. 1–9. DOI: https://doi.org/10.5334/johd.290

DOI: 10.5334/johd.290

(1) CONTEXT AND MOTIVATION

The advent of generative AI (genAI) systems and the publicity around large language models (LLMs) such as ChatGPT have not only sparked discussions around the meaning of concepts such as "creativity", "originality" and "ownership", but also on the data used to train such AI applications – especially where copyrights and related rights apply. The vast majority of digitised historical collections provided by CHIs and research libraries online are in the public domain as they mostly predate the 1920s. They are openly accessible, usually available through open licences, and this achievement should not be questioned. Already these assets are of high value to AI companies. Digitised books, for example, contain syntactically and lexically correct texts as opposed to what can usually be found on (and scraped from) the Internet. Copyright protected digital assets, on the other hand, are often left outside the digitisation planning of CHIs as their reproduction and (re)use come with legal barriers. However, within the European Union, CHIs are entitled to digitise works that are permanently in their collections (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, 2019, art. 6), irrespective of whether this work still falls under copyright restrictions or not (Janssens et al., 2021, p. 57). More recently, for example in Germany, funding agencies have launched calls to evaluate the possibility to digitise objects which are (still) copyright protected in pilot projects (German Research Foundation, 2024); such objects should be provided in interoperable presentation environments with restricted access. Both categories of assets, public domain digitised assets and copyright protected digital assets, are of high value for the AI industry; especially copyrighted material significantly improves the quality of trained machine learning models (de la Rosa et al., 2024). Against this backdrop, it should be noted that only last year, the KB – National Library of the Netherlands has "restrict[ed] access of commercial AI to KB collections" (Kleppe, 2024).

Within the European Union, the Directive on Copyright in the Digital Single Market, or CDSM Directive for short (Directive (EU) 2019/790) provides a modernised copyright framework. This Directive, which entered into force on April 19th, 2019, introduces exceptions important in the context discussed here and lays, with regard to TDM, the legal foundation for the AI Act which was passed in 2024. The new exceptions included in Title II of the CDSM Directive are mandatory for Member States and were therefore subsequently transposed into national laws. They are introduced in Articles 3 and 4 respectively of the CDSM Directive. Article 3 allows "for reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access". Research organisations and CHIs remain the only beneficiaries of the exception; for them, TDM activities may be solely carried out within the framework of research activities and public private partnerships and on works or other subject matter to which these institutions have lawful access. The CDSM Directive thus explicitly allows research organisations as well as CHIs to develop secondary products from their digital assets on their own, for example to train machine learning models, provide parallel corpora for machine translation, or for the development of genAI applications. Article 4 instructs Member States to provide for the exception formulated in Article 3 and states that 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 right holders in an appropriate manner, such as machine-readable means in the case of content made publicly available online"; it thus formulates the right to opt-out.

In the UK, a copyright exception for TDM was introduced in 2014 under sec. 29A of the Copyright, Designs and Patents Act 1988 (HM Government, 1988 as amended) for the "sole purpose of research for a non-commercial purpose" of copyright works to which a person already has lawful access. Where third parties or private companies hold the rights on the digitisation of cultural assets and/or the digital records themselves, CHIs and research libraries may need to negotiate contracts with them to ensure lawful access and be able to perform TDM (Ahnert al., 2023). In a recent article, Westenberger and Farmaki (2025) argue that the current TDM exception in sec. 29A is not fit for purpose for non-commercial research, particularly in heritage contexts, presenting a series of reasons such as the unclear concept of 'lawful access' and whether it includes web crawling/mining, the prohibition of transferring copies among

DOI: 10.5334/johd.290

parties and the existing problematic copyright status of cultural heritage datasets, arguing that the exception should be further clarified and expanded.

The UK government's recent Copyright and AI consultation, closed on 25 February 2025 (Secretary of State for Science, Innovation and Technology by Command of His Majesty, 2024), proposes a new TDM exception for non-commercial use with an opt-out for rights holders, supported by enhanced transparency measures. The new exception's impact remains uncertain, as the existing TDM framework already challenges heritage and research. Key issues, including rights reservations, compensation for rights holders, and AI transparency, are still under review. This consultation builds on past efforts, including the UK IPO consultation (Intellectual Property Office of HM Government, 2022) and a 24 October 2024 amendment to the Data (Use and Access) Bill, which tasked the Information Commissioner's Office (ICO) to regulate web crawler and AI systems' transparency (HM Government, 2025).

Finally, the recently introduced UK AI Opportunities Action Plan (Secretary of State for Science, Innovation and Technology by Command of His Majesty, 2025) proposes establishing a copyright-cleared public cultural asset training dataset through partnerships with heritage institutions alongside the development of a National Data Library as a digital infrastructure to ensure the development of a coherent data collection, access policies and exchange services to advance responsible AI development.

The development of capable genAI applications in the recent past has worked as a catalyst in the long-time controversy between right holders of digital assets on the one hand (i.e. the entertainment industry consisting of publishing houses, creative industries, individuals, and others) and the AI sector on the other hand. The crucial point is the right to perform TDM on digital assets, which is relevant for all machine learning applications, and this right becomes problematic especially with regard to foundational models developed outside the European Union and the United Kingdom. Major publishing houses like Elsevier, Springer, Wiley and Taylor & Francis holding copyright protected content as well as open access material have already started to reserve exclusive rights to perform TDM in their licence agreements reached with research libraries (Wiley-VCH GmbH et al., 2023, para 5.5). As this example shows, the interest of such publishing houses lies with training AI applications themselves and providing their customers with AI services which are subject to a fee (see also Elsevier & MPDL Services gGmbH, 2023; Springer Nature Customer Service Center GmbH & MPDL Services gGmbH, 2023). Even though the use of such digital assets is always possible for research purposes since the TDM exception "cannot be overridden by contract" (Geiger et al., 2019, p. 28), the result of such licence agreements is that CHIs and research libraries are not only struggling to offer transparency around TDM and reuse of their publicly available datasets to their patrons, but also their role as access brokers and knowledge enablers is impaired: they are no longer able to perform TDM themselves and unable to serve their patrons, such as start-ups, small and medium-sized enterprises (SMEs) and companies belonging to the cultural sector. With copyright protected content having become a contested resource, it is obvious that right holders aim for remuneration if AI companies want to use their content for TDM (Geiger & Iaia, 2024), and the same holds for collective management organisations (CMOs), who are given the mandate to licence their members' works on a blanket basis. These developments create a risk for CHIs and research libraries in that they will either have to remunerate right holders, which is a costly endeavour, or that licence agreements will exclude the right to perform TDM for CHIs, research libraries and their patrons.

In this situation, we wish to propose a set of solutions mediating between the interests of CHIs and research libraries, right holders (especially small and medium-sized publishing houses, but also the entertainment industry in general), and the AI sector including big tech as regards access to and reuse of licensed and copyright protected digital assets. To be fully clear in this respect: CHIs and research libraries are not in the position to define who should have access and to decide to impose restrictions on access; rather, they are bound to technically implement the right holder's terms and conditions (e.g. of CMOs). Furthermore, CHIs and research libraries are not entitled to make decisions on a TDM opt-out since this is the task of the right holders. However, in their role as access brokers and knowledge enablers, they are able to mediate between different interests and make suggestions that serve the purpose of open access, e.g. by proposing the creation of new licences.

DOI: 10.5334/johd.290

At the centre of our proposal is the creation of a licence allowing TDM on the condition that it is performed exclusively on the territory of the European Union and the United Kingdom respectively ("TDM yes, but only within EU/UK territory"). Such a licence applies only to cases where access is restricted and it builds on the grant of full transparency with regard to training data according to the EU AI Act that has entered into force in August 2024 (Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial (Artificial Intelligence Act), 2024). Technically, the machine-readable protocol as well as the infrastructure/operational framework we propose to express and support such a licence apply wherever access to and use of digital assets protected by copyright and related rights is restricted (with authentication as a minimum requirement) in trusted, interoperable presentation environments, or in repositories to which access is IP-controlled, or in the case of data spaces where CHIs and research libraries are able to define and control access and terms of use of the content provided by them. Ethically and politically, these solutions link the concepts of data spaces and data sovereignty to the idea of a commons. "Commons" designates a shared resource that is managed by a community for the benefit of its members (Lehmann, 2022, 2023). In this sense, the European Data Space for Cultural Heritage is explicitly conceptualised as "an open and resilient commons for the users of European cultural data, where data owners – as opposed to platforms – have control of their data and of how, when and with whom it is shared" (Davies et al., 2023).

(2) THE LICENCES

Licences are grants of rights issued by right holders to someone who wants to make use of the works covered by these rights. While we acknowledge that CHIs and libraries are not in the position to issue licences, we can definitely propose the creation of new ones. Interestingly, so far, no licence has been issued that reserves the right to opt-out of TDM and to be used with datasets published e.g. in interoperable presentation environments or in access-controlled repositories. This fact, alongside the lack of a robust machine-readable way to express it, has been a major hindrance in the implementation of the TDM exception and clearly identifies a desideratum.

Beyond this gap to be filled, we propose a new licence *allowing TDM on copyrighted material on the condition that it is performed exclusively on the territory of the European Union and the United Kingdom respectively* ("TDM yes, but on EU/UK territory"). This new licence could complement the set of standardised right statements provided by Rightsstatements.org, an international member-based consortium of CHIs and research libraries (RightsStatements.Org, 2022). It requires conformance to the transparency regulations laid out in the EU AI Act (Regulation (EU) 2024/1689, Arts. 11(1), 53(1), Annexes IV, XI, XII) as well as in the General-Purpose AI Code of Practice and, for the UK, in the Data (Use and Access) Bill 179 2024–25, sec. 135–139 (HM Government, 2025). It would be deployed within technical infrastructures that restrict access in order to be capable to proof who has accessed the content under question.

(3) THE TECHNICAL PROTOCOL

The licences we propose complement the way CHIs are currently enabling access and reuse of their digital collections and the implications this has for TDM uses, especially by commercial AI companies. The API most commonly used by CHIs is OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting). Initially, it was conceived as an interface to exchange metadata in interoperable formats between different institutions. Nowadays, OAI-PMH is the standard solution by which CHIs provide open access to the metadata for content hosted in their repositories. Licences and rights statements can be included in the metadata. OAI-PMH is currently being complemented by the International Image Interoperability Framework (IIIF) which also enables programmatic access to the content itself in addition to metadata, which is not the case with OAI-PMH. However, as the adjective "open" signals, the OAI-PMH is not designed to differentiate between users and was not foreseen to mask specific material from access; even if a restrictive licence is enclosed as part of the metadata, this may not prevent the assets to be copied. These circumstances can lead to a complicated situation if a bot scrapes content which might violate arrangements laid down in the licence agreement between, for example, a library and a CMO. We distance ourselves from the approach to restrict specific bots

DOI: 10.5334/johd.290

via the robots.txt, because this may also exclude access to digital assets which are in the public domain. What is in the public domain should remain in the public domain and should always be provided in open access.

As an alternative, we propose – as one of the obvious solutions to be used with OAI-PMH – to implement a protocol named "Text and Data Mining Reservation Protocol" or TDMRep for short, developed by a World Wide Web Consortium (W3C) community working group (Text and Data Mining Reservation Protocol Community Group, 2024). This specification defines a simple and practical web protocol which is capable of expressing the reservation of rights relative to TDM. Most interestingly, this specification is conceived of in a non-binary way. Beyond the provision of the right to opt-out (indication that TDM rights are reserved or TDM rights are not reserved), an option is provided to point to a TDM Policy. Technically, the latter option consists of an URL pointing to a TDM Policy set given by the right holder. 'Policies' are conceived as machine-readable structures referenced from the TDM Policy property defined in the specification. TDM actors are enabled to contact right holders, and the policy set offers details about available TDM licences. This solution can be implemented via the same infrastructure that lies behind OAI-PMH. The API provides information about available records as well as individual assets (images, full texts etc.), the latter of which are provided by the content server. The information about rights are usually stored in METS/MODS. One of the three options developed in TDMRep (TDM no, TDM yes, TDM Policy set) should be contained in the METS/MODS file. If a request for a METS/MODS file is received, the file is delivered containing machine-readable information, thereby allowing recipients of that declaration to adjust their scraping behaviour, or to reach a separate agreement with the right holder that satisfies all parties. This solution offers compatibility with existing CHI infrastructures, can be applied to diverse media, allows for the use of existing licences contained in the TDM Policy set, and it is machine readable. Wherever IIIF is deployed, information about usage rights can be included in machine-readable form in the manifest files holding the metadata (IIIF Presentation API 3.0, 2023).

However, there is a major constraint regarding the use of TDMRep which lies in the possibility that US-based companies and corporations may scrape the contents of EU- and UK-based CHIs and research libraries which are available via OAI-PMH. Such commercial actors may use these digital assets as training data irrespective of whether TDM rights are reserved or not and perform the training in the US and thus outside of the EU while claiming to make "fair use" of the material. This is a likely scenario, and even if it remains to be seen how jurisdiction in the USA pronounces sentences regarding the fair use of copyright protected training data within their legal framework (Karamolegkou et al., 2023; Samberg, 2024), there is a real chance that right holders will not be able to enforce their reservation of the right to opt-out.

(4) THE INFRASTRUCTURAL FRAMEWORK

All that has been explained above indicates that open solutions no longer provide viable ways of providing licensed and copyrighted digital assets for TDM. The idea of "open" therefore needs to be reconceptualised. Already in 2021, Paul Keller and Alek Tarkowski underlined in their essay entitled "The Paradox of Open" that "open resources are most likely to contribute to the power of those with the best means to make use of them" (Keller & Tarkowski, 2021, p. 5), and they conclude that the "only sensible action is to invest in alternatives to these platforms" (ibid., 4). We identified three such currently existing alternative infrastructures, all of which realise or support restricted access to licensed and copyright protected digital assets: 1) Establishing interoperable presentation environments where authentication is necessary for access; 2) using repositories to which access is IP-controlled; and 3) setting up data spaces. As for the first and second solutions, while interoperable presentation environments do not necessarily open up the possibility to perform TDM, licensed and copyright protected contents provided via IP-controlled repositories need to be accompanied by licences (or rights statements). To be fully clear with respect to the digital assets made available in the three named infrastructures: While all three may contain digital assets under copyright which were digitised under Article 3 of the CDSM Directive and sec. 29A of the CDPA (see above), this implies that such content is available solely for research. Beyond that, these infrastructures may provide copyright protected digital assets available for commercial use if CHIs and research libraries have negotiated license agreements directly with right holders that include the possibility to perform TDM.

DOI: 10.5334/johd.290

Regarding the third solution – data spaces –, it is best to illustrate the options which are available with a feasibility study conducted by Berlin State Library. The proof-of-concept aimed at exploring the possibilities of offering copyright protected data with secure access, providing them for compute-to-data and in general to enable handling copyright protected digital assets within a data space. The feasibility study was based on Gaia-X, an initiative to develop a federated secure data infrastructure for Europe (Gragert, 2023). Gaia-X grants full sovereignty over data, in this case licensed by CHIs; full control over the data always remains with the owner. It is a decentralised system, meaning that there is no central point in the network that is necessary for it to function, and where data can even remain on premise of the participating institutions. A set of rules to create trust and special services (available as open source) can be used by all network participants to check whether others are also complying with the rules. Gaia-X is therefore a federated system with implicit trust. It is interoperable as there is no prescribed technical infrastructure for participation, where the Pontus-X-Network uses the Ocean Protocol; furthermore, it is possible to create subject-specific data spaces and link them together. Algorithms can be sent to the data, and derivative products may be downloaded (Berlin State Library (SBB), 2024).

Therefore, it is possible to perform TDM within the data space. It is possible to implement access control and effectively handle licences; high granularity of access and use is possible. If right holders aim at remuneration, this can also be handled within the data space. Registered users like holders of a library card and researchers can even get access to such data, because they receive a wallet containing credits which can be spent in exchange for access. In this sense, such a data space could function as a commons: publishing houses, for example, can demand fees for their data (and might receive the larger part of the remuneration), while data consumers can be asked to contribute to the maintenance of the data space. This option conforms with the Data Governance Act (Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European Data Governance and Amending Regulation (EU) 2018/1724 (Data Governance Act) [2022], 2022, p. chap. 2, art. 6), saying that public sector bodies can demand fees which have to be oriented at the costs of the infrastructure to provide and enable reuse of the data. Moreover, a possibility for differentiation within the framework of commercial users may be created, in the sense that the digital assets may freely be used for innovation and research by start-ups, small and medium-sizes enterprises (SMEs) and companies belonging to the cultural sector, but for larger AI companies or for big tech the use of such data may only be possible if they pay fees. This tiered approach is foreseen in EU legislation; however, it might be necessary to define which tier a company belongs to, or even if an organisation has to be classified as a for-profit company or a non-profit organisation (European Union Intellectual Property Office, 2024). If data spaces are conceptualised as a commons, the role of CHIs and research libraries as access brokers and knowledge enablers becomes tangible: if right holders do not wish their data to be openly available for commercial AI companies, the former are able to negotiate differentiations in the use of such data. Moreover, data spaces are attractive for right holders who do not dispose of such infrastructures, e.g. small and medium-sized publishing houses, because remunerations can be processed therein.

As is the case with other data spaces like the European Health Data Space (EHDS), the UK could participate in such an EU data space on the basis of memorandums of understanding or international agreements within the framework of the EU-UK Trade and Cooperation Agreement TCA (European Commission, 2021). This data space would be attractive for EU and UK rights holders who do not dispose of such infrastructures, e.g. small and medium-sized publishing houses. They would profit from the technical infrastructure, because remunerations can be processed collectively therein, would profit from the transparency regulations laid out in the EU AI Act, and would be able to enforce existing regulations and directives via institutions such as the EU AI Office or to bring possible copyright infringements to court. Finally, this data space can operate as part or as a functionality of the UK National Data Library, as discussed above.

As with all technical solutions, there is a downside: such data spaces are usually complex and difficult to manage, which entails an obstacle especially for smaller CHIs and often results in the need for additional human resources. In this respect, the feasibility study conducted by Berlin State Library notes amongst the lessons learned, that while onboarding of CHIs and research libraries has to be regarded as a challenge, the publication of datasets is comparably

DOI: 10.5334/johd.290

easy. This insight shows that any institution aiming at setting up and maintaining a data space in the way described above should be well prepared to cope with this task.

(5) CONCLUSIONS

As digital assets from the recent past are mostly copyright protected but at the same time of high value for the AI industry, we propose the creation of a new licence "TDM yes, but on EU/UK territory" to be preferably used within the technical environment described above. Though this proposal is technically demanding, we wish to argue that there are clear benefits for all sides involved: CHIs and research libraries would be able to carry on providing responsible access and reuse of their licensed and copyright protected digital assets while ensuring transparency in their rights management as well as in cases of TDM exceptions. Small and medium-sized publishing houses as well as parts of the entertainment industry would profit from providing their content in such a data space. On the one hand they would be able to use an infrastructure which is capable of protecting their contents from unwanted downloads and/or which enables performing TDM within the data space, while on the other hand enabling them to process remuneration in a differentiated way, i.e. according to a variety of data consumers. For the AI industry, copyright protected digital assets would become available without the need to directly negotiate with each and every right holder. If these assets are not available for download, TDM may be performed within the data space and derivative products may be downloaded. This could also provide an incentive for AI companies to contribute to the maintenance of the data space. The maintainers of the data space would participate in the fees to be paid by data consumers, in the sense that the vast majority of the fees would go to the right holders and, in the sense of a commons, a small portion would be used for the maintenance. The licence we propose goes hand in hand with the technical options implemented in data spaces, and the current legal framework in the EU guarantees enforceability as well as transparency of what is being done with copyright protected digital assets being used as training data. Such a licence does not exclude any of the stakeholders involved. With regard to the AI industry, and depending on how the infrastructures are set up, it may be necessary for commercial AI companies to establish a subsidiary within the EU or the UK, which is already the case for many.

In the current, highly divisive controversy between the entertainment industry and the AI sector, there is a lot to gain if a differentiated approach is chosen. It is wise to introduce differentiations of the right to perform TDM according to the specific tasks on which a machine learning application is trained. While copyright protected digital assets might not be provided for the development of genAI applications or only in exchange for remuneration, TDM might be allowed for a broad range of machine learning applications without much ado. Similarly, it might be very helpful if TDMRep would allow the use of a granular vocabulary according to the possible uses which might be made of the material.

While we indicate with these hints and with this position paper in general that a lot of work needs to be done in the near future, we would welcome it if our proposal would be mediating between right holders opposing creators of genAI applications – but who are generally open to allowing TDM on the one hand – and the AI industry including big tech on the other hand, and we welcome it if the role and voice of CHIs and research libraries would gain more impact in the current controversy. More generally, we think that the advent of genAI forces us to rethink and differentiate principles of Open Access. If our proposals are read as a movement from the open to the closed, we counter that the minimum requirement for receiving access to technical infrastructures – authentication – seems to be the equivalent of the 21st century to holding a library card in the 20th century. Asking for authentication and then imposing a tiered system depending on the roles and the status of data consumers and the differentiations imposed by data providers may not change much for the individual user or researcher. However, it implies a departure from a conception of the Internet where any content is freely and openly accessible, a notion from which creators have begun to take leave since long ago.

ACKNOWLEDGEMENTS

The authors thank all persons who commented and thus contributed intellectually to a longer version of this article, even if they did not contribute to the text itself. These are, in alphabetical

order: Steven Claeyssens, Neil Fitzgerald, Paul Keller, Marco Klindt, Matthew Lambert, Ariadna Matas, Benjamin White, Alexander Winkler.

The authors acknowledge support by the Open Access Publication Fund of the Prussian Cultural Heritage Foundation.

FUNDING STATEMENT

The work of Jörg Lehmann was funded by the Federal Government Commissioner for Culture and the Media (BKM), project grant no. 2522DIG002. The work of Anna-Maria Sichani was support by the Bridging Responsible AI Divides (BRAID) programme with funds from the Arts and Humanities Research Council [grant number AH/X007146/1].

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Conceptualisation – all authors; Writing – original draft: Jörg Lehmann; Writing- review and editing – all authors.

AUTHOR AFFILIATIONS

Jörg Lehmann orcid.org/0000-0003-1334-9693
Staatsbibliothek zu Berlin – Berlin State Library, Berlin, Germany

Anna-Maria Sichani orcid.org/0000-0001-6778-4610
School of Advanced Study, University of London, London, United Kingdom

REFERENCES

Ahnert, R., Griffin, E., Ridge, M., & **Tolfo, G.** (2023). Collaborative historical research in the age of big data: Lessons from an interdisciplinary project. *Elements in Historical Theory and Practice*. https://doi.org/10.1017/9781009175548

Berlin State Library (SBB). (2024, March 20). *Berlin State Library Portal Demonstrator*. https://sbb.pontus-x.eu/

Davies, R., Leroi, M.-V., & **Verwayen, H.** (2023, November 6). *The data space and the Collaborative Cloud: Cooperation for mutual and collective benefit*. Europeana PRO. https://pro.europeana.eu/post/the-data-space-and-the-collaborative-cloud-cooperation-for-mutual-and-collective-benefit

de la Rosa, J., Mikhailov, S. A., Zhang, L., Wetjen, F., Samuel, D., Liu, P., Braaten, R., Mæhlum, P., Birkenes, M. B., Kutuzov, A., Enstad, T., Farsethås, H. C., Brygfjeld, S. A., Gulla, J. A., Oepen, S., Velldal, E., Østgulen, W., Øvrelid, L., & **Myhre, A. S.** (2024). *The Impact of Copyrighted Material on Large Language Models: A Norwegian Perspective.* https://doi.org/10.48550/arXiv.2412.09460

Directive 2019/790/EU 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, 2019. https://eur-lex.europa.eu/eli/dir/2019/790/oj#004

Elsevier, B. V., & **MPDL Services gGmbH.** (2023). *Projekt DEAL – Elsevier Publish and Read Agreement* (Version 1, p. 1142055). https://doi.org/10.17617/2.3523659

European Commission. (2021). *The EU-UK Trade and Cooperation Agreement.* https://commission.europa.eu/strategy-and-policy/relations-united-kingdom/eu-uk-trade-and-cooperation-agreement_en

European Union Intellectual Property Office. (2024, September 27). *Germany—Hamburg District Court, 310 O.22723, LAION v Robert Kneschke*. EUIPO. https://www.euipo.europa.eu/en/law/recent-case-law/germany-hamburg-district-court-310-o-22723-laion-v-robert-kneschke

Geiger, C., Frosio, G., & **Bulayenko, O.** (2019). *Text and data mining: Articles 3 and 4 of the Directive* 2019/790/Eu. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3470653

Geiger, C., & **Iaia, V.** (2024). The forgotten creator: Towards a statutory remuneration right for machine learning of generative AI. *Computer Law & Security Review, 52*, 105925. https://doi.org/10.1016/j.clsr.2023.105925

German Research Foundation. (2024, January 30). *Pilotphase "Digitalisierung und Bereitstellung (noch) rechtebewehrter Objekte"*. https://www.dfg.de/de/aktuelles/neuigkeiten-themen/info-wissenschaft/2024/ifw-24-10

Gragert, G. (2023, November 13). *From people reading to machines learning—How Gaia-x enables digital cultural heritage.* CrossAsia. https://blog.crossasia.org/from-people-reading-to-machines-learning-how-gaia-x-enables-digital-cultural-heritage/?lang=en

HM Government. (1988, November 15). *Copyright, Designs and Patents Act 1988*. Statute Law Database. https://www.legislation.gov.uk/ukpga/1988/48/section/29A

HM Government. (2025, February 26). *Data (Use and Access) Bill*. https://bills.parliament.uk/bills/3825

IIIF Presentation API 3.0. (2023). https://iiif.io/api/presentation/3.0/

Intellectual Property Office of HM Government. (2022, June 28). *Artificial intelligence and IP: Copyright and patents: Government response to consultation*. https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/outcome/artificial-intelligence-and-intellectual-property-copyright-and-patents-government-response-to-consultation

Janssens, M.-C., Gorbatyuk, A., & **Pajares Rivas, S.** (2021). Deliverable 2.1: Mapping of the relevant European IP legal framework. *Zenodo*. https://doi.org/10.5281/zenodo.5141438

Karamolegkou, A., Li, J., Zhou, L., & **Søgaard, A.** (2023). *Copyright violations and large language models* (No. arXiv:2310.13771). arXiv. https://doi.org/10.18653/v1/2023.emnlp-main.458

Keller, P., & **Tarkowski, A.** (2021, March 5). *The paradox of open*. https://openfuture.eu/publication/the-paradox-of-open/

Kleppe, M. (2024, January 9). *Statement on commercial generative AI (KB – National Library of the Netherlands)*. https://www.kb.nl/en/ai-statement

Lehmann, J. (2022). The tragedy of the cultural commons. Research Report and Data Publication. *Zenodo*. https://doi.org/10.5281/zenodo.6513596

Lehmann, J. (2023). *Digital commons as a model for digital sovereignty: The case of cultural heritage*. https://doi.org/10.34669/WI.CP/4.15

Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act), OJ L 152/1 (2022). http://data.europa.eu/eli/reg/2022/868/oj/eng

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). (2024). http://data.europa.eu/eli/reg/2024/1689/oj/eng

RightsStatements.org. (2022). https://rightsstatements.org/en/

Samberg, R., Vollmer, T., & **Teremi, S.** (2024, March 12). *Fair use rights to conduct text and data mining and use artificial intelligence tools are essential for UC research and teaching.* Office of Scholarly Communication. https://osc.universityofcalifornia.edu/2024/03/fair-use-tdm-ai-restrictive-agreements/

Secretary of State for Science, Innovation and Technology by Command of His Majesty. (2024, December 17). *Copyright and artificial intelligence: Consultation*. https://www.gov.uk/government/consultations/copyright-and-artificial-intelligence/copyright-and-artificial-intelligence

Secretary of State for Science, Innovation and Technology by Command of His Majesty. (2025, January 13). *AI opportunities action plan*. https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan (last accessed: February 28th, 2025).

Springer Nature Customer Service Center GmbH, & MPDL Services gGmbH. (2023). *Projekt DEAL – Springer Nature Publish and Read Agreement, 2024–2028* (Version 1, p. 11223199). https://doi.org/10.17617/2.3551270

Text and Data Mining Reservation Protocol Community Group. (2024, April 16). *TDMRep*. https://www.w3.org/community/tdmrep/

Westenberger, P., & **Farmaki, D.** (2025). *Artificial intelligence for cultural heritage research: The challenges in UK copyright law and policy*. SSRN. https://doi.org/10.2139/ssrn.5153757

Wiley-VCH GmbH, MPDL Services gGmbH, & **Max Planck Society.** (2023). *Projekt DEAL – Wiley Publish and Access Agreement, 2024–2028*. https://doi.org/10.17617/2.3551268

TO CITE THIS ARTICLE:
Lehmann, J., & Sichani, A.-M. (2025). A Position Paper on AI and Copyrights in Cultural Heritage and Research (EU and UK). *Journal of Open Humanities Data*, 11: 25, pp. 1–9. DOI: https://doi.org/10.5334/johd.290

Submitted: 05 December 2024
Accepted: 06 March 2025
Published: 02 April 2025

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.