STOCKHOLM INTELLECTUAL PROPERTY LAW REVIEW

#2 | 2024

STOCKHOLM
INTELLECTUAL PROPERTY
LAW REVIEW

CONTACT US
Do you want to publish in the review? For ordering, general comments and questions please contact us at inquiries@stockholmiplawreview.com

CONTENT EDITOR
Frantzeska Papadopoulou Skarp

BOARD OF DIRECTORS
Professor Frantzeska Papadopoulou Skarp, Department of Law, Stockholm University

Associate Professor Åsa Hellstadius, Vinge Law Firm

Jur. Dr. Richard Wessman, Partner, Vinge Law Firm

Associate Professor Marcus Holgersson, Chalmers University of Technology

Mats Lundberg, Managing Partner and Managing Director, Groth & Co

WEBPAGE
www.stockholmiplawreview.com

LINKEDIN
https://www.linkedin.com/company/stockholmiplawreview/

PRODUCTION
eddy.se ab

PRINT
The Faculty TF AB, Visby 2025

ISSN 2003-2382 (Online)
ISSN 2003-2390 (Print)

Content

Editorial

2024 was an important year both for legal and for technical developments relating to artificial intelligence (AI) and text and data mining (TDM). Landmark cases, such as the LAION ruling by the Hamburg court, provided early judicial interpretations of national TDM provisions post-implementation of the Copyright in the Digital Single Market (CDSM) Directive, while further litigation has recently also unfolded in the United States and the United Kingdom.

Against this backdrop, the Institute for Intellectual Property and Market Law (IFIM) at Stockholm University hosted a conference on TDM, AI, and Libraries in collaboration with Wikimedia Sverige and Swedish Library Association. The event was opened by the Dean of the Law Faculty, Professor Jane Reichel and closed by the National Librarian Karin Grönvall and Stockholm University Library's Head Librarian Wilhelm Widmark. It brought together legal scholars, researchers, and librarians all eager to examine the evolving legal framework surrounding TDM, AI-driven research, and its impact on knowledge dissemination.

A major theme of the conference was the complex interplay between copyright and AI-related research. While copyright may serve as a foundation for intellectual creation, it also presents a number of uncertainties and potential obstacles for researchers and libraries, particularly when it comes to digitization and access to materials to be used in TDM. Libraries hosting digitized materials are restricted by national copyright legislation when it comes to accessibility provided to researchers. The research exceptions to copyright are, in turn, hard to navigate and rarely interpreted in the national courts. The conference provided an important and rather unique platform to discuss if and how copyright needs to be amended to allow libraries to fulfil their role in supporting research and researchers.

Building on the very interesting debates in the conference, this issue of the Stockholm IP Law Review explores the legal, ethical, and practical challenges of TDM in the age of AI. The contributions examine TDM implementations across jurisdictions, the role of open-access knowledge, and the implications of AI for copyright law. Maja Bogataj Jančič and Ema Purkart analyze Slovenia's approach to TDM, highlighting both the progressive steps taken and the lingering legal uncertainties, while Konrad Gliściński provides insights into Poland's implementation, where conflicting interpretations have raised concerns over its compatibility with EU law. Branka Marušić explores how different EU Member States have navigated the harmonization of TDM exceptions, questioning whether the legal framework fosters cohesion or divergence across Europe.

Beyond legislative analysis, this issue also considers how AI is reshaping legal research itself. Professor Johan Lindholm examines the growing role of computational methods in legal scholarship, highlighting how natural language processing (NLP) and large-scale data analysis can transform traditional legal research methodologies. His work challenges the perception that doctrinal and empirical approaches are incompatible, arguing instead that data-driven legal analysis can provide deeper insights into legal texts, case law, and legal decision-making patterns. In a similar vein, Holli Sargeant and Professor Felix Steffek introduce a dataset of UK Court decisions, the Cambridge Law Corpus, and explore how AI models can predict outcomes in the UK Employment Tribunal, offering a glimpse into the future of computational legal analysis. These contributions reflect how AI is not only reshaping how legal professionals access and interpret the law but also redefining the nature of legal scholarship.

The role of open-access knowledge in AI training is another topic addressed. Eric Luth scrutinizes the use of Wikipedia and the other Wikimedia platforms as a source for AI training data, highlighting potential tensions between open-access licensing and proprietary and commercial AI development while arguing for the value and importance of open-access material in AI training. Relatedly, Ana Lazarova and Eric Luth examine the position of knowledge custodians—libraries, archives, and cultural heritage institutions—as enablers or gatekeepers in the AI era, exploring the legal and practical dilemmas they face in managing access to digital resources. A core

focus of their discussion is the opt-out mechanism of the CDSM Directive's TDM exception, assessing whether current legal structures empower rightsholders to control AI's use of copyrighted material—or rather introduce further legal uncertainty that could hinder research and innovation.

As AI-driven research accelerates, so does the urgency of ensuring that copyright law evolves to support and not stifle scientific inquiry. The contributions in this issue reflect the ongoing legal debates surrounding TDM, copyright, and AI, offering perspectives on how the law can better accommodate technological progress, respect the rights and interests of copyright holders, while safeguarding the ecosystem of free and open knowledge and its production.

Frantzeska Papadopoulou Skarp
Eric Luth
Lisa Gemmel

Frantzeska Papadopoulou Skarp

Frantzeska Papadopoulou Skarp is Professor of Intellectual Property Rights and the Head of the IP Law Group of Stockholm University. Papadopoulou is a member of the Research Council of the Law Faculty at Stockholm University and the Chair of IFIM (Research Institute for Intellectual Property Rights and Market Rights). She is the editor-in-chief and founder of the Stockholm Intellectual Property Law Review and a member of the Board of the National Library of Sweden.

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.

Lisa Gemmel

Lisa Gemmel is the press and policy officer at Swedish Library Association. She has a law degree from Stockholm University, and has a background in labour law and policy work in labour market and issues regarding culture workers.

Photo: Severus Tenenbaum.

CONTENT EDITOR

CONTENT EDITOR

Frantzeska Papadopoulou Skarp

ASSISTANT CONTENT EDITOR

ASSISTANT CONTENT EDITOR

Leonidas Fotiatis

STUDENT EDITORIAL TEAM

EDITOR

Constantin Berlage

EDITOR

Rashad Mahammadzada

EDITOR

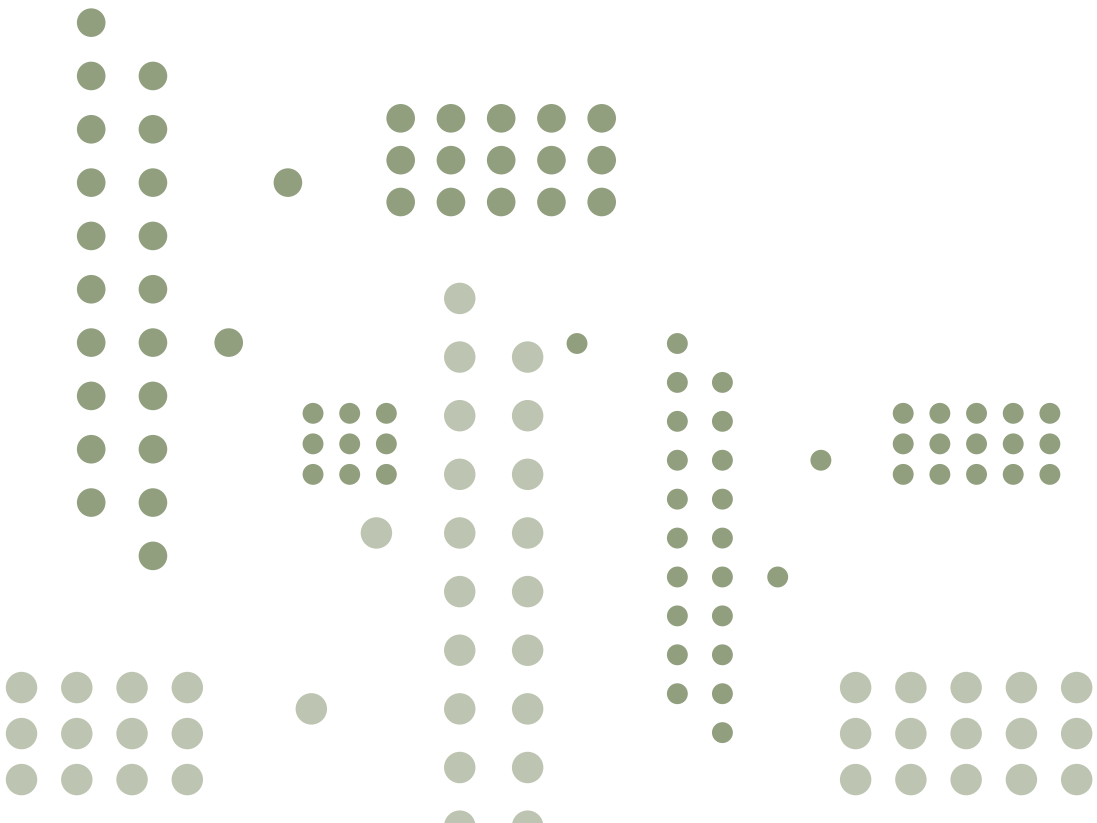
Melanie Lindgren

EDITOR

Sofía Sarmiento

EDITOR

Rahul Scharfenberg



Text and Data Mining in the Slovenian Legal System

Maja Bogataj Jančič and Ema Purkart

ABSTRACT

The Slovenian implementation of the text and data mining exceptions in Articles 57a in 57b of the Copyright Act provides both very progressive elements of the European TDM exceptions implementation and also problematic ones.

The TDM exceptions allow the digitization of analogue works for the purpose of TDM as well as the remote access to content and, in the case of the TDM exception for scientific research, also the sharing of the results for TDM purposes, which is a very progressive implementation worth repeating elsewhere. Rights holders also need to ensure that the beneficiaries of both exceptions can effectively perform TDM and need to act within 72 hours or face sanctions.

Consequently, the Slovenian legal order represents a favorable legal basis for building models of generative artificial intelligence.

The problematic aspect of the Slovenian implementation is that it does not explicitly consider access to the content freely available online as lawful access, as is otherwise explicitly stated in the Recital 14 of the DSM Directive. In this regard, artificial intelligence builders in Slovenia can be significantly worse off, and it is reasonable to expect that the legislators will correct this error in the future.

Despite this obstacle researchers who build open-access LLMs for Slovenian or other languages have a good legal basis for collecting texts and building datasets, sharing them with others, and building LLMs on the basis of the Slovenian exception.

1. INTRODUCTION

The Directive on copyright and related rights in the Digital Single Market[1] (hereinafter: "the DSM Directive") was implemented in Slovenia by the Copyright and Related Rights Act[2] (hereinafter: "the Copyright Act") and the Act Regulating Collective Management of Copyright and Related Rights[3] in autumn 2022.

The Slovenian implementation of the *text and data mining* (hereinafter: "TDM") *exceptions* in Arts. 57a and 57b of the Copyright Act provide a progressive example of the implementation. Both TDM exceptions allow the digitization of analogue works for the purpose of TDM. According to both exceptions the remote access to content is permitted. In the case of the TDM exception for scientific research also the sharing of the results for TDM purposes is allowed. Both exceptions provide that the rights holders need to ensure that the beneficiaries of both exceptions can effectively perform TDM and need to act within 72 hours or face sanctions.

2. EXCEPTION FOR TEXT AND DATA MINING FOR SCIENTIFIC RESEARCH PURPOSES

The exception for TDM for the purposes of scientific research grants to research organizations, publicly accessible archives, libraries, museums, film or audio heritage institutions, public broadcasting organizations, and persons belonging to research organizations and cultural heritage institutions (hereinafter: "beneficiaries of the exception for TDM for scientific research") the right to freely reproduce works to which they have lawful access and to carry out TDM on these works.[4] TDM means any

1 Directive (EU) 2019/790 of the European Parliament and of the Council on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC [2019] OJ L 130 (DSM Directive).

2 Zakon o avtorski in sorodnih pravicah (ZASP) (The Copyright and Related Rights Act), 1995.

3 Zakon o kolektivnem upravljanju avtorske in sorodnih pravic (ZKUASP) (The Act Regulating Collective Management of Copyright and Related Rights), 2016.

4 ZASP (n 2) art 57b(1), *"Research organisations, publicly accessible archives, libraries, museums, film or audio heritage institutions and public broadcasting organisations, as well as persons belonging to research organisations and cultural heritage institutions, may freely reproduce works to which they have lawful access under the conditions laid down in this Article and shall carry out text and data mining operations referred*

automated analytical technique aimed at analyzing text and data in electronic form to generate information such as patterns, trends, and correlations, including the digitization of analogue content and remote access to such content where this is necessary.[5]

The TDM exception for scientific research provides its beneficiaries the possibility of reproduction, which also includes the digitization of analog content, when necessary for the purposes of TDM.[6] Remote access to such content is also permitted under the same conditions. The right to reproduce and carry out TDM is limited to works to which beneficiaries of the exception for TDM for scientific research have lawful access. Lawful access includes access to works based on free licenses, contracts, or other legal bases[7]. When implementing Art. 3 of the DSM Directive to the Copyright Act, the Slovenian legislator narrowed the meaning of lawful access whereas, in accordance with the Recital 14 of the DSM Directive,[8] *lawful access* should also include access to content that is freely available online. Although the science and research stakeholders advocated for implementation that would include freely available online content, the Slovenian Copyright Office, which provided expert support to the Slovenian legislator, had a different much more pro-rightsholders view, which usually favors to limit the scope of exceptions and limitations as much as possible. They argued in favor of not specifically including the text from the Recital 14 into both articles of the Slovenian Implementation. Consequently, the Slovenian implementation of the lawful access provision does not include explicit mentioning that the access to the content that is freely available online represents as lawful access, which may greatly reduce the scope of this exception and limits scientific research in this field.[9]

Sharing and making available to the public of the results of TDM are also permitted under the exception. Such use is possible if four conditions are cumulatively met: (i) the extent of TDM must be limited by the intended purpose, (ii) it must be in accordance with fair practice, (iii) it must not conflict with the normal use of the work and (iv) does not unreasonably conflict with the author's legitimate

to in paragraph one of the preceding Article on works to which they have lawful access for the purposes of scientific research under the conditions laid down in this Article, including the digitisation of analogue content and remote access to such content where this is necessary for the purposes of text and data mining.".

5 Ibid art 57a(1), *"For the purposes of text and data mining, the reproduction of lawfully accessed works shall be free. Text and data mining shall mean any automated analytical technique aimed at analysing text and data in electronic form to generate information such as patterns, trends and correlations, including the digitisation of analogue content and remote access to such content where this is necessary for the purposes of text and data mining.".*

6 Maja Bogataj Jančič, 'Exceptions with teeth: the new Slovenian text and data mining provisions' (2023) <https://www.knowledgerights21.org/blog/exceptions-with-teeth-the-new-slovenian-text-and-data-mining-provisions/> accessed 15 October 2024.

7 Zakon o obveznem izvodu publikacij (ZOIPub) (The Legal Deposit Act), 2006.

8 DSM Directive (n 1) rec 14, *"Research organisations and cultural heritage institutions, including the persons attached thereto, should be covered by the text and data mining exception with regard to content to which they have lawful access. Lawful access should be understood as covering access to content based on an open access policy or through contractual arrangements between rightholders and research organisations or cultural heritage institutions, such as subscriptions, or through other lawful means. For instance, in the case of subscriptions taken by research organisations or cultural heritage institutions, the persons attached thereto and covered by those subscriptions should be deemed to have lawful access. Lawful access should also cover access to content that is freely available online.".*

9 Maja Bogataj Jančič, Sandra Koren, 'Avtorske pravice so zaščitene bolje. Javni interes pa slabše.' (2022) Sobotna priloga, Delo <https://www.delo.si/sobotna-priloga/avtorske-pravice-so-zascitene-bolje-javni-interes-pa-slabse> accessed 15 October 2024.

interests.[10] These four conditions[11] are specifically mentioned in this article although according to the Slovenian Copyright Act, the beneficiary of the exception must, in addition to the conditions specified for every exception, also take into account the conditions of Art. 46 of the Copyright Act to fulfill conditions for lawful use.[12]

The greatest feature of the Slovenian implementation of the exception for TDM for scientific research is the prohibition for "authors" to use disproportionate measures to ensure the security and integrity of their networks and databases, and that these measures must not prevent the effective implementation of TDM. In the event that the beneficiaries of the exception for TDM for scientific research could not, due to security and protection measures, perform actions that are permitted to them in accordance with the exception, the "author" must provide the beneficiary with access to and use of the works in accordance with the exception within 72 hours. According to the general rule contained in Art. 166c of the Copyright Act, the rights holder must provide the means for actors to exercise their rights under an exception within the shortest time possible, otherwise there is a possibility of a request for mediation. In contrast, the special regulation within the exception for TDM for scientific research provides a specific 72-hour deadline for the "author" to enable the beneficiary the access and use of works.[13] If the "author" does not enable TDM within this period, the conditions for awarding the author with a fine will arise.[14]

It is important to highlight that the provision uses the term "author" although it relates to the rights holder since in most cases the author will have no power over the removal of security and protection measures, and only the rights holder who (in commercial) practice apply these measures, do.

3. GENERAL EXCEPTION FOR TEXT AND DATA MINING

The general exception for TDM, which is regulated in Art. 57a of the Copyright Act, allows the free reproduction of lawfully accessed works for the purposes of TDM.[15] Here again, the permitted use also includes digitization of analog content and remote access to this content, when this is done for the purposes of TDM.[16] This exception constitutes the legal basis for all other purposes other than scientific research. The retention of copies of works created by TDM is also permitted, but is limited to the period when retention is necessary for the purposes of TDM. As with the exception for TDM for scientific research, the author (copyright holder?) must provide the beneficiary of the exception with access to the works within no more than 72 hours if the security and protection measures taken by the author prevent the beneficiary from exercising the exception.[17]

Authors may expressly and appropriately exclude the applicability of the TDM exception to their works. Unlike the DSM Directive, the Slovenian legislator used the term "author" and not "rights holder", which may represent a particular challenge for the implementation of this "opt-out" option in Slovenian legislation. It is also important to highlight that according to the current provision, all contractual stipulations contrary to this exception are null and void.[18] This means that "opting-out" via contracts is not possible according to Slovenian implementation.[19]

10 ZASP (n 2) art 57b(5), *"The sharing and making available to the public of the results of the text and data mining referred to in paragraph one of this Article shall be permissible provided that the extent of the text and data mining is limited by the intended purpose, is compatible with fair practice, does not conflict with normal use of the work and does not unreasonably prejudice the legitimate interests of the author."*.

11 In addition to the standard three -step test from the Article 9(2) Berne Convention Slovenian copyright law requires a fourth condition as well. This condition was otherwise recognized by the Stockholm revision of the Berne in 1967 but only for quotation exception.

12 Maja Bogataj Jančič, Avtorsko pravo v digitalni dobi : problematika zaščite avtorskih del s tehnološkimi ukrepi (Pasadena 2008) 46, *"An analysis of our legal system shows that the limitations of copyright in our country are very narrowly designed. The so-called three-step test, which is specified in international conventions as an instruction to the legislator on how to create exceptions in the law, is enacted in our country in Article 46 as a four-step test, which constitutes a binding instruction to the judge on the principles by which he should judge the validity of an individual limitation of exclusive rights. This means that the three-step test was not merely an instruction to the legislator on how to create exceptions, but must also be applied in each individual case."*.

13 ZASP (n 2) art 57b(4), *"An author may take appropriate measures to ensure the security and integrity of his networks and databases, but such measures may not be disproportionate and may not prevent the effective implementation of text and data mining as referred to in paragraph one of this Article. If the use of any security and protection measures prevents a person from carrying out acts permitted under this Article, the author shall provide that person with access to and use of the works in accordance with this Article within a time limit not exceeding 72 hours."*.

14 Ibid art 185, *"(1) A fine of between EUR 850 and EUR 3,000 shall be imposed for a minor offence on a legal entity that fails to provide a person that has lawful access to a copy of copyright work or to a subject matter of related rights with the means to enable that person the exercise of substantive limitations to rights (Article 166c). (2) A fine of between EUR 250 and EUR 1,500 shall be imposed on a sole trader or a self-employed person for the minor offence referred to in the preceding paragraph. (3) A fine of between EUR 250 and EUR 1,000 shall be imposed on the responsible person of a legal entity or the responsible persons of a sole trader or of a self-employed person for the offence referred to in paragraph one of this Article. (4) A fine of between EUR 250 and EUR 700 shall be imposed on an individual for the minor offence referred to in paragraph one of this Article."*.

15 ZASP (n 2) art 57(a)1.

16 Ibid.

17 ZASP (n 2) art 57a(4), *"An author may take appropriate measures to ensure the security and integrity of his networks and databases, but such measures may not be disproportionate and may not prevent the effective implementation of text and data mining as referred to in paragraph one of this Article. If the use of any security and protection measures prevents a person from carrying out acts permitted under this Article, the author shall provide that person with access to and use of the works or other protected subject matter in accordance with this Article within a time limit not exceeding 72 hours."*.

18 Ibid art 57a(5), *"Any contractual stipulation contrary to this Article shall be null and void."*.

19 Maja Bogataj Jančič, Laura Pipan, 'Text and Data Mining Copyright Exceptions Regulation in Central and Southeastern Europe' (2024) <https://www.odipi.si/wp-content/uploads/2024/07/TDM.pdf> accessed 15 October 2024.

4. CONCLUSION

The Slovenian legal order represents a favorable legal basis for building models of generative artificial intelligence because it provides for many favorable elements that will enhance machine learning in Slovenia. Unfortunately, the implementation also has certain problematic aspects: the most significant can turn out to be the omniscience of the express inclusion of the content freely available online in the definition of lawful access, as is explicitly stated in the Recital 14 of the DSM Directive. In this regard, artificial intelligence builders in Slovenia are significantly worse off, and it is reasonable to expect that the legislators will correct this error in the future.

This arrangement can also hinder the construction of a large open-access large language model for the Slovenian language, which is currently being built with public funds.[20]

Despite the obstacle that freely available content on the web is not expressly included in the exception, researchers who build open-access large language models for Slovenian or other languages have a good legal basis for their work for collecting texts in Art. 57b of the Copyright Act. Primarily, this is due to other available legislation (e.g. the Legal Deposit Act[21]), which allows lawful access to legally deposited materials for research purposes, which includes web harvesting of certain content as well.[22]

Additionally, the Slovenian article provides a good basis for the sharing of data sets, a topic that was recently touched upon in Europe's first TDM case, the German case of Robert Kneschke v. LAION.[23][24] Article 57b is also a very solid legal basis for the creation of an open-access large-scale language models,[25] which may frustrate rights holders and collective organizations that may have hoped for different new business models in such cases.

20 PoVejMo, 'Medijske objave' <https://povejmo.si/medijske-objave/> accessed 15 October 2024.

21 ZOIPub (n 7).

22 Ibid art 18(2), "*Primerki obveznega izvoda, ki nimajo statusa arhivskih izvodov, se uporabljajo za izvajanje knjižničnih informacijskih storitev ali morajo biti na voljo vsaj za študijske in raziskovalne namene v skladu s pravilnikom iz tretjega odstavka 13. člena tega zakona.*", Pravilnik o vrstah in izboru elektronskih publikacij za obvezni izvod, (2007), art 11, *"(1) Arhiv obveznega izvoda spletnih publikacij je praviloma javen in prosto dostopen. (2) Imetnik avtorske oziroma intelektualnih pravic lahko omeji dostop do obveznega izvoda svoje spletne publikacije, vendar mora biti zagotovljena prosta uporaba take publikacije za študijske in raziskovalne namene. (3) Prikaz spletnih publikacij na zaslonu in iskanje ter nalaganje datotek na delovno postajo so dovoljeni pri uporabi vseh arhiviranih obveznih izvodov spletnih publikacij vsaj za študijske in raziskovalne namene. (4) Indeksiranje spletnega arhiva obveznih izvodov NUK z uporabo spletnih iskalnikov ni dovoljeno.*".

23 Robert Kneschke v LAION eV [2024] 310 O 227/23 <https://pdfupload.io/docs/4bcc432c> accessed 15 October 2024.

24 Andres Guadamuz, 'LAION wins copyright infringement lawsuit in German court' (TechnoLlama, 28 September 2024) <https://www.technollama.co.uk/laion-wins-copyright-infringement-lawsuit-in-german-court> accessed 15 October 2024.

25 Paul Keller, 'LAION vs Kneschke, Building public datasets is covered by the TDM exception' (Open Future, 10 October 2024) <https://openfuture.eu/blog/laion-vs-kneschke/> accessed 15 October 2024.

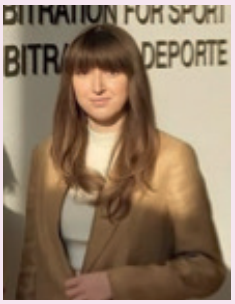
Maja Bogataj Jančič

Dr. Maja Bogataj Jančič is the founder and head of the Open Data and Intellectual Property Institute ODIPI based in Slovenia.

Maja is a copyright expert; her work focuses on open science, open data, data governance and artificial intelligence, as well as open science issues and the legal framework of copyright and data for research and science.

Maja is an Associate Research Fellow at the Berkman Klein Center for Internet and Society at Harvard. She co-chaired the Data Governance Working Group of The Global Partnership on Artificial Intelligence (GPAI) in 2020-2023. She is a board member of Communia and has been the representative and legal lead of Creative Commons Slovenia since 2004. Maja is the Knowledge Rights 21 (KR21) National Coordinator for Slovenia and the Regional Coordinator for the Central and South Eastern Europe. She has also been appointed by the Minister as the ERA (European Research Area) Action 2 promoter for Slovenia; ERA 2 actions focus on the impact of copyright and data regulation on research and innovation.

Maja graduated from the Faculty of Law in Ljubljana (1996), obtained her LL.M. from the Faculty of Law in Ljubljana (1999, Economics), Harvard Law School (2000, Law) and Facoltà di Giurisprudenza di Torino (2005, Intellectual Property), and her Ph.D. from the Faculty of Law in Ljubljana (2006, Copyright).



Ema Purkart

Ema Purkart is a Master Law student at the Faculty of Law at the University of Ljubljana. During her studies, she was an ERASMUS + exchange student at the EBS University in Germany. She also successfully represented her University at the Sports Law Arbitration Moot (SLAM).

Ema is a research assistant at the Open Data and Intellectual Property Institute ODIPI based in Slovenia. She is currently involved in the project Sustainable Digital Preservation of the Slovenian New Media Art and the research on the exceptions and limitations to copyright for scientific research and education.

Polish Implementation of TDM Exceptions – General Characteristics

Konrad Gliściński

ABSTRACT

The aim of this article is to analyse the implementation of Directive (EU) 2019/790 on copyright and related rights in the context of Text and Data Mining exceptions within Polish law. It highlights interpretative challenges and uncertainties arising from the regulations, potentially leading to legal disputes. The article begins with an overview of the Directive and then examines the specific provisions in Polish law that implement it, focusing on the general and research exceptions. It discusses the lack of clarity in definitions, the scope of exceptions, and the implications for potential beneficiaries. Additionally, it identifies uncertainties regarding the storage of copies, access conditions, and protections against technical measures. Ultimately, the article concludes with a summary of the main challenges presented by the implementation and their potential impact on the practical use of Text and Data Mining exceptions.

1. INTRODUCTION

The purpose of this article is to provide a general overview of how Polish law has implemented the exceptions related to text and data mining (TDM) as outlined in the CDSM Directive.[1] Two exceptions enabling TDM have been incorporated into Polish law: a general one, based on Art. 4 of the CDSM Directive, and a specific one for scientific research purposes, based on Art. 3 of the directive. Both exceptions are independent of each other. This means that beneficiaries of the research-specific exception will also be able to base their activities on the general exception, and vice versa, as long as the conditions set out in each exception are met.[2] In both instances, the legislator opted not to introduce compensation for the use of works for TDM purposes. The Polish legislator delayed the adoption of the relevant provisions, which only came into effect on 20 September 2024.[3] Although there was ample time for public consultations and adjustments, the current provisions raise concerns and may lead to interpretative disputes. Interestingly, numerous entities participated in these consultations,[4] but in many cases, their input was not reflected in the final version of the law. One significant exception in this regard was the proposal to exclude the use of both exceptions for the purpose of "creating generative artificial intelligence models." After criticism of this solution as potentially inconsistent with EU law, this exclusion was not included in the final text of the law.[5] Additionally, there may be aspects of the Polish regulations that conflict with EU law.

2. TDM – GENERAL INFORMATION

For the purposes of further analysis, it is worth explaining in simple terms what TDM (text and data mining) involves. It seems possible to outline three typical—though not always essential—steps in TDM processes: (1) accessing content, (2) extracting or copying content, and finally, (3) analysing the text or data to uncover knowledge. In the execution of Step 3, we can further distinguish, among others, Stage A (preliminary), which involves cleaning and normalising the texts, and Stage B, which involves the direct analysis of the data.[6] From the perspective of copyright law, we can identify that steps two and three may involve the right to reproduce

1 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 (Text with EEA relevance.) OJ L 130, 17.5.2019, p. 92–125.

2 See: E. Rosati, *Copyright in the Digital Single Market. Article-by-Article Commentary to the Provisions of Directive 2019/790* (OUP:2021), p. 41.

3 Ustawa o zmianie ustawy o prawie autorskim i prawach pokrewnych, ustawy o ochronie baz danych oraz ustawy o zbiorowym zarządzaniu prawami autorskimi i prawami pokrewnymi z dnia 26 lipca 2024 r. (Dz.U. z 2024 r. poz. 1254)

4 See: https://legislacja.rcl.gov.pl/projekt/12382002/katalog/13037394#13037394.

5 K. Gliściński, 'The Good, the Bad and the Missing – the new proposal for the implementation of the CDSM Directive into Polish law', (*Communia Association*, 1 March 2024) https://communia-association.org/2024/03/01/the-good-the-bad-and-the-missing/ accessed 10 August 2024.

6 E. Rosati, *Copyright in the Digital Single Market. Article-by-Article Commentary to the Provisions of Directive 2019/790* (OUP:2021), p. 68–71.

works. Before the introduction of exceptions for TDM, it was already clear that the copying of works as part of the preparatory activities for TDM constituted reproduction (Step 2). Such reproduction could be carried out with the rights holder's permission (a licence) or under permitted uses provisions.[7] The open question was the status of the analysis itself conducted within the TDM processes (Step 3). Specifically, the question was whether such activities constitute a form of reproduction of works or whether, e.g. due to the lack of human involvement and only machine use, these activities do not constitute reproduction within the meaning of copyright law.

This brings up an important issue. Before the introduction of the discussed exception into Polish law, did the copyright monopoly also cover the activities performed within the scope of Step 3? The Polish structure of economic rights is based on a (dynamic) construction of fields of exploitation.[8] According to Polish copyright law,[9] "...the author shall have an exclusive right to use the work and to dispose of its use throughout all the fields of exploitation and to receive remuneration for the use of the work." However, the Polish law does not introduce a definition of a field of exploitation. Art. 50 only provides examples of such fields.[10] These example fields of exploitation cover copyright provisions defined in EU directives but also include forms of use that have not been harmonised at the EU level.[11] The essence of the dynamic construction of fields of exploitation lies in the fact that, with technological development, new fields of exploitation may emerge,[12] which will automatically fall under the copyright monopoly. However, these fields of exploitation must first be distinguished in contractual practice.[13] However, it seems that (although this statement is not supported by empirical analyses) before the introduction of this exception, there was no widespread practice of distinguishing the activities that make up Step 3 as a separate field of exploitation in copyright agreements. Nor was it common to licence works in this area. Neither the current construction of new exceptions nor the content of Art. 50 directly answers the question of whether the activities carried out in the context of Step 3,[14] in themselves constitute a new field of exploitation. This situation supports the claim that, under Polish law, the activities previously carried out under Step 3 were not covered by the copyright monopoly.

This approach is also supported by the principle of public domain. According to this principle, the existence of exclusive rights imposing obligations on others to refrain from using works in a specific manner should not be presumed if such rights are not explicitly provided for by law.[15] Therefore, since it was not common practice to reserve certain types of activities for rights holders, those performing these activities should not be unexpectedly informed that they were infringing on copyright. This approach is particularly justified in light of the possibility of infringing copyright without fault. Such strict liability, although common in copyright law, should only apply to activities that are objectively defined in the law as falling within the scope of the monopoly but have been infringed without fault. However, if certain activities were not previously specified in the law, the possibility of judicially extending the copyright monopoly to those activities should not be allowed.

Of course, in practice, the issue of assessing the execution of activities involved in Step 3 will only be clarified through jurisprudence. The problem is unlikely to arise in situations where reproduction under Step 2 was carried out under the previously applicable provisions on permitted use, as these provisions could only serve as a basis for reproducing works in a limited range of situations. The issues in this area may particularly concern situations where the other party to the contract obtained, either through a transfer of rights or a licensing agreement, the right to use works in the field of digital reproduction.[16] If, according to the interpretation proposed here, we consider that the analytical activities within Step 3 do not constitute an act of exploitation, this means that such activities fall outside the scope of copyright monopoly. Consequently, performing these activities is not reserved for the rights holder, and simply acquiring or licensing the right to reproduction would be sufficient for conducting TDM. On the other hand, if we determine that the activities carried out within Step 3 are also covered by copyright, simply obtaining a licence or acquiring rights for digital reproduction would not be considered sufficient. Consequently, it would have to be recognised that such a person infringed the copyright of the work.

7 In the Polish legal system, exceptions and limitations to copyright are referred to as "dozwolony użytek," which in English translates to "permitted uses".

8 K. Gliściński, Komentarz do art 17, [w] A. Michalak, Ustawa o prawie autorskim i prawach pokrewnych. Komentarz, Warszawa 2019, p. 147–150.

9 Art. 17 Ustawa o prawie autorskim i prawach pokrewnych z dnia 4 lutego 1994 r. (Dz.U. z 2022 r. poz. 2509).

10 Ustawa o prawie autorskim i prawach pokrewnych z dnia 4 lutego 1994 r. (Dz.U. z 2022 r. poz. 2509).

11 According to it: "The separate fields of exploitation shall be, in particular: (1) within the scope of fixing and reproduction of works – the production of copies of a work using specific technologies, including printing, reprographics, magnetic fixing, and digital technology; (2) within the scope of trading the original or the copies on which the work was fixed – the introduction to trade, lending for use, or rental of the original or copies; (3) within the scope of dissemination of works in a manner different from that defined in subparagraph 2 – public performance, exhibition, screening, presentation, and broadcasting, as well as retransmission, and making the work publicly available in such a manner that anyone could access it at a place and time selected by them."

12 For example, through the mass identification of a specific method of using works in contracts as a separate source of economic benefits.

13 K. Gliściński, Wyodrębnianie się nowych pól eksploatacji i ich wpływ na obrót prawami do utworów, ZNUJ. PPWI 2010, nr 3, s. 45–60.

14 Ustawa o prawie autorskim i prawach pokrewnych z dnia 4 lutego 1994 r. (Dz.U. z 2022 r. poz. 2509).

15 K. Gliściński, Komentarz do art 17, [w] A. Michalak, Ustawa o prawie autorskim i prawach pokrewnych. Komentarz, Warszawa 2019, p. 145–147.

16 Another issue is to what extent and based on what form of permitted uses, before the introduction of the analyzed exception, it was possible to "reproduce" works for the purpose of performing step 2.

3. TDM – DEFINITIONS

According to the CDSM Directive TDM "means 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."[17] The Polish implementation enabling TDM is based on a definition that essentially resembles the definition contained in the CDSM directive. According to it: "The exploration of texts and data involves their analysis solely through the use of an automated technique designed for analysing texts and data in digital form, with the goal of generating specific information, including, in particular, patterns, trends, and correlations."[18] In the Polish translation of the Directive, the term "mining" has been translated as "*eksploracja*" ("exploration" in English). In the literature, certain doubts have been raised regarding the wording of this provision. It states that exploration must occur "solely through the use of an automated technique".[19] According to some, this wording may lead to uncertainties about whether preparatory activities such as pre-processing, data cleaning, or normalisation are covered by the provision Step 2. These activities are performed by humans and are not automated. The issue in this context concerns the use of the word "solely", which does not appear in the text of the CDSM Directive.[20] However, it appears that comparing this definition with the content of the relevant provision of the CDSM Directive introducing the TDM exception for scientific research provides grounds to assert that all reproduction activities are permitted as long as they serve the purpose of text and data mining (see below).

4. TEXT AND DATA MINING FOR THE PURPOSES OF SCIENTIFIC RESEARCH

a. Beneficiaries

The scope of beneficiaries indicated in the CDSM Directive refers to *research organizations* and cultural heritage institutions. As indicated in the literature, the approach taken in the Directive is based on a dual limitation: on one hand, the exception defined in Art. 3 applies only to "scientific research," and on the other hand, it must be carried out by *research organizations*. This means that independent researchers and other entities conducting "scientific research" (e.g., journalists or companies operating research centres) are outside the scope of this exception.[21] The exception in Polish law has three limitations: a formal list of beneficiaries, the purpose of TDM, and a prohibition on obtaining economic benefits (see below). Although it has not been definitively established, it seems that beneficiaries of this exception, according to Recital 11 of the CDSM Directive, can "rely on their private partners for carrying out text and data mining, including by using their technological tools".[22]

The Polish Act defines cultural heritage institutions similarly to how the CDSM Directive does. Consequently, such institutions are defined as: "a library, museum, archive, or a cultural institution whose statutory mission is to collect, protect, and promote collections of film or phonographic heritage."[23] A different legislative technique was used with respect to the second group of beneficiaries. The Polish Copyright Act, referring to the Act on Higher Education and Science, specifies a closed category of entities that are beneficiaries of this exception. They are (i) universities (both public and non-public); (ii) federations of higher education and science entities, scientific institutes of the Polish Academy of Sciences, research institutes;[24] (iii) International scientific institutes established under separate laws operating on the territory of the Republic of Poland; (iv) Łukasiewicz Center; (v) Institutes operating within the Łukasiewicz Research Network, hereinafter referred to as "Łukasiewicz Network institutes"; (vi) The Polish Academy of Arts and Sciences and other entities primarily engaged in scientific activities in an independent and continuous manner.[25] The same scope of beneficiaries has been provided for with respect to related rights[26] and databases protected by sui generis rights.[27]

Thus, this represents a narrower scope of beneficiaries compared to the broader category of *research organisa-*

17 Art. 2(2) 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 (Text with EEA relevance.) PE/51/2019/REV/1 OJ L 130, 17.5.2019, p. 92–125.

18 Art. 6(1)(22) Ustawa o zmianie ustawy o prawie autorskim i prawach pokrewnych, ustawy o ochronie baz danych oraz ustawy o zbiorowym zarządzaniu prawami autorskimi i prawami pokrewnymi z dnia 26 lipca 2024 r. (Dz.U. z 2024 r. poz. 1254).

19 Art. 6(1)(22) Ustawa o zmianie ustawy o prawie autorskim i prawach pokrewnych, ustawy o ochronie baz danych oraz ustawy o zbiorowym zarządzaniu prawami autorskimi i prawami pokrewnymi z dnia 26 lipca 2024 r. (Dz.U. z 2024 r. poz. 1254).

20 A. Matlak, M. Wyrwiński, B. Widła, Konsultacje publiczne projektu wdrożenia dyrektyw CDSM i SATCAP II [2024], https://ipwi.uj.edu.pl/documents/122195199/151128292/Konsultacje+publiczne+dotycz%C4%85ce+projektu+wdro%C5%BCenia+dyrektyw+CDSM+i+SATCAB+II+%5B2024%5D/ccbf017d-9501-46b6-94df-c5e04891f792 (10.08.2024), p. 7.

21 Thomas Margoni and Martin Kretschmer, 'A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology' (2021) 71(8) GRUR International 2022, 685–701, Available at SSRN: https://ssrn.com/abstract=3886695 or http://dx.doi.org/10.2139/ssrn.3886695 accessed 04 November 2024.

22 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 (Text with EEA relevance.), OJ L 130, 17.5.2019, p. 92–125, rec. 11 art. 77(1) Ustawa o zmianie ustawy o prawie autorskim i prawach pokrewnych, ustawy o ochronie.

23 Art. 6(1)(21) Ustawa o zmianie ustawy o prawie autorskim i prawach pokrewnych, ustawy o ochronie baz danych oraz ustawy o zbiorowym zarządzaniu prawami autorskimi i prawami pokrewnymi z dnia 26 lipca 2024 r. (Dz.U. z 2024 r. poz. 1254)

24 Ustawa o instytutach badawczych (Dz.U. z 2024 r. poz. 534).

25 Art. 7 Prawo o szkolnictwie wyższym i nauce (Dz.U. z 2024 r. poz. 1571).

26 Art. 100 Ustawa o zmianie ustawy o prawie autorskim i prawach pokrewnych, ustawy o ochronie baz danych oraz ustawy o zbiorowym zarządzaniu prawami autorskimi i prawami pokrewnymi z dnia 26 lipca 2024 r. (Dz.U. z 2024 r. poz. 1254).

27 Art. 8b Ustawa o zmianie ustawy o prawie autorskim i prawach pokrewnych, ustawy o ochronie baz danych oraz ustawy o zbiorowym zarządzaniu prawami autorskimi i prawami pokrewnymi z dnia 26 lipca 2024 r. (Dz.U. z 2024 r. poz. 1254).

tions as defined in the Directive. The Directive allows that the beneficiaries of this exception may also include entities whose "primary goal is to [...] carry out educational activities involving also the conduct of scientific research."[28] However, the aforementioned Polish catalogue does not include such educational institutions but only "entities primarily engaged in scientific activities in an independent and continuous manner".[29] Moreover, according to Recital 12 of the CDSM Directive, this definition should be interpreted broadly and include, among others, "hospitals that carry out research".[30] The current wording of the Polish implementation raises doubts as to whether it also covers such hospitals. This does not refer to hospitals run by universities (which should be considered as covered by this exception under Polish law) but rather to other hospitals that also engage in scientific research. Due to the mandatory nature of the exception outlined in Art. 3 of the CDSM Directive, such a narrow scope of beneficiaries is under the threat of being considered incompatible with EU law.

b. Permitted uses and subject matter

Polish law, similar to the CDSM Directive, permits reproduction for the purposes of TDM. This applies—*lege non distinguente*—to reproduction occurring as part of preparatory activities – Step 2 – (such as pre-processing, data cleaning, or normalisation), as well as directly within the TDM process itself (Step 3).

The Polish Copyright Act regulates the reproduction of works and objects of related rights for TDM purposes. Under the Polish Act, the term *works* also encompasses creative databases (protected under Chapter II of the Database Directive) and computer programs. Reproduction of such databases is thus covered by the exception in accordance with the CDSM Directive. At the same time, the Polish legislator, similar to the CDSM Directive, chose not to extend this exception to computer programs.[31] Polish copyright law includes the following under related rights: rights to performances, rights to phonograms and videograms (film fixations), rights to programme broadcasts, rights to first publications and scientific and critical publications, and rights to press publications within the framework of providing services by electronic means. This exception, with respect to all related rights, has been uniformly introduced and covers all related rights existing under Polish law, including those rights that have not been harmonised at the EU level.[32] An analogous exception—contained in the Database Act—allows for the reproduction (extraction) of data without restriction under *sui generis* rights.[33]

c. Direct and indirect economic benefits and TDM for scientific research purposes

The CDSM Directive generally does not prohibit TDM used for scientific research from providing economic benefits to the beneficiaries. It merely specifies that such beneficiaries must: (1) have as their primary goal the conduct of scientific research or carry out educational activities that also involve scientific research, and (2) operate on a non-profit basis[34] or reinvest all profits into scientific research[35] or pursuant to a public interest mission recognised by a Member State.[36] The provision allows

28 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 (Text with EEA relevance.), OJ L 130, 17.5.2019, p. 92–125, art. 2(1).

29 Art. 7(1)(8) Prawo o szkolnictwie wyższym i nauce (Dz.U. z 2024 r. poz. 1571).

30 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 (Text with EEA relevance.), OJ L 130, 17.5.2019, p. 92–125, rec. 12.

31 Art. 77(1) Ustawa o zmianie ustawy o prawie autorskim i prawach pokrewnych, ustawy o ochronie baz danych oraz ustawy o zbiorowym zarządzaniu prawami autorskimi i prawami pokrewnymi z dnia 26 lipca 2024 r. (Dz.U. z 2024 r. poz. 1254).

32 Art. 100 Ustawa o zmianie ustawy o prawie autorskim i prawach pokrewnych, ustawy o ochronie baz danych oraz ustawy o zbiorowym zarządzaniu prawami autorskimi i prawami pokrewnymi z dnia 26 lipca 2024 r. (Dz.U. z 2024 r. poz. 1254).

33 Art. 8b Ustawa o zmianie ustawy o prawie autorskim i prawach pokrewnych, ustawy o ochronie baz danych oraz ustawy o zbiorowym zarządzaniu prawami autorskimi i prawami pokrewnymi z dnia 26 lipca 2024 r. (Dz.U. z 2024 r. poz. 1254).

34 It is worth noting that the English wording of Art. 2(1)(a) of the CDSM Directive raises certain interpretative concerns. According to this definition, such an organization is one that operates "on a not-for-profit basis or by reinvesting all the profits in its scientific research." The issue with this phrasing lies in the fact that the Directive does not define what constitutes a "not-for-profit" organization. To my knowledge, European law also does not provide a clear definition of this term. In relation to certain types of activities, a distinction is often made between "not-for-profit" and "non-profit" organizations. Such distinctions often arise from the specific tax regulations adopted in different countries. In the US, it is noted that "A not-for-profit (NFPO) is an organization that, like a nonprofit, doesn't seek to turn a profit. However, unlike a nonprofit, a not-for-profit doesn't have to exist for the sole purpose of improving society." https://givebutter.com/blog/non-profit-vs-not-for-profit (04.09.2024). The European Commission's proposal includes a definition of organizations operating for "non-profit purposes." According to this definition, "non-profit purposes" means that, "regardless of whether the association's activities are of an economic nature or not, any profits generated are used solely to further the objectives of the organization as defined in its statutes, and are not distributed among its members." Art. 2(c) Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on European cross-border associations (Text with EEA relevance) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2023%3A516%3AFIN&qid=1693910621013 (04.09.2024). In the literature, one might encounter statements such as: "The essence of 'not-for-profit' activity is that, alongside its primary mission, it engages in ancillary commercial activities, which both foundations and educational institutions, including public ones, are permitted to undertake. This type of activity differs from 'non-profit' operations typical of administrative entities, which are never considered commercial activities and cannot generate profits." A. Bednarczyk-Płachta, Zysk założyciela szkoły wyższej niepublicznej jako inwestora w odniesieniu do zmian w prawie o szkolnictwie wyższym, PPP 2017, nr 3, s. 10–38. If we consider that a "not-for-profit" organization is one that can generate profit but must reinvest it into its activities, then the wording of Art. 2(1)(a) of the CDSM Directive may be superfluous. This assessment arises from the fact that the provision designates, in addition to "not-for-profit" organizations, another type of organization that can also generate profit but must reinvest it specifically in scientific research.

35 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 (Text with EEA relevance.), OJ L 130, 17.5.2019, p. 92–125, art. 2(1)(a).

36 Art. 2(1)(b). "Such a public-interest mission could, for example, be reflected through public funding or through provisions in national laws or public contracts." (recital 12). 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

research organisations, in principle, to generate profits. This further indicates that such organisations may also charge access fees for their analysis results as long as these fees only cover the costs of their activities (e.g., conducting analyses on behalf of external parties, including commercial entities). In this regard, the CDSM Directive only requires that: "access to the results generated by such scientific research cannot be enjoyed on a preferential basis by an undertaking that exercises a decisive influence over such an organisation."[37] This means that such results may be available to these entities, provided that other entities also have the opportunity to access these results under the same conditions, including the same financial terms. Furthermore, the directive directly provides that "research organisations should also benefit from such an exception when their research activities are carried out in the framework of public-private partnerships."[38]

In contrast, the Polish framework introduces a significant restriction. According to it, TDM for research purposes cannot be conducted "for the purpose of obtaining direct or indirect economic benefits."[39] In Polish law, this term appears in many provisions of copyright law. Generally, it is indicated that a financial benefit can be understood "as achieving profit or as reducing incurred costs."[40] This wording indicates that beneficiaries, contrary to the provisions of the CDSM Directive, will not be able to, for example, derive profits from using TDM for scientific purposes. An open question also remains as to whether they will be able to impose fees to cover the costs of providing access to such results and whether they will enter into public-private partnerships. Additionally, the Polish implementation completely overlooks the possibility of recognizing an entity conducting scientific research as a research organisation "pursuant to a public interest mission recognized by a Member State." Such a situation may occur, among other instances, when the research activity is funded by the public sector or is based on relevant provisions in national law or public contracts.[41] In summary, while the CDSM Directive allows for the possibility of deriving financial benefits under Article 3, outlining which entities and purposes are permitted, the Polish law implementing this exception outright prohibits obtaining any economic benefits. It seems that such a restrictive construction is inconsistent with the (already narrowly defined)[42] framework established in the CDSM Directive.

d. Storage and retention of copies created for TDM (Text and Data Mining) for the purpose of scientific research

The CDSM Directive specifies that the storage of copies of works and other subject matters must be done with "an appropriate level of security."[43] The Directive left the Member States the freedom to define the detailed rules for the storage of such copies.[44] The Polish law in this regard has detailed the general security requirement by specifying that: "The storage of works is conducted with a level of security that ensures access to these works is limited exclusively to authorised persons, taking into account authentication procedures."[45] The law itself does not specify who should be considered authorised persons. It seems that this term primarily refers to individuals involved in conducting scientific research on behalf of eligible beneficiaries. The decision of who qualifies as an authorised entity in the context of a particular study should be made by the beneficiary based on their internal procedures. Importantly, access to such copies is not limited solely to researchers directly participating in the study; it may also extend to other individuals (e.g., technicians, IT staff, librarians) who assist in conducting the research on behalf of the institution. Furthermore, according to Recital 11 of the CDSM Directive, beneficiaries of this exception "should also be able to rely on their private partners for carrying out text and data mining, including by using their technological tools". In this context, it can be understood that beneficiaries may designate authorised persons not only among their internal staff but also among private partners they engage for conducting text and data mining on the data copies of works. Given the requirement for "authentication procedures"[46] introduced in the Polish implementation, it seems that

96/9/EC and 2001/29/EC (Text with EEA relevance.), OJ L 130, 17.5.2019, p. 92–125.

37 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 (Text with EEA relevance.), OJ L 130, 17.5.2019, p. 92–125, art. 2(1).

38 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 (Text with EEA relevance.) PE/51/2019/REV/1 OJ L 130, 17.5.2019, p. 92–125, rec. 11.

39 Art. 26^2(1) Ustawa o zmianie ustawy o prawie autorskim i prawach pokrewnych, ustawy o ochronie baz danych oraz ustawy o zbiorowym zarządzaniu prawami autorskimi i prawami pokrewnymi z dnia 26 lipca 2024 r. (Dz.U. z 2024 r. poz. 1254).

40 J. Marcinkowska [w:] Komentarz do ustawy o prawie autorskim i prawach pokrewnych [w:] Ustawy autorskie. Komentarze. Tom I, red. R. Markiewicz, Warszawa 2021, art. 31.

41 Recital 12 of the 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 (Text with EEA relevance.) PE/51/2019/REV/1 OJ L 130, 17.5.2019, p. 92–125.

42 See: Thomas Margoni and Martin Kretschmer, 'A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology' (2021) 71(8) GRUR International 2022, 685–701, Available at SSRN: https://ssrn.com/abstract=3886695 or http://dx.doi.org/10.2139/ssrn.3886695 accessed 04 November 2024.

43 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 (Text with EEA relevance.), OJ L 130, 17.5.2019, p. 92–125, art. 3(1).

44 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 (Text with EEA relevance.), OJ L 130, 17.5.2019, p. 92–125, rec. 15.

45 Art. 26^2(2) Ustawa o zmianie ustawy o prawie autorskim i prawach pokrewnych, ustawy o ochronie baz danych oraz ustawy o zbiorowym zarządzaniu prawami autorskimi i prawami pokrewnymi z dnia 26 lipca 2024 r. (Dz.U. z 2024 r. poz. 1254).

46 Art. 26^2(2) Ustawa o zmianie ustawy o prawie autorskim i prawach pokrewnych, ustawy o ochronie baz danych oraz ustawy o zbiorowym zarządzaniu prawami autorskimi i prawami pokrewnymi z dnia 26 lipca 2024 r. (Dz.U. z 2024 r. poz. 1254).

access to such copies should be granted individually to specific persons.

Similarly to the CDSM Directive, the Polish implementation specifies that works reproduced under this exception "may be stored for scientific research purposes, including the verification of research results."[47] Polish law does not impose any time limits on the storage of copies of works reproduced under this exception.[48] Such a solution should be considered desirable, both from the perspective of the specific nature of conducting scientific research in general and sustainability goals. It is not possible to determine in advance from which point in time duplicated works will no longer be needed. Given the ongoing nature of scientific research, access to such copies may be necessary and desirable at any future time. Therefore, rather than deleting such copies, they should be preserved for future scientific research needs.

The CDSM Directive distinguishes between the "verification" of scientific research and its "review."[49] In the context of Polish law, this distinction can lead to problematic situations. While the TDM exception for scientific research allows entities involved in the verification of research results to access all copies of works used in the TDM process, the situation will be different if a researcher is interested in reviewing those results. In this case, access to these data will not be possible under the TDM exception for scientific research but rather under the general exception for research purposes. This second exception has been recently amended and now allows for the reproduction of "published small works or excerpts from larger works not exceeding 25% of the work's volume."[50] This means that the researcher will be able to physically view the data in its entirety (as long as it does not require the reproduction of data) but will not be permitted to make a complete copy of the data for the purpose of conducting the review. Certainly, such a situation is undesirable from the standpoint of research integrity and transparency. At the same time, this example highlights that the distinction introduced by the CDSM Directive seems unjustified. If the entity conducting TDM research is interested in verifying the results, it will be able to involve third parties to whom it can provide the collected copies. However, if a researcher not affiliated with the original entity conducting the research wishes to review the results, they will only be able to do so based on a limited excerpt of the collected copies.

Certainly, in practice, it will be challenging to distinguish whether a given activity constitutes the verification of research results or their review. Should the determination of whether an activity is one or the other be decided solely by the entity that originally conducted the research (e.g., by specifying a verification stage in the research protocol)? Can a scientist not affiliated with the original entity claim to independently verify the results, and how would such verification differ from a rigorous review of scientific results? Additionally, beyond the scope of this exception's regulations remains the issue of access to such data. Exceptions to the right of reproduction may only grant beneficiaries the right to make copies of certain data but do not impose an obligation on any entities to create such copies. In other words, if a scientist wishes to verify results, but the entity that created the data is unwilling to provide access, the verification cannot be enforced.[51]

e. Measures to ensure the security and integrity of the networks and databases

Following the CDSM Directive, the Polish implementation stipulates that: "Rightholders, in order to ensure the security and integrity of networks and databases in which works are stored, may use only the measures necessary to achieve this goal."[52] The Polish legislation does not specify exactly which measures can be employed by authorised entities, nor does it indicate which measures are considered impermissible. According to Recital 16 of the CDSM Directive, such measures could, for example, "be used to ensure that only persons having lawful access to their data can access them, including through IP address validation or user authentication". These issues are expected to be resolved in practice by judicial rulings at the level of the Court of Justice of the European Union (CJEU). However, it appears that impermissible measures would include those that either prevent or significantly hinder the extraction of data from databases for the purpose of TDM used in scientific research. Currently, there are no publicly known actions by the Polish authorities aimed at fulfilling the obligations arising from Art. 3(4) of the CDSM Directive, including those specified in Art. 3(2) and 3(3) thereof.

f. Protection against contractual override

Art. 7(1) of the CDSM Directive provides that any contractual provision contrary to the exceptions for TDM for scientific research "shall be unenforceable." Consequently,

47 Art. 26²(2) Ustawa o zmianie ustawy o prawie autorskim i prawach pokrewnych, ustawy o ochronie baz danych oraz ustawy o zbiorowym zarządzaniu prawami autorskimi i prawami pokrewnymi z dnia 26 lipca 2024 r. (Dz.U. z 2024 r. poz. 1254).

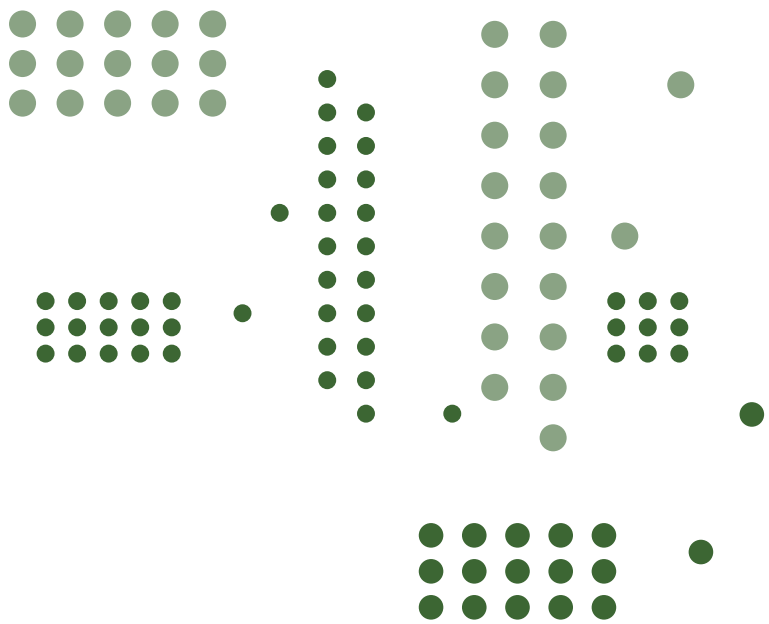
48 A time limit for storing such copies has been introduced in German law, for example, in Section 60d(5) Copyright Act of 9 September 1965 (Federal Law Gazette I, p. 1273) https://www.gesetze-im-internet.de/englisch_urhg/englisch_urhg.html (18.08.2024).

49 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 (Text with EEA relevance.), OJ L 130, 17.5.2019, p. 92–125, rec. 15.

50 Art. 27(1) Ustawa o zmianie ustawy o prawie autorskim i prawach pokrewnych, ustawy o ochronie baz danych oraz ustawy o zbiorowym zarządzaniu prawami autorskimi i prawami pokrewnymi z dnia 26 lipca 2024 r. (Dz.U. z 2024 r. poz. 1254).

51 This issue highlights that copyright law—while it affects scientific activity—does not resolve all the problems associated with it. In this context, it seems important to explore other legal instruments aimed at comprehensively regulating scientific activities in the digital context.

52 Art. 26²(3) Ustawa o zmianie ustawy o prawie autorskim i prawach pokrewnych, ustawy o ochronie baz danych oraz ustawy o zbiorowym zarządzaniu prawami autorskimi i prawami pokrewnymi z dnia 26 lipca 2024 r. (Dz.U. z 2024 r. poz. 1254).



Member States are obligated to safeguard this exception against contractual override. This is especially important in the context of licensing agreements entered into by beneficiaries of this exception with database providers. Polish law does not contain a specific provision implementing such protection. In the course of preparing the legislation, it was indicated that: "provisions of the Copyright Act concerning permitted use (Art. 23–35) leave no doubt that they apply regardless of the will of the rights holders, and thus also regardless of any contractual provisions between the rights holder and the beneficiary of the permitted use."[53] The approach adopted by the Polish legislator is difficult to consider correct. First, it is important to highlight that there is a divergence of views in the doctrine on this issue. Some legal scholars argue that the provisions on permitted use are indeed imperative (or semi-imperative), while others believe that it is possible to contractually exclude their application. The lack of consistency in the doctrine in this area, coupled with the absence of case law addressing this issue, means that the position adopted by the legislator lacks strong justification and is not, in itself, a source of law.[54]

Second, even if one assumes that contractual provisions cannot effectively limit the scope of permitted use, using a work in violation of such a provision may still result in liability for breach of contract. This situation creates legal uncertainty and may have a chilling effect. It is crucial to directly regulate this issue, as users often lack knowledge about the legal nature of exceptions and base their decisions on the wording of the provisions.

This problem affects both individual users, such as ordinary citizens who typically accept the terms of agreements automatically, and public institutions that enter into contracts with clauses limiting the scope of permitted use. For such institutions, the legal uncertainty as to whether violating a contractual provision leads to an infringement of copyright law (assuming the non-imperative nature of the provisions) or merely to contractual liability is not so important. In both cases, it may lead public institutions to refrain from using works within the scope of permitted use.

5. GENERAL EXCEPTION OR LIMITATION FOR TEXT AND DATA MINING

a. Beneficiaries, permitted uses and subject matter

The TDM exception for scientific research is based on an open formula indicating that, in the absence of a specific reservation, "it is permissible to reproduce disseminated works for the purpose of text and data mining".[55] This construction means that any entity can benefit from this exception. Such an entity can, therefore, reproduce works of any type (textual, musical, graphic, video, etc.) and in any form and format (particularly in digital formats) for the purpose of TDM. However, the use of computer programs for TDM purposes may be problematic. While the exception allows for the reproduction of such programs, the exclusive rights also cover "translations, adaptations, rearrangements, or any other modifications of the computer program."[56] In many cases, utilising computer programs in this context will require stepping into rights beyond just the right to reproduce.[57]

b. Lawful access v. disseminated work

The only limitation introduced by the Polish legislator is that these works must have been previously disseminated. According to Polish copyright law,[58] a *disseminated work* is that "which, with the permission of its author, has been made available in any manner to the public". However, the Polish concept of a *disseminated work* is not equivalent to the condition of a "lawfully accessible work" as used in the CDSM Directive. The dissemination of a work pertains to the status of the work itself rather than the status of individual copies of it. A work could, therefore, be considered disseminated under Polish law while simultaneously not being a "lawfully accessible work" by the beneficiary. The condition specified in the directive will, therefore, be met first when the rights holder grants the beneficiary appropriate permission to access the work

53 Tabela zgodności, https://legislacja.rcl.gov.pl/docs//2/12382002/13037388/13037389/dokument656773.pdf, p. 14.

54 See: K. Gliściński, Komentarz do art 17, [w] A. Michalak, Ustawa o prawie autorskim i prawach pokrewnych. Komentarz, Warszawa 2019, p. 205–206.

55 Art. $26^2(1)$ Ustawa o zmianie ustawy o prawie autorskim i prawach pokrewnych, ustawy o ochronie baz danych oraz ustawy o zbiorowym zarządzaniu prawami autorskimi i prawami pokrewnymi z dnia 26 lipca 2024 r. (Dz.U. z 2024 r. poz. 1254).

56 Art. 74(4)(2) Ustawa o prawie autorskim i prawach pokrewnych z dnia 4 lutego 1994 r. (Dz.U. z 2022 r. poz. 2509).

57 B. Widła, Programy komputerowe jako przedmiot eksploracji tekstów i danych w kontekście dyrektywy 2019/790, Europejski Przegląd Sądowy nr 3(210)/2023, p. 13–14.

58 Art. 6(1)(3) Ustawa o prawie autorskim i prawach pokrewnych z dnia 4 lutego 1994 r. (Dz.U. z 2022 r. poz. 2509).

(e.g., through a licensing agreement or an open access policy) or second when the work is available without any legal restrictions (e.g., placed on the internet by the rights holder). On the other hand, if a work has been disseminated with the rights holder's permission (e.g., in digital form), but the beneficiary accesses an electronic version of the book from an illegal source, the condition specified in the directive is not met (even though the work is considered disseminated under Polish law). From this perspective, it can be stated that the condition of disseminating a work protects the creator from situations where works are used under permitted uses before their first public release. It is, therefore, related to the moral right of the author "to decide about making the work available to the public for the first time."[59] On the other hand, the condition specified in the Directive pertains to the protection of economic interests related to lawful access to individual copies of the work. As a consequence, the introduction of the requirement for "dissemination of works" in place of "lawful access" may be regarded as incompatible with EU law. In this context, it was pointed out that the absence of this requirement is not necessarily an issue, as Polish law includes a clause referring to the three-step test. Thus, under permitted use, one cannot use works that have been made available illegally.[60] However, such an approach may raise certain doubts.

c. Opt-out mechanism

Art. 4(3) of the CDSM Directive stipulates that the general exception for TDM applies unless it has been expressly reserved by the right holder in an appropriate manner. The Polish legislator, when implementing this solution, specified that such reservations must be made "explicitly and in a manner appropriate to the way in which the work was made available. In the case of works made publicly available in such a way that anyone can access them at a time and place of their choosing, the reservation must be made in a machine-readable format as defined in Art. 2(7) of the Act of 11 August 2021 on open data and the re-use of public sector information),[61] along with metadata".[62] According to this latter provision, a machine-readable format means "a file format structured in such a way that computer programs can identify, recognize, and retrieve specific data and their internal structure."[63] This article, in turn, implements the definition of "machine-readable format" as outlined in Art. 2(13) of Directive 2019/1024. Examples of such formats include XML, JSON, RDF, and CSV.[64]

The legislator has not specified how such a reservation should be made when making works available through other means. Essentially, according to Recital 18 of the CDSM Directive, this can occur through, among other means, "contractual agreements or a unilateral declaration".[65] While in the case of access to works in electronic format, a contractual reservation seems conceivable (e.g., in licensing terms), it is less likely to occur with works available in analogue formats (e.g., printed books). In this latter case, unilateral reservations become significant. It seems that such a reservation should be made on every copy of the work in question. A general reservation, for instance, on the publisher's website or in accompanying materials, may prove to be insufficient. From a practical standpoint, such a reservation can be made alongside the traditional copyright notice typically found in books.

At the same time, in both cases, the legislator did not determine the specific wording of such a reservation. He merely indicated that it should be *explicit*. This means that the content of the reservation should clearly state the prohibition against reproducing the works for text and data mining purposes. On the one hand, it can be argued that using the traditional phrases *all rights reserved* or *no copying allowed*, without explicitly linking them to a prohibition on using the work for TDM purposes, would not meet the requirement for an explicit reservation. On the other hand, it does not seem necessary to cite specific articles from the law or directive to fulfil this requirement. For works distributed digitally but not made publicly available in a manner where anyone can access them at any time and place of their choosing (e.g., music on CDs). However, it seems that the requirement for an *explicit* reservation supports the view that such a reservation should also be made in natural language (e.g., on the packaging of a CD) so that it can be reviewed before purchase. In cases where the reservation does not meet the aforementioned requirements, it should be considered ineffective against individuals conducting TDM activities based on improperly marked or unmarked copies of the work. It is difficult to assert that a purchaser of a work is obliged to seek such a reservation beyond the copy being acquired. Of course, issues related to the effective manner of making reservations have already been raised at the level of the directive itself. However, the Polish legislator did not decide to introduce any specific regulations in this regard.

At the same time, it should be emphasised that opting out does not preclude conducting TDM for scientific

59 Art. 16(4) Ustawa o prawie autorskim i prawach pokrewnych z dnia 4 lutego 1994 r. (Dz.U. z 2022 r. poz. 2509).

60 Raport z konsultacji publicznych projektu ustawy o zmianie ustawy o prawie autorskim i prawach pokrewnych oraz niektórych innych ustaw – załącznik do Oceny Skutków Regulacji https://legislacja.rcl.gov.pl/docs/2/12360954/12887995/12887998/dokument587349.pdf (19.07.2024), p. 15.

61 Ustawa o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego (Dz.U. z 2023 r. poz. 1524).

62 Art. 26^3(2) Ustawa o zmianie ustawy o prawie autorskim i prawach pokrewnych, ustawy o ochronie baz danych oraz ustawy o zbiorowym zarządzaniu prawami autorskimi i prawami pokrewnymi z dnia 26 lipca 2024 r. (Dz.U. z 2024 r. poz. 1254).

63 Art. 2(7) Ustawa o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego (Dz.U. z 2023 r. poz. 1524).

64 Art. 2 OtwDaneU red. Sibiga/Sybilski 2022, wyd. 1/Garstka/Gos/Sibiga/Sybilski/Szelenbaum, G. Sibiga, D. Sybilski (red.), Ustawa o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego. Komentarz, Warszawa 2022.

65 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 (Text with EEA relevance.), OJ L 130, 17.5.2019, p. 92–125, rec. 18.

research, nor does it limit activities that are not covered by exclusive rights or those performed with unprotected elements of works.

d. The retention period for copies of reproduced works

Following the text of the directive, the Polish legislator stated that works reproduced under the discussed exception "may be stored solely for the purpose of text and data mining, and only for as long as is necessary to achieve that purpose."[66] This construction, however, leaves some uncertainty regarding the duration for which such copies may be stored. On the one hand, a narrow interpretation of this purpose suggests that once the TDM process is completed, these copies should be deleted. On the other hand, the TDM process can be understood more broadly, encompassing not only the preparation phase and the TDM itself but also subsequent verification activities. These verification activities may be carried out shortly after the TDM or much later.

6. PROTECTION OF BENEFICIARIES FROM TPMS

According to Art. 7(2) of the CDSM Directive, the discussed exception is subject to Art. 6(4) of the InfoSoc Directive. The Polish Copyright Act does not explicitly regulate mechanisms for protecting the beneficiaries of exceptions from Technological Protection Measures (TPMs). In the justification for the draft implementing the CDSM Directive, it was indicated that there is no need to implement Art. 7(2).[67] This approach is based on the assumption that Polish copyright law provisions regarding liability for the removal or circumvention of TPMs (Art. 79(6)) allow for "the removal and circumvention of technical protections if it is intended for the lawful use of works (e.g., within the scope of exceptions for public use)".[68] While it may be agreed that such behaviour is permissible under the current Polish legal framework, the question remains whether this solution complies with Art. 6 of the InfoSoc Directive.[69]

7. CONCLUSIONS

As I have explained throughout this article, the Polish implementation of both TDM exceptions may raise certain concerns. Given that the implementation has only just come into effect, there is a lack of extensive commentary in the legal doctrine on this matter. The chosen method of implementation, largely based on a copy-paste approach, also fails to address many of the issues that were raised concerning the text of the directive. In particular, it does not resolve the issues related to the process of opting out. It remains an open question as to how these provisions will be applied in practice. Will the concerns outlined in this presentation actually translate into practical difficulties in their use? Specifically, will they give rise to legal disputes? All these questions will expectedly find their answers in time.

Konrad Gliściński

PhD in Law. Researcher at the Department of Intellectual Property Law at the Jagiellonian University. Legal advisor to the Board of the Jagiellonian Center of Innovation. Intellectual property expert at Centrum Cyfrowe. Lecturer at the H. Grotius Center for Intellectual Property Rights. He collaborates with the Kalecki Foundation. He completed doctoral studies at the Faculty of Law and Administration of the Jagiellonian University in Cracow and post-graduate studies in company law at the Warsaw School of Economics. He holds an LL. M. of the University of Turin in the field of intellectual property law. A graduate of the Top500 Innovators program at Stanford University on management and commercialization of scientific research. Laureate of the 2010 Minister of Science and Higher Education award for best master's thesis organized by the Patent Office of the Republic of Poland. Author of "All rights reserved. The history of disputes over copyright. 1469 – 1928". Co-author of the commentary on the Industrial Property Law Act (2016) and the Copyright and Related Law Act (2019).

66 Art. 26³(2) Ustawa o zmianie ustawy o prawie autorskim i prawach pokrewnych, ustawy o ochronie baz danych oraz ustawy o zbiorowym zarządzaniu prawami autorskimi i prawami pokrewnymi z dnia 26 lipca 2024 r. (Dz.U. z 2024 r. poz. 1254).

67 Tabela zgodności, https://legislacja.rcl.gov.pl/docs//2/12382002/13037388/13037389/dokument656773.pdf, p. 14.

68 A. Matlak, T. Targosz, E. Traple [w:] Komentarz do ustawy o prawie autorskim i prawach pokrewnych [w:] Ustawy autorskie. Komentarze. Tom II, red. R. Markiewicz, Warszawa 2021, art. 79, s. 1188.

69 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, 22/06/2001, p. 0010 – 0019.

TDM Exception or Limitation – Methodology of Implementation in the EU Member States: Creating Cohesion or Diversion?

Branka Marušić

ABSTRACT

This article examines the margin of appreciation of the EU Member States on the choice and formulation of the E&Ls when implementing them into their national law. It does so, firstly by explaining the methods and terminology used to assess implementation of directives. It then continues with the cartography of E&Ls prior to and after the enactment of the DSM Directive in the research sector. Finally, this article concludes with remarks on the future viability of the TDM exception.

1. INTRODUCTION

The main reason for the introduction of the text and data mining (TDM) exception within 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 (DSM Directive)[1] was to support the European research organisations' scientific work. The problem that the research organisations *were* and *are* facing – all over the globe and not exclusively in Europe, is the legal uncertainty as to whether TDM activities are infringing copyright. The problem is mostly vested in the diverging national solutions that address this problem in a form of the existence – or lack – of exceptions that cover such activities.

For clarification purposes, the activities of TDM – in a broader sense, could be understood as different types of computational processes that aim at discovering patterns in large databases and/or collections of textual content, as well as extracting information from previous sources (e.g., existing dataset and collection of journal articles) and transforming it into information that can be used for further purposes (e.g., analysis or pattern discovery). From a copyright perspective, these types of activities forming part of the computational process can attract several different economic rights of rightsholder – be it in copyright or related rights. These economic rights can be right of reproduction – right to copy parts or whole items of protected objects; adaptation right to change and transform protected objects; translation – right to translate from one language to another; extraction and re-utilisation of the *sui generis* database right– parts of database sets; and making available – creating and enabling access online to protected objects. copyright framework, to some extent, can shield users from copyright and related rights infringement claims by providing exceptions or limitations (E&L) to these rights.

The margin of appreciation in the choice of creating E&Ls – as well as formulating them in national laws, is the bedrock in the aforementioned TDM activities problem. On the European Union (EU) level, the margin of appreciation of the EU Member States on the choice and formulation of the E&Ls is somewhat bound within the EU-wide harmonisation measures. This article aims to explore the boundaries of the 'margin of appreciation' –by examining how harmonisation measures, specifically directives, are assessed and implemented into the national laws of the EU Member States. In order to achieve this, this article first explains the methods and terminology used to assess an implementation. It then continues with the cartography of E&Ls prior to and after the enactment of the DSM Directive in the research sector. Finally, this article concludes with remarks on the future viability of the TDM exception.

2. METHODS AND TERMINOLOGY FOR ASSESSING IMPLEMENTATION

On the EU level, harmonisation of copyright has been predominantly achieved by the use of directives by the EU

1 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] OJ L 130/92 ('DSM Directive').

legislator. The EU directives, by virtue of Art. 288 (3) of the Treaty on the Functioning of the European Union,[2] bind EU Member States regarding the result to be achieved, but they leave it to national authorities to choose the form and methods for implementation. These 'results' – of implementation – need to be the same for the territory of the EU, but form and method of implementation is in the national purview of the EU Member States.[3] Generally speaking there are two ways of assessing whether these countries have achieved the 'results' aimed by the directive. The first one is the *prima facie* assessment, where the result is measured on the contextual analysis of the national law. The second one is the *impact assessment,* where the result is measured on the 'law in action, to put it differently, on how the implemented directive operates in practice. This article focuses on the first type of assessment—the *prima faciae*

In the *prima faciae* assessment, the two most prominent methods of contextual assessment are one of literal transposition and one of a flexible approach. A provision of a directive has been literally transposed, if it has been adopted verbatim into national law, meaning 'copied and pasted'. The provisions most likely to be transposed in that manner are provisions which should be exactly or to a high degree worded the same as in the directive. These provisions are the ones consisting of definitions contained in the directive, start with 'shall', or contain full harmonisation and/or maximum standards.

The flexible approach, as the name suggests, provides leeway to EU Member States with the wording and framing of the provision of a directive when transposed or reflected into national law. These are the provisions that start with 'may'; allow EU Member States to provide 'more detailed or stricter rules', and contain partial harmonisation and/or minimum standards. The flexible approach hides a danger of EU Member States going beyond the 'results to be achieved' by providing more favourable terms in the form of gold-plating provisions.

For ease of clarity regarding the terminology used in the previous paragraphs, the scope and intensity of harmonisation requires explanation, as well as what does a gold-plating provision entail. Harmonisation is 'full' in scope when there is comprehensive or exhaustive harmonisation in a specific area; harmonisation will otherwise be 'partial' in scope.[4] Partial harmonisation can be vertical or horizontal in scope.[5] The former referring to harmonising rules for specific products or services, for example databases in copyright in Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.[6] The latter referring to a legal act covering all or several different products and services, for example 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 (InfoSoc Directive).[7] In addition, the notion of targeted harmonisation refers to measures that provide only very selectively for harmonised rules. An example of this can be found in Art. 17 of the DSM Directive – which is a specific *sui generis* liability regime for platform copyright infringement liability.

Distinct from the scope of harmonisation, the standard(s) set may also vary in their intensity. They may provide for 'full' (or 'maximum' or 'total') harmonisation, in the sense of setting standards which nation-states cannot derogate from, or they may provide for 'minimum' harmonisation only, leaving some discretion to nation states to go beyond.[8] A nation-state implementing this standard can go above it, but not below.[9] Conversely, when implementing a standard of maximum harmonisation, nation-states may not introduce stricter rules. A maximum standard therefore serves as a regulative limit.[10]

Last, the term gold-plating describes a transposition or implementation EU directives in its national law, where the EU Member State uses the opportunity to impose additional requirements, obligations, or standards on the addressees of its national law that go beyond the requirements or standards foreseen in the EU directives.[11]

In a situation where a directive is an amendment directive or a part of the legislative package/series of directives, and even more so if it implements and/or reflects international obligations of the EU and/or the EU Member States (or both), the case law of the Court of Justice of the European Union (CJEU) is incorporated in the contextual assessment. The case law that is incorporated in such contextual assessment is the one that defines words that are found in the directives – and these words are found in the new amendment directive. These words defined by the CJEU can be seen in the case law labelled

2 Treaty on the Functioning of the European Union [2008] OJ C 326/47 (the 'TFEU').

3 For a detailed account on form and method see Richard Král, 'On the choice of methods of transposition of EU Directives' (2016) 41(2) ELR 220.

4 Marcus Klamert, 'What We Talk About When We Talk About Harmonisation' (2015) 17 CYELS 360, 362–363.

5 Ibid, 362.

6 Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [1996] OJ L 77/20 ('Database Directive').

7 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 [2001] OJ L 167/10 ('InfoSoc Directive').

8 Klamert (n 4) 362.

9 Stephen Weatherill, 'Maximum versus Minimum Harmonization: Choosing between Unity and Diversity in the Search for the Soul of the Internal Market' in Niamh Nic Shuibhne and Laurence W Gormley (eds), *From Single Market to Economic Union: Essays in Memory of John A. Usher* (OUP 2012) 175, 176.

10 Ibid.

11 European Commission 'Better Regulation Guidelines' SWD (2017) 350 1, 88.

Table 1

State-of-the-art of E&L prior to the DSM Directive in the research sector (The legislative landscape for E&L for research activities)

Object of Protection	Rightholder	Economic Right	E&L
Original Database (copyright)	Author	Reproduction (CJEU defined autonomous legal concept concept)[16] Translation, adaptation, arrangement and any other alteration Distribution (CJEU defined autonomous legal concept concept) [17] Any communication, display or performance to the public	Art. 6(2)(b) Database Directive 'sole purpose of illustration for teaching or scientific research'
Sui Generis Database (CJEU defined concept) [18]	Database rightsholder	Extraction (CJEU defined concept)[19]	Art. 9(b) Database Directive 'sole purpose of illustration for teaching or scientific research'
Original works (CJEU defined autonomous legal concept concept)[20]	Authors	Reproduction (CJEU defined autonomous legal concept concept)[21] Communication to the public including making available (CJEU defined autonomous legal concept concept) [22]	Art. 5(3)(a) InfoSoc Directive 'sole purpose of illustration for teaching or scientific research'
Fixations of performances Phonograms Original and copies of films-Fixations of broadcast	Performers (CJEU defined concept) [23] Phonogram producers Producers of the first fixations of films Broadcasting organisations	Reproduction (CJEU defined autonomous legal concept concept)[24] Making available (CJEU defined autonomous legal concept concept) [25]	Art. 5(3)(a) InfoSoc Directive 'sole purpose of illustration for teaching or scientific research'
Fixations of performances Phonograms Fixations of broadcast	Performers (CJEU defined concept) [26] Phonogram producers Broadcasting organisations	Communication to the public (CJEU defined autonomous legal concept concept) [27]	Art. 10(1)(d) Rental and Lending Rights Directive[28] 'solely for the purposes of teaching or scientific research'

as concepts,[12] autonomous legal concepts of EU law,[13] and as principles.[14]

Thankfully in copyright we have the whole set. The DSM Directive is the add on to the existing series of directives. Art. 3 and 4 of the DSM Directive – the new mandatory TDM exceptions contain the concept of 'lawful access' (through recital 8 and 11) which resembles the CJEU's defined concept of 'lawful use'[15] from the transient copy exception found in Art. 5 (1) (b) of the InfoSoc Directive.

Notwithstanding the above explanation of terminology, it is relevant to examine the types of E&Ls contained in the EU copyright framework that shield users from TDM activates infringing copyright. More importantly, the *prima faciae* assessment provides valuable clues on their uniformity and viability.

3. CARTOGRAPHY OF E&LS PRIOR TO AND AFTER THE ENACTMENT OF THE DSM DIRECTIVE

According to a very simple portrayal, there first needs to be – in order to attract the protection of copyright or a related right – an object of protection linked with the

12 See 'extraction' and 're-utilisation' in Judgment in *The British Horseracing Board and Others*, C-203/02, EU:C:2004:695 paras 47–53.

13 See fair compensation in Judgment in *Padawan*, C-467/08, EU:C:2010:620 para 33.

14 See exhaustion in Judgment in *Vereniging Openbare Bibliotheken*, C-174/15, EU:C:2016:856 paras 58–59.

15 See Judgment in *Football Association Premier League Ltd and Others v QC Leisure and Others (C-403/08) and Karen Murphy v Media Protection Services Ltd (C-429/08)*, C-403/08 and C-429/08, EU:C:2011:631 paras 167–173; Judgment in *Stichting Brein (Filmspeler)*, C-527/15, EU:C:2017:300 paras 64–71.

16 See Judgment in *Infopaq International (I)*, C-5/08, EU:C:2009:465 para 27.

17 See Judgment in *Dimensione Direct Sales and Labianca*, C-516/13, EU:C:2015:315 para 22.

18 See Judgment in *Apis-Hristovich*, C-545/07, EU:C:2009:132 paras 62–73.

19 See Judgment in *Directmedia Publishing*, C-304/07, EU:C:2008:552 paras 22–47.

20 See Judgment in *Cofemel*, C-683/17, EU:C:2019:721 paras 29–35.

21 See Judgment in *Infopaq International (I)* (n 16).

22 See Judgment in *Circul Globus București*, C-283/10, EU:C:2011:772 para 32.

23 See Judgment in *Recorded Artists Actors Performers*, C-265/19, EU:C:2020:677 paras 49–54.

24 See Judgment in *Infopaq International (I)* (n 16).

25 See Judgment in *Circul Globus București* (n 22).

26 See Judgment in *Recorded Artists Actors Performers* (n 23).

27 See Judgment in *Circul Globus București* (n 22).

28 Directive 2006/115/EC of the European Parliament and of the Council of 12 December 2006 on rental right and lending right and on certain rights related to copyright in the field of intellectual property (codified version) [2006] OJ L 376/28 ('Rental and Lending Rights Directive').

Table 2

State-of-the-art of E&L prior to the DSM Directive in the research sector (The legislative landscape for L&E with a research activities flavour)

Object of Protection	Rightsholder	Economic Right	E&L
Computer programs (original) (CJEU defined autonomous legal concept concept)[29]	Author	Reproduction (CJEU defined autonomous legal concept concept)[30] Translation, adaptation, arrangement and any other alteration	Art. 5 & 6 Software Directive[31] Interoperability and decompilation of software in individual research activities (mandatory exception) Decompilation (CJEU defined concept) [32]
Sui Generis Database (CJEU defined concept) [33]	Database rightsholder	Extraction and re-utilisation (CJEU defined concept)[34]	Art. 8 Database Directive lawful users can extract or re-utilise insubstantial
Original works (CJEU defined autonomous legal concept concept)[35] Fixations of performances Phonograms Original and copies of films Fixations of broadcast	Authors Performers (CJEU defined concept) [36] Phonogram producers Producers of the first fixations of films Broadcasting organisations	Reproduction (CJEU defined autonomous legal concept concept)[37]	Art. 5(1) InfoSoc Directive Transient copies (CJEU defined concept)[38] (mandatory exception)
Original works (CJEU defined autonomous legal concept concept)[39] Fixations of performances Phonograms Original and copies of films Fixations of broadcast	Authors Performers (CJEU defined concept) [40] Phonogram producers Producers of the first fixations of films Broadcasting organisations	Reproduction (CJEU defined autonomous legal concept concept)[41] Making available (CJEU defined autonomous legal concept concept) [42]	Art. 5(2)(b) InfoSoc Directive Private copy exception (CJEU defined concept)[43] Art. 5(2)(c) InfoSoc Directive Reprography exception (CJEU defined concept)[44] For both: Fair compensation (CJEU defined autonomous legal concept) [45]
Original works (CJEU defined autonomous legal concept concept)[46]	Authors	Reproduction (CJEU defined autonomous legal concept concept)[47] Communication to the public including making available (CJEU defined autonomous legal concept concept) [48]	Art. 5(3)(d) InfoSoc Directive Quotation (CJEU defined concept) [49]
Fixations of performances Phonograms Original and copies of films Fixations of broadcast	Performers (CJEU defined concept) [50] Phonogram producers Producers of the first fixations of films Broadcasting organisations	Reproduction (CJEU defined autonomous legal concept concept)[51] Making available (CJEU defined autonomous legal concept concept) [52]	Art. 5(3)(d) InfoSoc Directive Quotation (CJEU defined concept) [53]
Original works (CJEU defined autonomous legal concept concept)[54]	Authors	Communication to the public including making available (CJEU defined autonomous legal concept concept) [55]	Art. 5(3)(n) InfoSoc Directive Use for the purpose of research or private study by dedicated terminals on the premises of establishments (CJEU defined concept) [56]
Fixations of performances Phonograms Original and copies of films Fixations of broadcast	Performers (CJEU defined concept) [57] Phonogram producers Producers of the first fixations of films Broadcasting organisations	Making available (CJEU defined autonomous legal concept concept) [58]	Art. 5(3)(n) InfoSoc Directive Use for the purpose of research or private study by dedicated terminals on the premises of establishments (CJEU defined concept) [59]

rightholders (who are involved in the creation or existence of the object of protection), followed by rights that derive from this protection, and finally limits to these rights. The patchwork of the copyright legislative framework in the EU, together with its interpretation by the CJEU link the aforementioned four broad categories together in an aim to create a coherent system. Prior to the enactment of the DSM Directive, the cartography of E&Ls in the research sector that shielded researchers from TDM activates copyright infringement claims can be seen in Table 1.

Furthermore, there also existed the E&Ls with a research sector 'flavour' that shielded researchers from TDM activates copyright infringement claims and they are listed below in Table 2.

In the *prima faciae* assessment of these provisions, a flexible approach of contextual assessment was taken. This was since the E&Ls with a TDM flavour, are optional – save from the exception on temporary reproduction and the E&Ls contained in the Software Directive. This means that EU Member States are able to cherry-pick the exact scope of the E&L, which has resulted in varying formulation and intensity[60] of the 'TDM flavour' E&L. To put it simply, EU Member States were given an option on linking an E&L with a specific object of protection, rightsholder and economic right and providing a variety of different solutions for essentially the same TDM activity.[61] This, in turn, has been criticised by some scholars,[62] who contend that the aim of harmonising copyright on the EU-wide level has not been met in its full form because the optional E&Ls have only a minimal harmonising character; and without the implementation guidelines, Member States have often implemented a narrower scope than was foreseen by the directives.[63] However, there are two limits to the EU Member State margin of appreciation: The first one is in the form of an interpreted concept by the CJEU, and the second relies on the fact that the list of E&Ls is a closed one. Providing an E&L that is outside of the enumerated list in the copyright harmonisation framework on the EU level could amount to a gold-plating provision.

The introduction of the mandatory TDM E&Ls in Art. 3 and 4 of the DSM Directive adds to the variety of national solutions without bringing uniformity. This primarily relates to the fact that, here as well, the contextual assessment method is one of a flexible approach. More importantly, the flexibility starts with the definition of the TDM in Art. 2(2) of the DSM Directive where the EU Member States can add to it (e.g.by including elements of recital 8 and 11) and/or subtract (by omitting parts of the definition). The problem that arises from this approach is that there is an uneven scope of the definition itself found in the national transposing measures. Nevertheless, this flexible approach becomes more stringent in the assessment of the body of text of Art. 3 and 4 of the DSM Directive. The narrowing of the flexibility in the assessment approach can be seen for example in the approach that rightsholders and economic rights are full harmonisation – Member States have no discretion in defining what they are. On the other hand, broadening or removing the scope of economic rights as well as rightsholders by Member States in instances of full harmonisation, by adding or subtracting could be considered gold-plating provision.

29 See Judgment in *Cofemel* (n 20).

30 See Judgment in *Infopaq International (I)* (n 16).

31 Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs (Codified version) (Text with EEA relevance) [2009] OJ L 111/16 ('Software Directive').

32 See Judgment *Judgment in Top System SA v État belge,* C-13/20, EU:C:2021:811 para 40.

33 See Judgment in *Apis-Hristovich* (n 18).

34 See Judgment in *The British Horseracing Board and Others* (n 12).

35 See Judgment in *Cofemel* (n 20).

36 See Judgment in *Recorded Artists Actors Performers* (n 23).

37 See Judgment in *Infopaq International (I)* (n 16).

38 See Judgment in *Football Association Premier League* (n 15) paras 161–179.

39 See Judgment in *Cofemel* (n 20).

40 See Judgment in *Recorded Artists Actors Performers* (n 23).

41 See Judgment in *Infopaq International (I)* (n 16).

42 See Judgment in *Circul Globus București* (n 22).

43 See Judgment in *Copydan Båndkopi*, C-463/12, EU:C:2015:144 paras 68–73.

44 See Judgment in *Eugen Ulmer*, C-117/13, EU:C:2014:2196 paras 47–49.

45 See Judgment in *Padawan* (n 13) paras 29–37.

46 See Judgment in *Cofemel* (n 20).

47 See Judgment in *Infopaq International (I)* (n 16).

48 See Judgment in *Circul Globus București* (n 22).

49 See Judgment in *Painer*, C-145/10, EU:C:2011:798 paras 130–137.

50 See Judgment in *Recorded Artists Actors Performers* (n 23).

51 See Judgment in *Infopaq International (I)* (n 16).

52 See Judgment in *Circul Globus București* (n 22).

53 See Judgment in *Painer* (n 49).

54 See Judgment in *Cofemel* (n 20).

55 See Judgment in *Circul Globus București* (n 22).

56 See Judgment in *Eugen Ulmer* (n 44) paras 38–40.

57 See Judgment in *Recorded Artists Actors Performers* (n 23).

58 See Judgment in *Circul Globus București* (n 22).

59 See Judgment in *Eugen Ulmer* (n 56).

60 Thomas Dreier, 'Limitations: The Centrepiece of Copyright in Distress' (2010) 1(2) JIPITEC 50,52.

61 For exact formulation, scope and intensity of national solutions please see de Francquen A, Dusollier S, Triaille J-P, Hubin J-B, Depreeuw S, Coppens F, 'Study on the application of Directive 2001/29/EC on copyright and related rights in the information society (the "Infosoc Directive") (2013)'; Brigitte Lindner and Ted Shapiro (eds), *Copyright in the Information Society: A Guide to National Implementation of the European Directive* (2nd edn, Edward Elgar 2019); Caterina Sganga et al, 'Copyright flexibilities: mapping and comparative assessment of EU and national sources' (2023) https://zenodo.org/record/7540511#.Y8Uss3bM.

62 Lucie Guibault, 'Why Cherry-Picking Never Leads to Harmonisation: The Case of the Limitations on Copyright under Directive 2001/29/EC' (2010) 1(2) JIPITEC 55; Mireille van Eechoud, Bernt P. Hugenholtz, Stef van Gompel, Lucie Guibault, Natali Helberger, *Harmonizing European Copyright Law: The Challenges of Better Lawmaking* (Kluwer Law International 2009) 94–120.

63 Christophe Geiger and Franciska Schönherr, 'Defining the Scope of Protection of Copyright in the EU: The Need to Reconsider the Acquis regarding Limitations and Exceptions' in Tatiana-Eleni Synodinou (ed), *Codification of European Copyright Law: Challenges and Perspectives* (Kluwer Law International 2012)139.

4. FINAL REMARKS

Taking a flexible approach in the *prima facie* methodology for assessing the implementation of provisions of mandatory TDM E&Ls does not add to the creation of legal certainty for the researchers that wish to avail themselves to them. This narrow add-on to the existing cartography of 'chaotic' E&Ls with a TDM flavour can be labelled as a missed opportunity to make a narrow yet mandatory provision functionally harmonised in the territory of the EU. This is since the flexible approach provides leeway in the 'form and method' of implementation to EU Member States.

Adding to this, there is still no clear guidance on situations where a computational process of a TDM falls outside of the scope of Art. 3 or 4 of the DSM Directive and into the scope of another E&L – a question that is still quite dependent on the territory of the computational process and its cross-border reach.

Moving forward, we can expect at least three scenarios. The first scenario is that the TDM E&Ls could become redundant by the licensing schemes between publishers and research and cultural heritage institutions. Alternatively, they can also become redundant by open-access initiatives. In both scenarios the publishers benefit from these outcomes since in their role as middle management – between creators and users – they receive remuneration in form of licensee fees and open access fees. The third scenario is that the CJEU interprets the TDM E&Ls in a manner in which such interpretation would not be operational for the researchers.

Branka Marušić

Branka Marušić, Stockholm University, Sweden, is an Associate senior lecturer in intellectual property law.

Textual Insights: What Can Computers Teach Legal Scholars About Law?

Johan Lindholm

ABSTRACT

Legal research has historically relied on the manual and systematic study of authoritative texts, a methodology that has remained largely unchanged despite technological advancements. However, recent developments in natural language processing and other data-driven approaches present new opportunities for legal scholars. This essay examines whether and how these computational tools can complement doctrinal approaches and explores the potential of computational methods to enhance and transform legal scholarship. In emphasizing the compatibility of computational and doctrinal approaches, it argues that by integrating these approaches, legal scholars can make scientific discoveries beyond the scope of either method alone. The essay concludes by outlining the steps necessary for legal scholarship to fully embrace and benefit from these emerging technologies.

Keywords: legal scholarship, computational methods, empirical legal studies, natural language processing, large language models.

1. INTRODUCTION

For centuries, lawyers, judges, law students, and legal scholars alike would gather in law libraries to 'do legal research'. By 'doing legal research' they meant roughly the same thing: to carefully study authoritative texts in order to determine what the law governing a particular issue 'is', that is to say what is commonly referred to as conducting doctrinal legal research. They all used more or less the same methodology which centered around the manual and systematic reading of authoritative legal sources, although they did this for slightly different motives, e.g. to advise and represent clients, to justify judgments, to learn the law, and to improve the legal system.[1]

At the turn of the twentieth century, legal realists on both sides of the Atlantic challenged traditional conceptions of law. Rejecting what they viewed as the metaphysical elements of law, they instead saw it as an inherently social phenomenon. For instance, under this perspective, a statement about what the law is could be understood as a prediction of how judges will apply it to specific facts.[2] One could imagine that this would change how legal research was conducted. In America, this was also to some extent the case. Realists accepting law as inherently entangled in a messy social and political reality[3] would pave the way for a methodological turn towards the empirical.[4] This was less true in Europe where legal scholars and practising lawyers largely continued going to the law library to find and read the authoritative legal documents, largely in the same way as each other, and largely using the same methods as before.

Even the technical revolution that took place in the 1990s did not fundamentally change how legal research was conducted. The main contribution that general access to affordable personal computers, the invention of CD-roms, and even the introduction of the Internet made to legal research was that the authoritative sources traditionally studied now could be accessed digitally while the paper versions collected dust on library shelves.[5] This

1 See e.g. Rob van Gestel and Hans-Wolfgang Micklitz, 'Why Methods Matter in European Legal Scholarship: Methods in European Legal Scholarship' (2014) 20 European Law Journal 292; Terry Hutchinson and Nigel Duncan, 'Defining and Describing What We Do: Doctrinal Legal Research' (2012) 17 Deakin Law Review 83.

2 Alf Ross, *On Law and Justice* (Jakob vH Holtermann ed, Uta Bindreiter tr, Oxford University Press 2019) 156; see also Éric Millard, 'Alf Ross and Realist Conceptions of Legislation' in Pierre Brunet, Éric Millard and Patricia Mindus (eds), *The Theory and Practice of Legislation* (Hart Publishing 2013); Oliver Wendell Holmes, 'The Path of the Law' (1997) 110 Harvard Law Review 991, 994 ("The prophecies of what the courts will do in fact, and nothing more pretentious, are what I mean by the law.").

3 Gregory S Alexander, 'Comparing the Two Legal Realisms-American and Scandinavian' (2002) 50 The American Journal of Comparative Law 131, 133.

4 Michael Heise, 'The Past, Present, and Futute of Empirical Legal Scholarship: Judicial Decision Making and the New Empiricism' (2002) 2002 University of Illinois Law Review 819, 822–824.

5 Despite some valiant efforts towards change. For a historical overview, including of the efforts that were made, see Peter Wahlgren, 'The Quest for Scientific Methods: Sociology of Law, Jurimetrics and Legal Informatics' in Håkan Hydén and others (eds), *Combining the Legal and the*

was true for practising lawyers as well as for most legal scholars.

We find ourselves more recently in the middle of another technological revolution. Increasingly easy access to large, accurate, and accessible datasets on law and legal institutions combined with a methodological development that can best be described as dizzying presents legal scholars with a rich toolbox of exciting computational methods at their disposal.[6] The term 'computational methods', as it is used in this contribution, refers to a broad range of data-driven approaches developed in the field of computer science.[7] The development in natural language processing (NLP) methods, in particular the introduction of word embeddings[8] and transformers,[9] deserve special attention as they are, on their face, ideally suited for a text-focused discipline like law.

Will this development change how legal scholarship is done? I will address whether and to what extent legal scholarship[10] can benefit from using computational methods, i.e. what is sometimes referred to as data science in law,[11] law-as-data,[12] or computational legal studies (CLA).[13] The vein of scholarship that I will focus on shares the theoretical and epistemological foundations of many types of empirical legal scholarship,[14] but falls closer to legal informatics with regards to method.[15] My main point is that computational and doctrinal approaches are compatible and that combining them allows for scholarly discoveries beyond the reach of either by themselves.[16] After presenting the reasons for this position, I will discuss the steps needed to move forward.

2. TRACING THE ROOTS OF MISDIRECTED EMPIRICAL SCEPTICISM

It seems that there is considerable scepticism among European legal academics regarding the use of empirical methods in legal research and that this is the source of some tension between scholars that employ empirical methods and those who do not.[17] An example of such scepticism can be found in Hesselink's claim that "[i]f one wants to know what the right answer is to a question of law then empirical research of whatever kind will simply not be helpful".[18] Even more bluntly formulated, Kochenov believes that "while there is law and there is empirical research, doing the latter in order to pretend to say anything about the former... is both methodologically and theoretically dubious, if not nonsensical".[19] While I wholeheartedly believe (i) that it is important to answer the type of research questions that legal scholars have traditionally asked, (ii) that legal scholars by answering such questions fill an important role in society, and (iii) that a thorough understanding of law of the kind that one attains through legal education and training is required in answering those questions,[20] I respectfully disagree

Social in Sociology of Law: An Homage to Reza Banakar (Hart Publishing 2023).

6 Elena Kantorowicz-Reznichenko, 'Computational Methods for Legal Analysis: The Way Forward?' (2021) 14 Erasmus Law Review.

7 Cf. ibid. The terminology in the field is both vast and complicated and, in order to not unnecessarily confuse the reader, I will try to keep it as simple as possible. The type of methods discussed in this contribution includes what is sometimes refereed to as data science and artificial intelligence (AI) methods. This broad category includes, among other things, machine learning (ML) – which *inter alia* includes deep learning – and natural language processing (NLP) – which in turn includes, among other things, large language models (LLMs). It also includes "non-AI" methods, including some of the methods used in quantitative text analysis (QTA). See Bao Kham Chau and Michael A Livermore, 'Studying Judicial Behavior with Text Analysis' in Lee Epstein and others (eds), *Oxford Handbook of Comparative Judicial Behavior* (Online version, Oxford University Press 2024), including so-called text and data mining (TDM), as well as for example network analysis (NA). It does not however include more traditional frequentist statistical methods (not that there is anything wrong with these methods, I use them myself all the time).

8 Tomas Mikolov and others, 'Efficient Estimation of Word Representations in Vector Space' <http://arxiv.org/abs/1301.3781> accessed 29 September 2023. Word embeddings are representations of words in a continuous vector space, where words with similar meanings have similar vector representations.

9 Ashish Vaswani and others, 'Attention Is All You Need' <http://arxiv.org/abs/1706.03762> accessed 17 March 2023. Transformers are context-aware embeddings, i.e. the embedding of a word depends on the context it is used, and is serves as the basis for many state-of-the-art models like BERT and GPT.

10 As discussed in Section 2, I here chose to define 'legal scholarship' broadly based on the knowledge it tries to produce rather than by the methods it (traditionally) uses. I here deliberately do not use the more established term 'legal research', even though they could be synonymous, in order to avoid confusion with the type of legal research that non-scholar lawyers engage in. Although scholarly and non-scholarly legal research *may* significantly overlap with regard to aim, theory, and method, a crucial point of departure for this essay is that they do not *necessarily* do so. In a Swedish context, it would be natural to use the term 'legal science' (*rättsvetenskap*), but I fear that it might spark connotations to and questions about whether legal scholarship is sufficiently scientific, which is not this contributions' subject and might detract from its actual one. Finally, it is also worth clarifying that I do not even entertain the idea that legal scholarship should only use computational methods, nor herein seek to address the appropriateness of using "AI" in law outside the scientific domain, for example automated decision-making.

11 Jinzhe Tan and others, 'Data Science Applications and Implications in Legal Studies: A Perspective Through Topic Modelling' [2023] Journal of Data Science 57, 2.

12 Michael A Livermore and Daniel N Rockmore (eds), *Law as Data: Computation, Text, and the Future of Legal Analysis* (Santa Fe Institute Press 2019); Bao Kham Chau and Michael A Livermore, 'Computational Legal Studies Comes of Age' (2024) 1 European Journal of Empirical Legal Studies; Jens Frankenreiter and Michael A Livermore, 'Computational Methods in Legal Analysis' (2020) 16 Annual Review of Law and Social Science 39, 4–6.

13 Kantorowicz-Reznichenko (n 6).

14 I here consciously refrain from making distinctions between subfields, such as socio-legal studies and law and economics, qualitative and quantitative ELS etc.

15 Thomas Margoni, 'Computational Legal Methods: Text and Data Mining in Intellectual Property Research' in Irene Calboli and Maria Lillà Montagnani (eds), *Handbook of Intellectual Property Research* (Oxford University Press 2021) 490–493; see also Wahlgren (n 5).

16 Cf. Margoni (n 14) 493.

17 Gestel and Micklitz (n 1) 293–297, 300.

18 Martijn Hesselink, 'A European Legal Method? On European Private Law and Scientific Method' (2009) 15 European Law Journal 20, 28.

19 Dimitry Kochenov, 'Counting Swines at a Satan's Ball: Book Review of Jan Zglinski's Europe's Passive Virtues' <http://dx.doi.org/10.2139/ssrn.4086668>.

20 Cf. Richard A Posner, 'The State of Legal Scholarship Today: A Comment on Schlag' (2008) 97 Georgetown Law Journal 845, 854.

with Hesselink's and Kochenov's blanket rejections of the usefulness of empirical methods when it comes to saying something novel about law.[21] Computational methods not only can but have already improved legal scholarship.

It seems that the under-appreciation of computational methods in law can be traced back to certain incorrect ideas and assumptions. There is a presence in legal academia of a certain understanding of what constitutes legal scholarship and that in my opinion is unduly restrictive and scientifically counterproductive. At the root of much traditionalist rejection of empiricism lies a dichotomous distinction between doctrinal legal scholarship that seeks to answer normative questions about the law from a legal-internal perspective and empirical legal scholars that are interested in answering descriptive questions related to law's external effects and relations.[22] This is reflective of a view that it is possible and important to uphold a distinction between doctrinal legal scholars and other scholars interested in law. For example, it is commonplace in legal literature to distinguish between, on the one hand, doctrinal legal research and doctrinalists and, on the other hand, empirical legal studies, empirical social science, and multidisciplinarians.[23]

I think this dichotomous thinking is based on an incorrect belief that legal scholars are primarily interested in normative doctrinalism and deductive analysis, whereas legal scholars in fact frequently make empirical claims.[24] One could even make the case that much (supposedly doctrinal) legal research employs a type of empirical approach in so far that makes a prognosis about how the law will be applied[25] on the basis of what has been said and done in the past.[26] A common strategy employed by legal scholars that make empirical claims – for example about shifts in the law, in legal reasoning, legal culture, or legal institutions – is to provide a few examples. This can essentially be characterized as small-n empirical studies.[27] I have no wish to debate the appropriate terminology for different methodological approaches. My point is that empiricism is not fundamentally alien in legal scholarship and that we should therefore discuss *when*, not *if*, we should use empirical approaches in legal scholarship.

Another assumption concerns what computational and other empirical approaches are and what they can be used for. There is a risk that these ideas are based on an outdated understanding both of what empirical legal scholars do and of the methodological state of the art. It seems that this view stems from the idea that empirical approaches are exclusively capable of saying something about the context surrounding law (the external perspective), and not about the law as such (the internal perspective). I will not deny that much empirical legal scholarship, possibly even the majority, focuses on questions, factors, and phenomena that can be characterised as external to the law. To the extent that there is a dominant view of what empirical legal scholarship can be used for, it might in this way be based on empirical observations of the type of empirical scholarship that has been conducted.

The inclusion of external factors has been promoted as one of the strengths of empirical approaches. By studying, *inter alia*, how people experience interacting with the legal system, how law affects behavior on individual and group levels, the efficacy of policy implemented through law under various conditions, and the micro- and macroeconomic impact of legal rules and procedures, empirical approaches to law have produced important knowledge that could not have been attained using exclusively a doctrinal approach.[28] Some of these studies can be characterised as interesting in something different than what has traditionally interested legal scholars. For example, if a law-and-economics scholar using an empirical approach argues in favor of a particular regulatory solution based on market efficiency, this can be seen as distinctly different from doctrinal legal scholarship, something "outside the realm of legal analysis".[29]

That some empirical legal scholars *have historically been* interested in legal-external questions or that some empirical approaches are *unsuitable* for answering legal-internal questions is however irrelevant when it comes to determining whether current state-of-the-art computational approaches are *capable* of answering legal-internal questions.[30] While empirical approaches can be used to uncover information about the context in which law is situated, it does not conversely follow that it is only good for this. I now will provide some concrete and illustrative examples to the contrary.

21 In all fairness it should be pointed out that much development has taken place in the decade-and-a-half that has passed since Hesselink made his claim and I do not know to what extent he would stand by it today.

22 See e.g. Hesselink (n 17) 28–39; Gareth Davies, 'The Relationship Between Empirical Legal Studies and Doctrinal Legal Research' (2020) 13 Erasmus Law Review 3, 9.

23 See e.g. Davies (n 20); Sanne Taekema, 'Methodologies of Rule of Law Research: Why Legal Philosophy Needs Empirical and Doctrinal Scholarship' (2021) 40 Law and Philosophy 33; Gestel and Micklitz (n 1).

24 Lee Epstein and Gary King, 'The Rules of Inference' (2002) 69 University of Chicago Law Review 1, 2–4; Gestel and Micklitz (n 1) 302–303.

25 See fn 2 and accompanying text.

26 I make this point knowing that not all would share this description.

27 In social science it is commonplace to distinguish between qualitative and quantitative empirical studies. The differences between these approaches and the basis for the distinction, in terms of methodology, epistemology, what they can say about 'reality' etc., are not settled questions. However, one basic difference lies in the number of observations (n) that they study where quantitative studies can be characterized as large-n studies.

28 Deborah R Hensler and Matthew A Gasperetti, 'The Role of Empirical Legal Studies in Legal Scholarship, Legal Education and Policy Making: A US Perspective' in Rob van Gestel, Hans-W Micklitz and Edward L Rubin (eds), *Rethinking Legal Scholarship: A Transatlantic Dialogue* (Cambridge University Press 2017) 469–474.

29 Hesselink (n 17) 30.

30 Much like legal scholarship should not be defined by its dominant methodology, empirical legal studies and the use of empirical methods in legal scholarship should not on principle be limited to legal-external questions.

3. USEFULNESS OF COMPUTATIONAL METHODS IN LEGAL SCHOLARSHIP: SOME EXAMPLES

3.1 Generative AI to the Rescue?

Computational methods are not new, not even new in the field of law. Researchers in the field of legal informatics, as well as commercial actors in the Legal Information Retrieval Systems (LIRS) market and the so-called LegalTech sector, have been applying computational methods to law for some time.[31] As I will elaborate on, these methods have also been used in legal scholarship for quite some time.

The introduction of GPT-3 in 2020 likely provided many lawyers' first direct experience with the power of applying computational methods to law.[32] ChatGPT and other chatbots that are based on generative, pre-trained large language models (LLMs) have proved capable of accurately answering quite sophisticated questions about the law.[33] It is true that even state-of-the-art LLM-based chatbots specifically fine-tuned on legal data are prone to hallucinations,[34] and are not perfect at conducting statutory reasoning.[35] While LLMs make mistakes, so do LL.M.s, and it is equally clear that they are not incapable of answering legal questions or nonsensical.[36] I would think that the existence of these LLM-based chatbots should help convince sceptics that computational methods can be useful in legal scholarship, even scholarship that seeks to answer legal-internal questions.

Impressive as they are, LLM-based chatbots are not prone to conduct legal scholarship in the sense that they can generate novel insights about the law. This is clearly the case with the current state of the technology, but it appears to be an inherent limitation of how they are trained. By virtue of being limited to the data that they have been trained on, LLMs are capable of generating information based on what has already been concluded, the type of legal answers that one can find in textbooks. Such answers are clearly not worthless, and because of their ability to generate such information LLMs are valuable tools for scholars conducting research, but they cannot as such produce boundary-pushing research. I shall now provide some concrete and illustrative examples of how computational approaches, including the use of LLMs, can be useful in legal scholarship, drawing on my own and others' research.

3.2 A Helicopter Perspective on the Law

Some of the most important contributions that conducting computational and other large-n studies can provide come from the type of questions that they allow us to ask. While these benefits may come across as somewhat "soft", they should not be underestimated. My first experience with using computational methods was born out of a dissatisfaction over the natural, cognitive limitations on the size of the dataset that one can analyze using a purely doctrinal approach. My dissertation had left me with the impression that the Court of Justice of the European Union (CJEU) was inconsistent in how it cited and used its own case law,[37] but to identify the existence and absence of citation patterns on a large scale was impossible using traditional legal methods. However, by using network analysis we were able to study all references in and between all decisions. I have since come to appreciate that just as some important aspects of the law can only be understood through a close reading, others, like the Nazca Lines of Peru, only make sense when viewed from high above.

Doctrinal legal research largely rests on deductive reasoning, that is to say that it departs from predefined prin-

31 Margoni (n 14) 490–493.

32 Not every reader may be aware that the use of pre-trained LLMs in law predates GPT-3. See e.g. Ilias Chalkidis and others, 'LEGAL-BERT: The Muppets Straight Out of Law School', *Findings of EMNLP* (Association for Computational Linguistics 2020).<https://www.aclweb.org/anthology/2020.findings-emnlp.261> accessed 17 March 2023).

33 Daniel Martin Katz and others, 'GPT-4 Passes the Bar Exam' [2023] SSRN Electronic Journal; Jonathan H Choi and others, 'ChatGPT Goes to Law School' [2023] SSRN Electronic Journal <https://www.ssrn.com/abstract=4335905> accessed 17 March 2023; Michael James Bommarito and Daniel Martin Katz, 'GPT Takes the Bar Exam' [2022] SSRN Electronic Journal.

34 Varun Magesh and others, 'Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools'.

35 Andrew Blair-Stanek, Nils Holzenberger and Benjamin Van Durme, 'Can GPT-3 Perform Statutory Reasoning?' <http://arxiv.org/abs/2302.06100> accessed 4 February 2024.

36 Their current level of competence can perhaps be compared to that of an experienced law student or recent law graduate.

37 Johan Lindholm, State Procedure and Union Rights: A Comparison of the European Union and the United States (Iustus Förlag 2007).

ciples, concepts, and rules. In this regard computational approaches enhance our ability to conduct inductive legal research by identifying patterns in very large legal datasets.[38] Of particular interest in this regard are so-called unsupervised approaches, that is to say approaches that allow computers to identify patterns unrestrained from any preconceived notions or theories. By allowing computers to "run free" they can identify patterns in empirical data that challenges lawyers' existing theories about the law. For example, using network analysis, we clustered all of the CJEU's decisions into communities based on how they are connected to each other through citations. Those communities are functionally comparable to areas of law, but because we used an unsupervised approach they could and in some important regards did differ from the areas of EU law one encounters in textbooks.[39] Citations are not the only method that can be used for categorizing legal sources in an unsupervised manner. A commonly used computational approach is to identify topics in large legal text collections and to categorize individual legal text on these topics using topic modelling.[40] This has a number of useful applications. For example, I have used unsupervised topic modelling to split sports arbitration cases into novel categories,[41] whereas Yannis Panagis, Martin Lolle Christensen and Urška Šadl[42] used it to track legal change in European courts.

This is also an illustration of how computational methods can assist in theorizing about law.[43] It is important to point out that machine-identified patterns are not absolute truths and that they should always be subject to human analysis. However, if the facts do not seem to fit the theory, it is good scientific practice to change the theory, and computational approaches can help us test to what extent theories and facts fit together. Also, conducting empirical research can help us sharpen legal theories and concepts. A theory can only be tested and a concept captured or quantified if it is clear. By requiring sharp legal theories and concepts, the use of computational approaches reveal ambiguities and inconsistencies. As Zglinski acutely observes, "[a]nalysing large numbers of decisions forces us to be precise about what it is that we are looking for [and i]t, thereby, indirectly benefits the conceptual work."[44] For example, in one study, we used network analysis to identify which CJEU decisions are the most "important". This required us to clarify the different ways in which a case can be legally important and turn them into measurable variables.[45] Similarly, in his study of judicial deference, Zglinski developed a conceptual framework in order to study its presence in the CJEU over time.[46]

3.3 Generating Research Data

Empirical research is only as good as the data that is based on. While legal scholars are very skilled at collecting, systematizing, and analyzing legal authorities, the methods traditionally used in law are poorly suited for generating accurate and reproducible large legal datasets that capture relevant internal aspects of the law.[47] A fundamental challenge in this regard is that the variables that we want to capture are hidden in complicated, technical, and nuanced language found in documents that are, at best, semi-structured.[48] For example, a legal scholar or a legally-trained research assistant reading CJEU decisions can determine whether the Court deferred to a national court to make the final decision, but to do this on a large scale is prohibitively time consuming.[49] What if we can train computers to be perfect research assistants: perfectly consistent, highly effective, low-cost, and able to work indefinitely without taking breaks?[50]

Significant progress towards this becoming reality has been made in recent years. One example of this that should interest legal scholars is Joana Ribeiro De Faria, Huiyuan Xie and Felix Steffek[51] who successfully employed GPT-4 to extract key legal aspects from case law text, such as claims, references, case outcomes, and

38 Margoni (n 14) 491–492.

39 Atieh Mirshahvalad and others, 'Significant Communities in Large Sparse Networks' (2012) 7 PLoS ONE e33721; Mattias Derlén and others, 'Coherence Out of Chaos: Mapping European Union Law by Running Randomly Through the Maze of CJEU Case Law' (2013) 16 Europarättslig Tidskrift 517; see also Martin Lolle Christensen, Henrik Palmer Olsen and Fabien Tarissan, 'Identification of Case Content with Quantitative Network Analysis: An Example from the ECtHR', vols 29th International Conference on Legal Knowledge and Information Systems (JURIX'16) (2016) <https://hal.science/hal-01386810> accessed 20 August 2024.

40 See e.g. M Mohammadi and others, 'Combining Topic Modelling and Citation Network Analysis to Study Case Law from the European Court on Human Rights on the Right to Respect for Private and Family Life' <http://arxiv.org/abs/2401.16429> accessed 14 August 2024; Tan and others (n 11).

41 Johan Lindholm, 'Court of Arbitration for Sport: En framgångsrik trettiofemåring med begynnande medelålderskris?' (2019) 2019/20 Juridisk Tidskrift 482, 146–159.

42 'On Top of Topics: Leveraging Topic Modeling to Study the Dynamic Case-Law of International Courts' (2016) 294 Frontiers in Artificial Intelligence and Applications 161.

43 Cf. Michael Heise, 'The Importance of Being Empirical' (1999) 26 Pepperdine Law Review 807, 813 ("The development of good theories is made even more difficult without the benefit of good data.").

44 Jan Zglinski, Europe's Passive Virtues: Deference to National Authorities in EU Free Movement Law (Oxford University Press, Oxford 2020), 7.

45 Mattias Derlén and Johan Lindholm, 'Goodbye *van Gend En Loos*, Hello *Bosman*? Using Network Analysis to Measure the Importance of Individual CJEU Judgments' (2014) 20 European Law Journal 667.

46 Ibid. See also Michal Ovádek, Phillip Schroeder and Jan Zglinski, 'Where law meets data: a practical guide to expert coding in legal research' European Law Open (forthcoming); Jan Zglinski, 'What is the Point of Empirical Legal Research in EU Law?' in Empirical Legal Studies in EU Law (Cambridge University Press, Cambridge, forthcoming).

47 Cf. e.g. Mark A Hall and Ronald F Wright, 'Systematic Content Analysis of Judicial Opinions' (2008) 96 California Law Review 63; Frankenreiter and Livermore (n 12) 40.

48 The task of extracting the valuable information is sometimes referred to as text and data mining or TDM. See e.g. Margoni (n 14) 487.

49 The coding task that is ideal for automation is one that is sufficiently clear that humans can reliably do it but it takes a lot of time.

50 Cf. Alessandro Contini and others, 'Recognising Legal Characteristics of the Judgments of the European Court of Justice: Difficult but Not Impossible' [2022] Legal Knowledge and Information Systems.

51 'Automatic Information Extraction from Employment Tribunal Judgements Using Large Language Models' [2024] SSRN Electronic Journal <https://www.ssrn.com/abstract=4776160> accessed 4 June 2024.

reasons for the decision. Ivan Habernal and others[52] were similarly able to use LLMs to 'mine' different types of legal arguments, such as different methods of interpretation, in the case law of the European Court of Human Rights. A third and final example is Jonathan H Choi who developed a computational method for measuring the clarity of legal texts.[53]

Named entity recognition (NER) is a computational task that should be of great interest to legal scholars. NER involves the identification of unique identifiers of 'entities' in text, such as proper nouns, names referring to people or places, but which in principle can be any type of text element.[54] Legal texts are full of entities that are of central relevance when it comes to understanding the text, including legal-internal entities such as sources, actors, rules, principles, and legal concepts. I would think that the ability to reliably and effectively identify such legal entities in very large amounts of legal text makes NER valuable to most legal scholars.[55] Scholars have been developing methods for NER in legal text and successfully applied these to extract a variety of legal contexts across multiple jurisdictions.[56] Being able to annotate references to legal concepts in legal text automatically, reliably, and on a large scale creates a number of opportunities for legal scholars. In addition to the value of information about legal entities as such, using NER-annotated text data can enhance other computational methods in law.[57]

Another example of how computational methods can be useful in generating valuable legal research data involves 'issue splitting'. It is not uncommon that judgments address multiple, distinct legal issues and for each such issue contain reasoning and holding. This makes entire judgments a non-ideal unit of observation for the purpose of empirically studying case law.[58] Judgments often contain extraneous information, such as details about the parties, quotes from relevant legislation, and costs, that may not only be irrelevant but that for analytical purposes constitutes "noise" and that ideally should be removed. Paragraphs or sentences on the other hand are too fine of a unit as important contextual information is lost. Schroeder and I therefore propose the concept of legal issues as a 'Goldilocks' layer, a more efficient level of analysis that balances comprehensiveness and specificity. While it is possible to hand-split judgments by legal issues, it is an immensely resource-intensive task. As an alternative, we hand-coded a relatively small set of judgments by the CJEU and used this data to train a neural network to identify where the Court starts and stops discussing a legal issue. We then use this model to quickly, cheaply, and with high accuracy 'issue split' a much larger number of CJEU decisions.[59]

These examples of successful automated coding of legal data – using computers as research assistants – worked well because the coding tasks were relatively easy, that is to say that the concepts of interest were clear and rather simple and that they were expressed in the text in a transparent and reliable fashion. This will not always be the case. In fact, scholars are often most interested in complex and vague concepts that are difficult to reliably code even by hand and after much training. Although advances in machine learning techniques constantly moves the frontier forwards, some tasks will be beyond machines' reach for a long time (and forever unless scholars do the necessary conceptualization and theorization). Humans and machines are good at and should be used for different tasks: whereas machines are ideal research assistants solving many tedious tasks, the hardest problems should be left to humans.[60]

3.4 Predicting Citations

A critical aspect of legal research and practice involves identifying relevant sources, such as case law, that support legal propositions. Given the rapidly expanding volume of legal documents – such as the more than 800 judgments issued by the CJEU each year – this task is becoming increasingly difficult to perform effectively without computer assistance, and eventually possibly impossible.[61] This raises the question: can we predict citations for legal propositions at a paragraph level? To address this, a group of scholars that included myself attempted to mimic CJEU citation patterns by estimating the probability that the CJEU would cite a particular paragraph in support of a legal statement, based on previous citations. Our approach involves training a BERT-

52 Ivan Habernal and others, 'Mining Legal Arguments in Court Decisions' [2023] Artificial Intelligence and Law.

53 'Measuring Clarity in Legal Text' (2024) 91 The University of Chicago Law Review 1.

54 Mónica Marrero and others, 'Named Entity Recognition: Fallacies, Challenges and Opportunities' (2013) 35 Computer Standards & Interfaces 482.

55 Cf. Christopher Dozier and others, 'Named Entity Recognition and Resolution in Legal Text' in David Hutchison and others (eds), *Semantic Processing of Legal Texts*, vol 6036 (Springer Berlin Heidelberg 2010) 1–3 <http://link.springer.com/10.1007/978-3-642-12837-0_2> accessed 26 June 2024.

56 See e.g. Elena Leitner, Georg Rehm and Julián Moreno-Schneider, 'Fine-Grained Named Entity Recognition in Legal Documents' [2019] SEMANTiCS 2019 272; Vitor Oliveira and others, 'Combining Prompt-Based Language Models and Weak Supervision for Labeling Named Entity Recognition on Legal Documents' [2024] Artificial Intelligence and Law; Andreas Östling and others, 'The Cambridge Law Corpus: A Corpus for Legal AI Research' <http://arxiv.org/abs/2309.12269> accessed 22 September 2023; Milagro Teruel and others, 'Legal Text Processing Within the MIREL Project' in Georg Rehm, Victor Rodríguez-Doncel and Julián Moreno-Schneider (eds) (2018); Ilias Chalkidis, Ion Androutsopoulos and Achilleas Michos, 'Extracting Contract Elements', *Proceedings of ICAIL '17* (2017).

57 Irene Benedetto and others, 'Boosting Court Judgment Prediction and Explanation Using Legal Entities' [2024] Artificial Intelligence and Law.

58 E.g. when answering research questions about courts and law or for offering well-informed recommendations.

59 Philipp Schroeder and Johan Lindholm, 'From One to Many: Identifying Issues in CJEU Jurisprudence' (2023) 11 Journal of Law and Courts 163.

60 See also, for a similar argument in math, interview with Terrence Tao in Matteo Wong, "We're Entering Uncharted Territory for Math", The Atlantic, 4 October 2024.

61 Cf. Benjamin Alarie, 'The Path of the Law: Towards Legal Singularity' (2016) 66 University of Toronto Law Journal 443; Simon Deakin and Christopher Markou, 'From Rule of Law to Legal Singularity' in Simon Deakin and Christopher Markou (eds), *Is Law Computable?: Critical Perspectives on Law and Artificial Intelligence* (Hart Publishing 2020).

based encoder model on both positive (cited) and negative (not cited) text data to predict citation links between paragraphs. Our model, when tested, on average ranks the actually cited paragraph as number 2. This method enables us to predict references, identify surprising references, and model relationships between legal statements and their supporting sources. Ultimately, this approach provides valuable tools for offering recommendations, as well as for detecting, studying, and explaining unexpected judicial reasoning.[62]

An accurate citation prediction model has multiple potential uses in legal scholarship. One of these is to measure whether a decision is 'good law', that is to say whether it is a good authority for a legal proposition, or whether it for example has been overruled or become obsolete. To do so has important legal implications and tangible practical uses. For example, while explicit overruling is rare on the CJEU, the Court frequently implicitly or 'covertly' overrules its own case law.[63] Consequently, it can be difficult to know whether CJEU case law is good law. Hand coding whether a case is good law is possible,[64] but to do so is prohibitively expensive. This begs the question, can 'case law health' be measured computationally? We have previously experimented with network analysis, more specifically various network centrality measurements, to capture whether a CJEU decision has been subsequently overruled.[65] Access to a model capable of accurately predicting the probability that a judgment would be cited in support in a particular textual context and being able to measure the difference between a case's predicted and observed citation rate provides an exciting new avenue in this field.

4. WHAT DO WE NEED GOING FORWARD?

I hope that I have convinced the reader that it is both appropriate and useful to employ computational methods in legal research, and that Chau and Livermore are correct in that computational methods, "[u]sed in conjunction with traditional legal research methodologies,... promise to open new avenues of research that could revolutionize the study of law."[66] It seems to me that state-of-the-art computational approaches are ideally suited for law that rule-based approaches, due to law's indeterminate features, fails to capture accurately and fully.

My position on this matter is supported by an arguably liberal understanding of legal scholarsip. The core mission of social sciences and scientists is to produce novel insights about social phenomena. In the specific case of

62 Henrik Palmer Olsen and others, 'Re-Framing Case Law Citation Prediction from a Paragraph Perspective' in Giovanni Sileno, Jerry Spanakis and Gijs Van Dijck (eds), *Legal Knowledge and Information Systems* (IOS Press 2023).

63 Daniel Sarmiento, 'The 'Overruling Technique' at the Court of Justice of the European Union' [2023] European Journal of Legal Studies 109; Jan Komárek, 'Judicial Lawmaking and Precedent in Supreme Courts' [2011] SSRN Electronic Journal 32–33 <http://www.ssrn.com/abstract=1793219> accessed 6 May 2022.

64 Some American actors in the LIRS space provide this service.

65 Mattias Derlén and Johan Lindholm, 'Is It Good Law? Network Analysis and the CJEU's Internal Market Jurisprudence' (2017) 20 Journal of International Economic Law 257.

66 Chau and Livermore (n 12) 10.

legal scholarship and scholars, that social phenomenon is the law. Whereas producing new knowledge about the law requires the scientific community to employ certain methods,[67] it does not, on principle or in practice, exclude other methods. Every method is obviously not a good fit for answering every research question, but methodological conservatism also has no value *per se*. Moreover, I hope that I, through my examples, have been able to convince some readers that computational methods can and have helped produce novel insights into law, even from a legal-internal perspective.

Frankenreiter and Livermore write that "[a]s these tools continue to advance, and law scholars become more familiar with their potential applications, the impact of computational methods is likely to continue to grow."[68] While I hope that this will be true, I am somewhat sceptical about the ease of the transition. In order to advance the use of computational methods in legal research we must, first, improve access to the infrastructure on which it is based. Legal data access has improved significantly in recent years, but access to open, reliable, and comprehensive legal datasets still constitutes a bottleneck. Such data needs to include not only statutes and court precedent, but also, *inter alia*, preparatory works, other documents from the legislative process, legal literature, decisions by lower courts and administrative agencies, and party court filings. Two major obstacles to the development of such datasets is that not all legal sources are collected in a freely and publicly accessible archive and that access to important information about law and legal institutions are blocked by commercial actors that hold intellectual property rights. It is however not sufficient to make text data available, one must also ensure that it is accurate and of high quality. This means ensuring that legal texts are curated, clean, correct, accurate, and organized. Additionally, it is essential to enhance it with rich and accurate metadata, including the use of unique and stable identifiers, assigning legally relevant labels to text elements, and tagging of natural and legal entities. Achieving this requires the collaboration of multiple actors including libraries, parliament, government, courts, government agencies, and commercial actors. On the academic side, the creation of shared and open legal datasets will require pooling the skills and efforts of legal scholars, computer scientists, and other academics.

Law schools and legal scholars also have an important role to play when it comes to capacity building. Currently, most European law students graduate without serious exposure to empirical methods or research design, creating a "closed loop" from which professors and doctoral students are drawn, perpetuating stagnant methodological capacities. To break this loop, it is essential to introduce doctoral students to empirical legal thinking and computational methods, thus fostering a new generation of scholars equipped with the tools necessary for modern legal research. Not every future legal scholar will or should learn to master state-of-the-art computational methods, but if we can provide them with a basic understanding of the tools, their possibilities, and their limitations, we can facilitate fruitful collaboration between legal scholars and computer scientists.[69]

Johan Lindholm

Professor of Law, Umeå University, Department of Law. email: johan.lindholm@umu.se.
This essay is based on talks that he gave at the Conference on Text and Data Mining and Artificial Intelligence, 14 June 2024, at Stockholm University, and at a seminar at the Department of Law at Uppsala University. The author is grateful for comments and suggestions provided by the participants, as well as by Isak Nilsson and Jan Zglinski.

67 Most obviously to engage in traditional descriptive and normative jurisprudence.

68 Frankenreiter and Livermore (n 12) 39.

69 Kantorowicz-Reznichenko (n 6) 5; Heise (n 4) 828–829.

Researching Legal AI: The Cambridge Law Corpus and Predicting Decisions of the UK Employment Tribunal

Holli Sargeant and Felix Steffek

ABSTRACT

This contribution introduces the Cambridge Law Corpus (CLC) and a research project benchmarking the prediction of UK Employment Tribunal decisions, which is based on the CLC data. The CLC is a dataset containing more than 320,000 UK court decisions. This article explains the need for legal datasets, the creation of the CLC and the ethical considerations concerning the dataset's construction and distribution. Subsequently, an experiment engaging with legal judgment prediction using the dataset is reported. The decisions predicted are those of the UK Employment Tribunal, which is the first instance for conflicts between employees and their employers. The experiment compares baselines of different AI models and human experts predicting whether the employee will win, partly win, lose or whether the Tribunal will render another decision.

1. THE CAMBRIDGE LAW CORPUS: A DATASET FOR LEGAL AI RESEARCH

The Cambridge Law Corpus (**CLC**) represents a groundbreaking advancement for legal AI research in the UK. We present the first and only large-scale dataset of machine readable UK court cases for computational research. This dataset of over 320,000 UK court cases spans from the 16th century to the present, with most cases originating in the late 20th and early 21st centuries. The CLC establishes the research infrastructure required to advance legal AI research traditionally hindered by access to large-scale, structured legal data. It has been created by an interdisciplinary team, consisting of Andreas Östling, Holli Sargeant, Huiyuan Xie, Ludwig Bull, Alexander Terenin, Leif Jonsson, Måns Magnusson and Felix Steffek. The paper introducing the CLC has been published by Advances in Neural Information Processing Systems 36 (NeurIPS 2023): Datasets and Benchmarks Track.[1]

Recent advancements in AI and natural language processing (NLP) have been remarkable, especially with the development of transformer-based models like BERT and large language models such as GPT. These models have achieved or even surpassed human performance in various language tasks. While their application to the legal domain is a rapidly developing area, it is limited by the scarcity of specialised legal datasets. One of the primary strategies for enhancing the capabilities of legal AI involves pre-training language models. Therefore, legal AI development hinges substantially on the availability and quality of legal data, which is distinct from general corpora. First, case law contains complex, nuanced, and domain-specific language. Second, it is jurisdiction-specific, making it challenging to develop models that are specific to different legal systems. Third, the inherent lack of metadata or structure in UK case law further complicates the application of AI, which thrives on large, well-structured data.

The CLC aims to bridge this gap by providing a rich, structured dataset tailored for legal AI research. It currently contains case law from 53 courts and tribunals across the UK, particularly focusing on England and Wales. It is continuously updated, for example, judgments from Scotland and Northern Ireland will be added in due course. The dataset is organised by court and year, where each case is stored as a single XML file containing the legal text and certain metadata including an assigned unique identifier (CLC-ID) and neutral citation. Additionally, we include a small set of expert annotations for case outcomes to assist advanced research tasks like outcome prediction and extraction. Using our annotated data, we have trained and evaluated case outcome extraction with GPT-3.5, GPT-4 and RoBERTa models to provide benchmarks for future research.

1 Andreas Östling, Holli Sargeant, Huiyuan Xie, Ludwig Bull, Alexander Terenin, Leif Jonsson, Måns Magnusson and Felix Steffek, *The Cambridge Law Corpus: A Dataset for Legal AI Research,* Advances in Neural Information Processing Systems 36 (NeurIPS 2023): Datasets and Benchmarks Track, available at <https://papers.nips.cc/paper_files/paper/2023/hash/819b8452be7d6af1351d4c4f9cbdbd9b-Abstract-Datasets_and_Benchmarks.html>.

The CLC can be used for diverse research tasks and applications; we consider two in our paper. Case outcome extraction, for example, allows models to locate judgment outcomes within lengthy documents, a challenging task well-suited to automation. In early experiments, transformer-based models and large language models show differing levels of accuracy in identifying outcome-related information. Another example for computational analysis about case law includes topic modelling. This research enables analysis of long-term trends in legal areas, such as contract disputes and employment law, shedding light on the evolving factors influencing UK court decisions and access to the legal system. The CLC also opens up a multitude of research opportunities in the field of legal AI and broader computational analysis of law. By providing a comprehensive and structured dataset, the CLC provides the research infrastructure to explore such opportunities.

The legality and ethics of collecting, processing and releasing the corpus is of paramount importance. We have undertaken considerable analysis of the relevant considerations for lawful and ethical design of this project. One core concern with the release of large legal datasets is the personal information they contain. To uphold principles of open justice, UK court cases are generally not anonymised. However, where necessary for the proper administration of justice or to protect certain parties—such as children, victims of sexual offences or asylum seekers—the court will anonymise identities. Privacy regulations, specifically the Data Protection Act 2018 and UK implementation of the European Union's General Data Protection Regulation, detail how personal data can be handled. We have prioritised the use of this corpus in a way that is in the public interest and does not pose risks to individuals' rights, freedoms or interests. By balancing the public availability of all cases in the dataset in other repositories and the principle of open justice, with our prohibition of research identifying individuals, the requirement of ethical clearance and our mechanisms for the erasure of data, we believe these are appropriate safeguards to avoid harm to any individuals.

Against this background, the CLC is not open access. Only researchers can gain access through a straightforward application form.[2] We ask that university-affiliated researchers provide a research plan, university ethical approval and agree to the Terms and Conditions. These requirements help ensure the corpus is used responsibly, aligning with UK laws and ethical research standards.

The CLC has established critical infrastructure for legal AI research in the UK. We are committed to the continuous improvement of the CLC. Future updates will include additional cases, enhanced annotations, and new features based on user feedback and emerging research needs. As more researchers engage with this corpus, the opportunities for impactful insights and transformative advancements in legal AI will continue to expand, reshaping the future of legal research and accessibility.

The work on the CLC is part of the UK Economic and Social Research Council (ESRC) and JST (Japan Science and Technology Agency) funded project on Legal Systems and Artificial Intelligence. The support of the ESRC and JST is gratefully acknowledged.

2. BENCHMARKING CASE OUTCOME PREDICTION FOR THE UK EMPLOYMENT TRIBUNAL: THE CLC-UKET DATASET

Employment tribunals play a critical role in resolving disputes between employers and employees, yet the volume and complexity of cases create challenges for timely and consistent resolution. Predicting case outcomes through advanced AI can enhance access to justice, streamline legal processes and help stakeholders make better-informed decisions. In a recent paper published by the Association for Computational Linguistics in the Proceedings of the Natural Legal Language Processing Workshop 2024, Huiyuan Xie, Felix Steffek, Joana Ribeiro de Faria, Christine Carter and Jonathan Rutherford explore the intersection of technological innovation and access to justice, focusing on the development of benchmarks for predicting case outcomes within the UK Employment Tribunal (**UKET**).[3]

Despite the potential benefits of predictive models in legal contexts, there remains a notable gap in available legal data that hampers AI advancements. Publicly accessible, comprehensive datasets are rare, particularly those that offer standardised annotations of legal decisions. Addressing this gap, the CLC-UKET dataset created as part of this project offers a solution by providing an extensive, curated collection of UKET cases, annotated and organised to enhance predictability and transparency within employment dispute resolution.

The CLC-UKET dataset was curated from the Cambridge Law Corpus,[4] compiling approximately 19,000 UKET cases. The dataset includes intricate legal annotations across multiple facets, making it a comprehensive resource for legal AI applications. Manual annotation by legal experts is a time-consuming and costly process. To alleviate this burden, we explored the use of large language models (**LLMs**) to automate the annotation process. By utilising LLMs, specifically the GPT-4-turbo model, we efficiently handled vast quantities of data without compromising on the accuracy or depth of information. Through an iterative approach to prompt design, we

2 The application form and associated information are available at <https://www.cst.cam.ac.uk/research/srg/projects/law>.

3 Huiyuan Xie, Felix Steffek, Joana De Faria, Christine Carter and Jonathan Rutherford, *The CLC-UKET Dataset: Benchmarking Case Outcome Prediction for the UK Employment Tribunal*, Proceedings of the Natural Legal Language Processing Workshop 2024, pp. 81–96, available at <https://aclanthology.org/2024.nllp-1.7/>.

4 Andreas Östling, Holli Sargeant, Huiyuan Xie, Ludwig Bull, Alexander Terenin, Leif Jonsson, Måns Magnusson and Felix Steffek, *The Cambridge Law Corpus: A Dataset for Legal AI Research*, Advances in Neural Information Processing Systems 36 (NeurIPS 2023): Datasets and Benchmarks Track, available at <https://papers.nips.cc/paper_files/paper/2023/hash/819b8452be7d6af1351d4c4f9cbdbd9b-Abstract-Datasets_and_Benchmarks.html>.

optimized the LLM's performance for annotating the following details: (1) facts, (2) claims, (3) references to legal statutes, (4) references to precedents, (5) general case outcomes, (6) general case outcomes labelled as "claimant wins", "claimant loses", "claimant partly wins", and "other", (7) detailed orders and remedies and (8) reasons. We report on this process in more detail in another paper available on SSRN and arXiv.[5]

The annotated CLC-UKET dataset allows for case outcome prediction, a challenging but valuable task in legal AI. Acknowledging discussion on task terminology,[6] we use the term "prediction" rather than "classification" because we specifically focus on predicting case outcomes using only facts and claims, without including explicit outcome information in the input data. In this prediction task, given a set of case facts and claims, the model generates an outcome label that falls into one of four categories: "claimant wins", "claimant loses", "claimant partly wins" or "other". This task relies solely on the description of facts and claims, intentionally excluding any explicit details about the tribunal's final decision to test the model's predictive capabilities based on input case summaries alone. To establish a baseline for model performance, human predictions were collected by providing experts access to the same facts and claims without the actual case outcomes. Comparing human predictions to model outputs is crucial for understanding the limitations and strengths of AI in this domain.

Four types of approaches were used to benchmark the dataset's predictive potential. Each type offers a unique approach, and their comparative performances shed light on the effectiveness of model customisation for complex legal tasks.

1. Performance of Finetuned Transformer Models

- **Highest F-Scores Overall**: Among all models, **finetuned transformer models**, particularly T5, achieved the best results, showing superior accuracy in predicting outcomes. The T5 model displayed the highest F-scores across most categories, highlighting the advantage of training models specifically on the CLC-UKET dataset.
- **Precision and Recall Strengths**: The T5 model achieved strong precision and recall scores across the categories of "claimant wins" and "claimant loses." For instance, T5 attained an F-score of **0.650 for "claimant wins"** and **0.734 for "claimant loses"**. This accuracy underscores how model fine-tuning on specific legal annotations can enhance precision in interpreting complex tribunal judgments.
- **Gaps in Specific Categories**: Despite its overall performance, the T5 model struggled with the categories "claimant partly wins" and "other", where it achieved low F-scores. The "other" category in particular yielded an F-score of zero, suggesting that even advanced models face challenges with underrepresented or very complex outcomes. This outcome indicates that finer distinctions in nuanced cases may require additional tailored training or refined annotation strategies.

2. Comparative Analysis of GPT-3.5 and GPT-4 Models

- **Small but Notable Improvements with GPT-4**: Between the two GPT-based models, **GPT-4 consistently outperformed GPT-3.5**, although the margin was relatively small. This improvement highlights the incremental advancements in newer LLM versions and how refined language models contribute to higher accuracy in complex legal tasks.
- **Impact of Few-Shot Examples on GPT-3.5's Accuracy**: Interestingly, incorporating **task-specific few-shot examples** significantly enhanced GPT-3.5's performance. For instance, using few-shot examples that matched the legal area of the target case improved its F-score in outcome prediction more effectively than randomly sampled examples. This result emphasises the importance of contextual relevance when leveraging few-shot learning, especially in specialised fields like legal AI where case-specific nuances matter.
- **GPT-4 Zero-Shot Precision**: Notably, **GPT-4 achieved the highest precision in its zero-shot setting** among all baseline models, indicating that it can accurately predict outcomes without task-specific fine-tuning when given the right context. Providing task-related examples in few-shot settings (specifically the **"juris-2"** setting, where two examples from similar legal areas were provided) boosted GPT-4's F-score. However, the relatively modest gains suggest that simply adding more examples does not drastically improve performance, pointing to a need for high-quality, highly relevant few-shot examples.

3. Benchmarking Against Human Expert Predictions

- **Human Predictions Outperform AI**: A critical reference point for the model's efficacy was **human expert predictions**, which outperformed the AI models by an approximately **19% higher F-score** over the best-performing model, T5. This gap highlights

5 Joana Ribeiro de Faria, Huiyuan Xie and Felix Steffek, *Automatic Information Extraction for Employment Tribunal Judgements Using Large Language Models*, available at <https://ssrn.com/abstract=4776160> and <https://arxiv.org/abs/2403.12936>, submitted to journal.

6 Masha Medvedeva and Pauline McBride, *Legal Judgment Prediction: If You Are Going to Do It, Do It Right*, Proceedings of the Natural Legal Language Processing Workshop 2023, pp. 73–84, available at <https://aclanthology.org/2023.nllp-1.9/>.

the value of human expertise in interpreting legal nuances that current AI models struggle to replicate.

- **Strength in Judgment-based Decisions**: Human expert annotators demonstrated the highest F-scores for both "claimant wins" and "claimant loses" categories, indicating that the subjective analysis of case nuances may require human interpretation that AI has yet to achieve. On the other hand, GPT-4 outperformed the human experts when predicting "claimant partly wins" and "other", i.e., in more complex cases.

4. Benchmarking Hard Cases

- **Predicting Hard Cases: The human experts were asked to identify those cases that they considered as hard to predict. This allowed comparing the models' and human performance as regards hard cases. As expected, both AI models and human experts achieved worse scores for hard cases.**
- **Finetuned Transformer Models are Best in Predicting Hard Cases: Interestingly, the finetuned transformer models, in particular T5, outperformed both the GPT-based models and the human experts in predicting hard cases.**

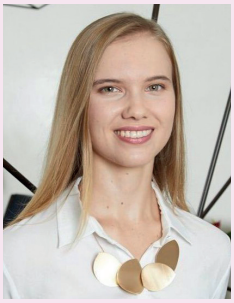
Whilst the study provides valuable insights into the prediction of dispute outcomes for the UK Employment Tribunal, it is important to acknowledge certain limitations of our findings. First, information leakage, one example of bias in legal data,[7] may arise from using LLM summaries of judge written case judgments as we are unable to use neutral information. This information might reflect the judges' post-hoc knowledge and subjective perspectives that shape their written judgments and any information leakage from the LLM summary. Second, while GPT-4 was used for efficient annotation, automated extraction may contain minor inaccuracies, and more detailed factual data could improve predictions. Finally, the dataset spans 2011-2023, during which legal rules and principles evolved, possibly affecting model accuracy over time, as decision dates were indirectly inferred. Future research will address these aspects for more robust prediction models.

The CLC-UKET dataset establishes a meaningful benchmark in legal AI, offering a robust resource for advancing outcome prediction in employment tribunals. Access to the CLC-UKET dataset is available through the Cambridge Law Corpus.[8] While AI models demonstrate promising accuracy, particularly with fine-tuning, human expertise still outshines AI in relevant areas. As we move forward, exploring ways to bridge this gap and improve AI's adaptability will be key to realizing a future where predictive AI and human judgment work seamlessly to enhance access to justice.

This project received funding support from the Cambridge Centre for Data-Driven Discovery and Accelerate Programme for Scientific Discovery, made possible by a donation from Schmidt Futures.

7 Holli Sargeant and Måns Magnusson, *Bias in Legal Data for Generative AI*, 2nd Workshop on Generative AI and Law (GenLaw '24), available at <https://icml.cc/virtual/2024/39169>.

8 At <https://www.cst.cam.ac.uk/research/srg/projects/law>.



Holli Sargeant

Holli Sargeant is a PhD Candidate in the Faculty of Law, University of Cambridge funded by the General Sir John Monash Foundation. Her research examines the consequences of algorithmic decision-making. Her research focuses on the intersection of artificial intelligence and law, exploring two key areas: the necessary adaptations of legal frameworks to address AI-related risks, and the potential applications of AI to the legal system itself. Holli works with various international organisations and not-for-profits to provide legal advice on the use of emerging technology to improve access to justice and uphold human rights.

Prior to commencing her PhD, Holli was an Australian solicitor working in digital law, technology transactions and human rights at Herbert Smith Freehills and the Australian Human Rights Commission. Holli holds a Bachelor of Laws with first class Honours and a Bachelor of International Relations from Bond University. She has previously studied at the National University of Singapore and worked at the Singapore Academy of Law as a New Colombo Plan Scholar.

https://www.law.cam.ac.uk/people/research-students/h-sargeant/79151.

Felix Steffek

Felix Steffek is Professor of Law at the University of Cambridge and Senior Member of Newnham College. He serves as Co-Director of the Centre for Corporate and Commercial Law (3CL) and holds a JM Keynes Fellowship in Financial Economics awarded by the University of Cambridge. He is Global Distinguished Professor of Law at the University of Notre Dame.

His research interests cover corporate finance and insolvency law, artificial intelligence, dispute resolution and commercial law. He has advised international organisations, governments, parliaments and courts in these areas. He represents law on the Academic Publishing Committee of Cambridge University Press.

Felix Steffek is leading multiple research projects on artificial intelligence and law, among them the Nuffield Foundation funded project on 'Access to Justice Through Artificial Intelligence' and the AHRC funded project on 'Explainable and Ethical Legal Artificial Intelligence'.

For further information please see https://www.law.cam.ac.uk/people/academic/f-steffek/6136.

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

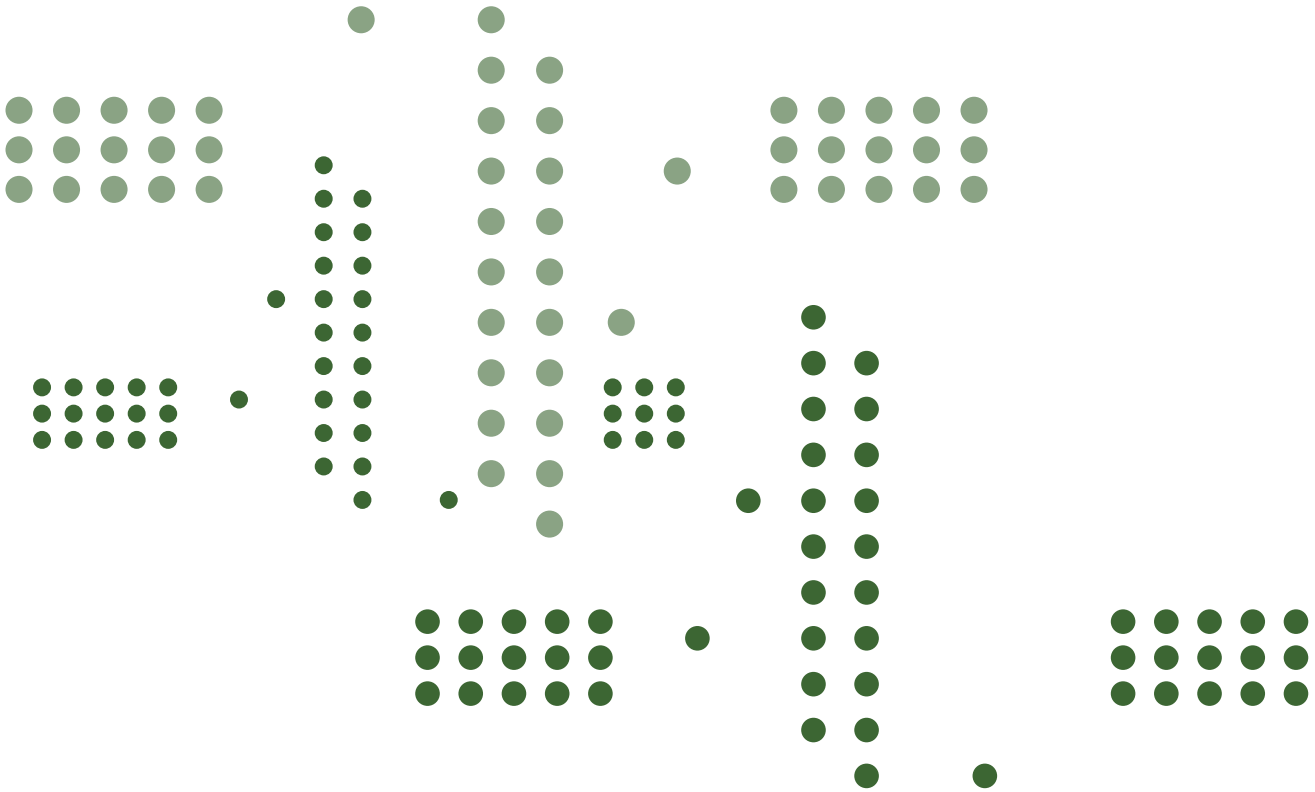
1 Selena Deckelmann, 'Wikipedia's Value in the Age of Generative AI' (*Wikimedia Foundation*, 12 July 2023) <https://wikimediafoundation.org/news/2023/07/12/wikipedias-value-in-the-age-of-generative-ai/>, accessed 17 October 2024.

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 Communi*cation <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, CCo.[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 Res* 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,

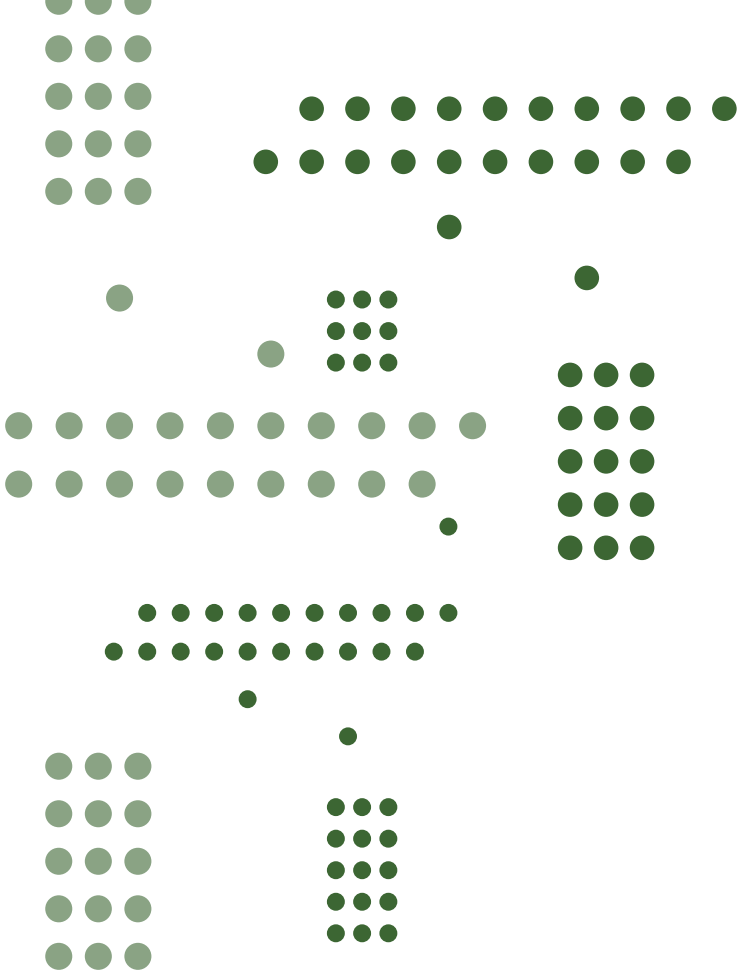
ssrn.com/sol3/papers.cfm?abstract_id=4963711> accessed 17 October 2024.

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 CCo 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 CCo 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).

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 study from 2023 analysed 100 licensing contracts between scientific publishers and data vendors, on the one hand, and public libraries and research institutions, on the other, 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

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, Licenses, Limitations: An Empirical Insight into the Contractual Conditions Regulating Text and Data Mining for Research' (2024) <https://digrep.bg/en/copyright/libraries-licenses-limitations/.

47 Ibid.

means of an appropriate machine-readable rights reservation that the Periodicals may not be used for text and 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

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>.

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 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.

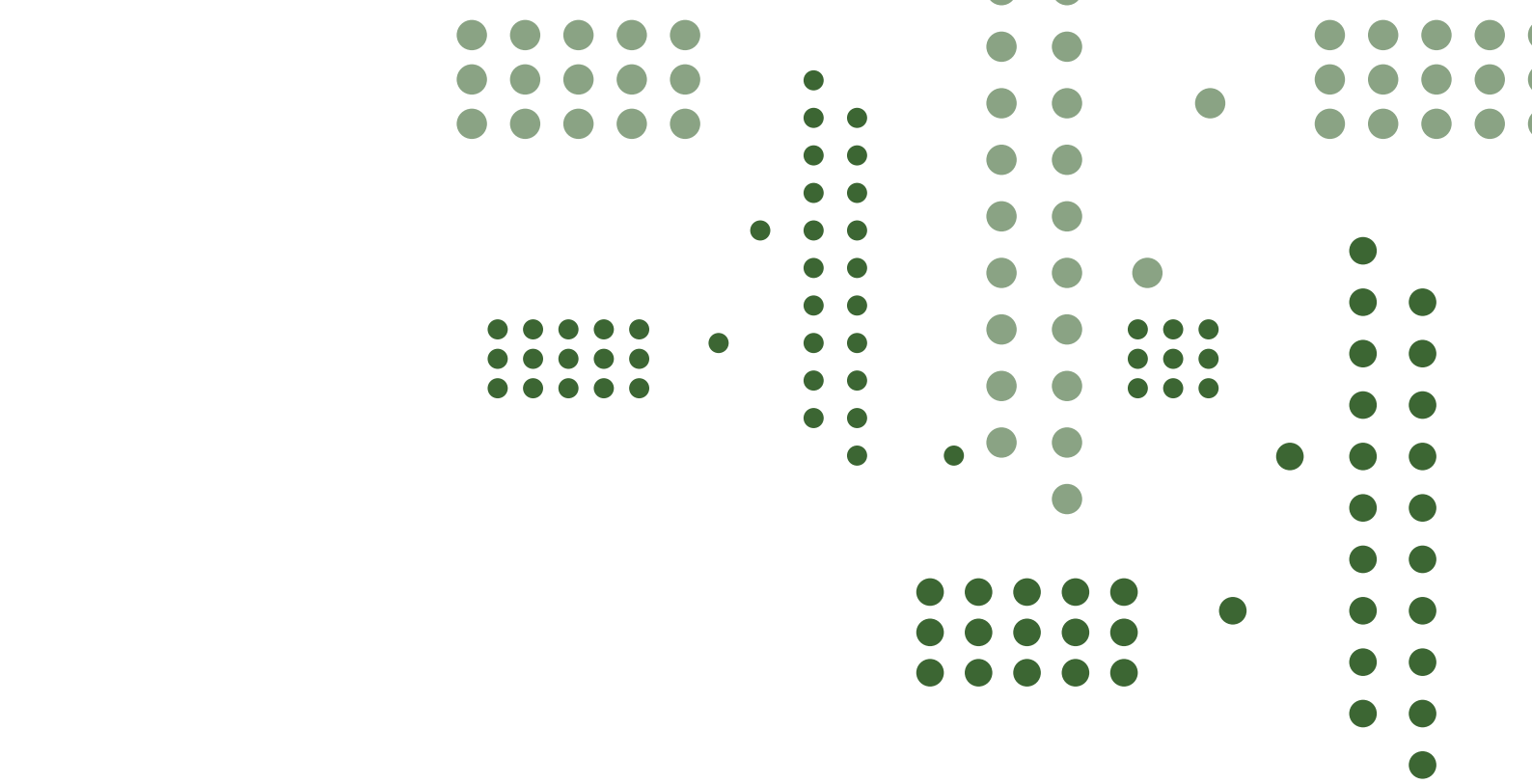
Nonetheless, the challenges posed by AI training on digital cultural heritage, including legal considerations related not only to copyright but also to privacy and other concerns, must be carefully addressed. Knowledge custodians should not be left to navigate these issues alone. The EU has made an initial 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.

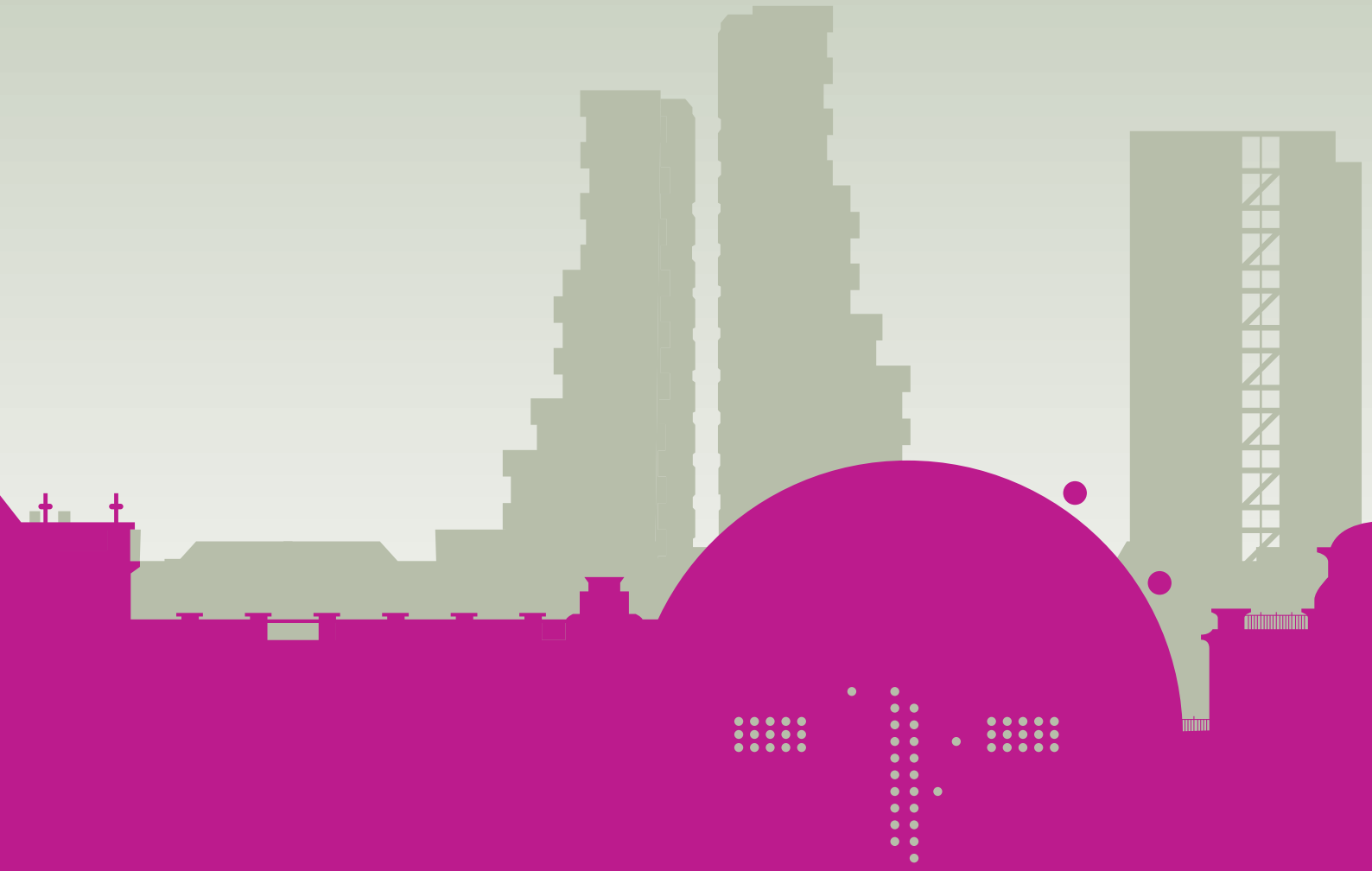
Ana Lazarova

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.





Produced with the support of STIFTELSEN JURIDISK FACKULTETSLITTERATUR and the sponsorship of

Groth & Co

Sandart&Partners

VINGE

AWA

ROSCHIER

CIRIO