

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 Market¹ (hereinafter: “the DSM Directive”) was implemented in Slovenia by the Copyright and Related Rights Act² (hereinafter: “the Copyright Act”) and the Act Regulating Collective Management of Copyright and Related Rights³ 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 bases⁷. 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 conditions¹¹ 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 Act²¹), 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 ZOlPub (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.



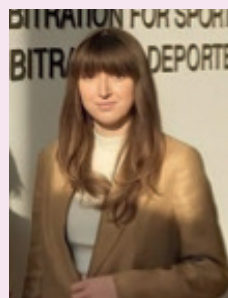
Maja Bogataj Jančič

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.