LEGAL IMPLICATIONS OF DATA MINING FOR JOURNALISTIC PURPOSES

Ana Lazarova, PhD
Sofia University „St. Kliment Ohridski"

Abstract

Access to information, and data in particular, is a necessary condition for carrying out journalistic activity as an important pillar of free and democratic societies. And while freedom of expression and of the press traditionally may come into conflict with the protection of privacy or intellectual property as fundamental rights, the use of new technological forms such as text and data mining for journalistic purposes is creating an entirely new set of legal implications for the exercise of freedom of information. The present study focuses on the issues arising in the fields of copyright, personal data protection, open data, data governance etc. in connection with the use of data and creative materials as an input to machine learning models in the context of journalism.

Key words: Freedom of information; text and data mining; machine learning; copyright; privacy; journalism.

In recent decades, the journalistic profession was fundamentally impacted by technological developments. These new opportunities give rise to the so-called algorithmic journalism[1]. And while automated content production, also referred to as „synthetic media"[2] and „robojournalism"[3] steals the spotlight, new technologies are largely incorporated mostly in the context of journalistic investigations and research. Journalists often find themselves working with datasets too massive for humans to comprehend and data mining is the only viable option in order to uncover connections between variables with high significance. This, in turn, can allow journalists to test complex ideas and hypotheses and discover new social trends[4]. Covering the

---

1 Kotenidis E, Veglis A. (2021) Algorithmic Journalism - Current Applications and Future Perspectives. *Journalism and Media* 2(2), p.244 <https://doi.org/10.3390/journalmedia2020014>.

2 See inter alia Ufarte-Ruiz, M.J., Murcia-Verdú, F.J. and Túñez-López, J.M. (2023) Use of artificial intelligence in synthetic media: first newsrooms without journalists. *Profesional de la informaciyón, 32*(2).

3 For studies dealing with the copyright protectability of outputs generated by, or with the help of, Artificial Intelligence (AI), see Trapova, A. and Mezei, P. (2022) Robojournalism - A Copyright Study on the Use of Artificial Intelligence in the European News Industry. *GRUR International, 71*(7), p. 589.

4 Kotenidis, Veglis (n 1).

world's largest whistleblower case to date - the *Panama papers*[5], would not have been possible without using data mining.

Notwithstanding whether certain data or content are publicly accessible or uploaded online with the initial consent of the concerned party or not, mining can still have legal implications in several aspects at both the EU and the national level.

This study focuses, without claiming to being exhaustive, on some of the normative requirements concerning the processing and use of data and content at the EU level. Thus, the research does not comprise an in-depth analysis of the legislation in other jurisdictions[6] or the potential use of private-ordering mechanisms to restrict mining of content, including for public interest purposes and for investigative journalism.

Text and Data Mining

The research technique of gathering information and extracting patterns from large amounts of digital data using automated software tools is called text mining or data mining, respectively. Commentators define data mining as the extraction of useful information from a larger subset of data and consider it a central part of a broader process called „knowledge discovery"[7]. Data mining is also defined as „the process of using computers and automation to search large sets of data for patterns and trends, turning those findings into business insights and predictions"[8].

In the field of copyright, the technique is referred to as „text and data mining" (TDM), however in practice, there is a difference between data mining, which is the computational process of discovering and extracting knowledge from structured data, and text mining, which is the computational process of discovering and extracting knowledge from unstructured data, usually referring to information created by a human in a natural language, representing unstructured data in a machine-readable format. Textual data can also be created and generated by software programs. The term „text and data mining" was recently granted a formal legal definition in the CDSM Directive

---

[5] For more information regarding how repositories linked to Panama Papers LeaksDB uncovered patterns of relationships, see Zhuhadar, L. and Ciampa, M. (2019) Leveraging learning innovations in cognitive computing with massive data sets: Using the offshore Panama papers leak to discover patterns. *Computers in Human Behavior*, *92*, p. 507.

[6] For example, the legal source comprising the major potential chilling effect for 'algorithmic journalism' in the United States must be the Computer Fraud and Abuse Act (CFAA), which provides for both civil and criminal liability for unauthorized access to networked computers. According to Molly Shaffer Van Houweling, operators of internet platforms have argued, sometimes successfully, that the CFAA prohibits access *even to publicly accessible information* if that access violates a platform's terms of service or continues in the face of a cease-and-desist letter. See Yildirim, E., Van Houweling Shaffer, M., Lazarova, A. and Vézina, B. (2023) *Freedom to Share: How Government's Data Sharing Policies Concerning Publicly Available Data Impact Academic Research and Journalism in the Public Interest.* Creative Commons Medium Blog <https://medium.com/creative-commons-we-like-to-share/freedom-to-share-how-governments-data-sharing-policies-concerning-publicly-available-data-impact-d09cb736aebf>.

[7] Bramer, M. (2007) *Principles of data mining* (Vol. 180, p. 2). London: Springer.

[8] Rutgers Bootcamps (2022) What Is Data Mining? A Beginner's Guide <https://bootcamp.rutgers.edu/blog/what-is-data-mining/>.

of 2019[9]. The meaning is defined in para 2 of Art. 2 of the directive 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". According to Recital 8, it is a technology that enables the processing of large amounts of information with a view to gaining new knowledge and discovering new trends. The recitals further describe the technology as enabling „the automated computational analysis of information in digital form, such as text, sounds, images or data".

The term that the proposal for an amendment of the Bulgarian Copyright and Neighbouring Rights Act (CNRA) uses to denote the technology is „automated text and information analysis". The new concept is defined in a proposed § 2, item 3a of the Additional Provisions of the CNRA, as „any automated analytical method used for the analysis of text and data in digital form, for the creation of patterns, trends, relationships and other information".

Processing of Personal and Non-Personal Data

Personal data[10], regardless of whether it was publicly available or if it was shared by users voluntarily, falls within the scope of the General Data Protection Regulation (GDPR)[11]. This means that whenever mining datasets of personal data, or even mixed datasets that include personal data[12], utilising such data will most likely constitute „processing"[13] and be scrutinised under the GDPR regime, notwithstanding the public availability of the data mined or the initial consent of the data subject to the publication of this data. This will also make the journalist or the respective organisation, i.e., news provider, a data „controller"[14]. In addition, mining can often

---

9 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, OJ L 130.

10 Under art 4 of the GDPR, 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

11 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) OJ L 119, 4.5.2016, p. 1-88.

12 According to art 2, para 2 of the Regulation on free flow of non-personal data, 'In the case of a data set composed of both personal and non-personal data, this Regulation applies to the non-personal data part of the data set. Where personal and non-personal data in a data set are inextricably linked, this Regulation shall not prejudice the application of Regulation (EU) 2016/679.'

13 Under art 4 of the GDPR 'processing' means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.

14 Under art 4 of the GDPR a 'controller' means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law.

involve sensitive personal data, the processing of which is subject to an even stricter regime[15].

The GDPR, however, acknowledges that journalistic expression is a form of protected expression and obliges Member States to reconcile the right to the protection of personal data with the rules governing freedom of expression and information as per Art.11 of the EU Charter of Fundamental Rights[16]. It is worth noting that in order to take account of the importance of the right to freedom of expression in the context of balancing fundamental rights, the GDPR sets a requirement for a *broad interpretation* of notions relating to that freedom, including the notion of *journalism*[17]. Furthermore, according to Art. 85 of the GDPR, the processing of personal data solely for journalistic purposes, including in the audio-visual field and in news archives and press libraries, is subject to derogations or exemptions from certain general provisions under the GDPR. The adoption of concrete legislative measures which lay down the exemptions and derogations necessary for the purpose of balancing fundamental rights - on, *inter alia*, general principles, the rights of the data subject and specific data-processing situations - is within the discretion of Member States.

The Bulgaria law handles the issue in Art. 25h of the Personal Data Protection Law[18]. The provision, introduced in 2019, states that „processing of personal data for journalistic purposes [...] is lawful when it is carried out for the purpose of exercising freedom of expression and the right to information, while respecting privacy." Furthermore, para 3 of Art. 25h introduces a derogation for journalistic uses from the obligations under Art. 6, 9, 10, 30, 34 and chapter five of Regulation (EU) 2016/679, as well as Art. 25c of the Bulgarian law. Also, the data controller or the data processor may in such cases refuse the full or partial exercise of the data subjects' rights under Arts. 12 to 21 of the GDPR. Moreover, para 4 limits the exercise of the powers of the European commission under Art. 58, para 1 of the GDPR in a way that may cause disclosure of information identifying a source. Lastly, according to para 5, when processing personal data for the purposes of creating a photographic or audio-visual work by photographing a person in the course of their public activity or in a public place, Arts. 6, 12 to 21, 30 and 34 of the GDPR do not apply.[19]

It must be taken into account, however, that GDPR exceptions do not always constitute blanket exemptions for the use of personal data, even for public interest

---

[15] See e.g., art 9 of the GDPR on Processing of special categories of personal data.

[16] According to art 11 of the EU Charter, '(1) Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. (2) The freedom and pluralism of the media shall be respected.'

[17] See GDPR, Recital 153, *in fine* - 'In order to take account of the importance of the right to freedom of expression in every democratic society, it is necessary to interpret notions relating to that freedom, such as journalism, broadly'.

[18] See Personal Data Protection Law, amendment published in SG No. 17 of 2019.

[19] This study will not tackle the decision of the Bulgarian Constitutional Court of 2019, striking the provision of para 2 of art 25h of the Bulgarian Personal Data Protection Law as unconstitutional, because it only covers dissemination of personal data and does not directly concern data mining.

purposes. For example, under Art. 14 of the GDPR, data controllers have the obligation to provide information to the data subject. Para 5, p.(b) of Art. 14 sets an exemption to that requirement if „the provision of such information proves impossible or would involve a disproportionate effort, in particular for processing for archiving purposes in the *public interest*, scientific or historical research purposes or statistical purposes, subject to the conditions and safeguards referred to in Art. 89(1) or in so far as the obligation referred to in paragraph 1 of this article is likely to render impossible or seriously impair the achievement of the objectives of that processing". In the popular *Bisnode* case[20] the Polish DPA fined a company for scraping data from publicly available resources, finding that the fulfilment of its obligation to provide information *did not require a disproportionate effort*.

Furthermore, under Art. 35, para 3, a systematic evaluation of personal aspects relating to natural persons, based on automated decision-making, including profiling, is subjected to the requirement of a *prior assessment* by the controller of the impact of the envisaged processing operations on the protection of personal data. Although the provision of para 1 requires taking into account the nature, scope, context and purposes of the processing on a case-by-case basis, it does not exempt certain activities of the requirement solely based on the public interest nature of the activity or mission of the controller. In the *EU DisinfoLab* case[21] the Belgian DPA fined researchers for publishing raw data in a disinformation analysis on the possible political origin of tweets concerning the „Benalla affair" in France, without conducting a prior data protection impact assessment.

Finally, it should be mentioned that the collection and processing of non-personal data at the EU level are expected to also be affected by different legal instruments resulting from the European Strategy for data[22], such as the Regulation on free flow of non-personal data[23], the Open Data directive[24], as well as upcoming legislation such as the Data Governance Act[25], the Data Act[26], the Interoperable Europe Act[27], etc.

---

[20] Polish Personal Data Protection Office (UODO) v. Bisnode, ZSPR.421.3.2018 (2019) <https://uodo.gov.pl/decyzje/ZSPR.421.3.2018>.

[21] Décision quand au fond 13/2022 du 27 janvier 2022 de la Chambre Contentieuse de l'Autorité de protection des données, N° de dossier: DOS-2018-04433 <https://www.autoriteprotectiondonnees.be/publications/decision-quant-au-fond-n-13-2022.pdf>.

[22] <https://digital-strategy.ec.europa.eu/en/policies/strategy-data>.

[23] Regulation (EU) 2018/1807 on a framework for the free flow of non-personal data in the European Union (Regulation on free flow of non-personal data).

[24] Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast), OJ L 172, 26.6.2019.

[25] Proposal for a Regulation on European data governance (Data Governance Act) COM(2020) 767 final, 2020/0340(COD).

[26] Proposal for a Regulation on harmonised rules on fair access to and use of data (Data Act) COM(2022) 68 final, 2022/0047(COD).

[27] See proposal for a Regulation laying down measures for a high level of public sector interoperability across the Union (Interoperable Europe Act), COM(2022) 720 final, 2022/0379 (COD).

Use of Creative Content

Another crucial legal implication that has to be considered in the process of text and/or data mining is, of course, copyright and database protection regulations.

The relationship between mining and copyright is not always self-evident. The common conception is that once a subject has access to certain content and can read it physically, they should also be entitled to „read" it via a computer[28]. In other words, content that has been made publicly available, incl. online, should be also open to mining[29]. This intuitive approach can be legitimately applied or not depending on the level of formalisation of copyright protection across jurisdictions, the scope of the so-called ontological public domain[30] and the availability of a flexible (open) exception like the *fair use* doctrine.

In the U.S., in the majority of cases, text and data mining would not require the rightsholder's sanction. Some commentators argue that on fundamental level, transitory copies made in the process of TDM may not implicate the rightsholder's exclusive rights at all[31]. However, if they did, U.S. case law seem to consistently suggest, that such use would be *fair*[32]. When assessing the fairness of an unauthorised use, American courts tend to prioritise two of the four legal criteria of fair use, namely - for the use to be „transformative", and for it to not directly compete with the rightsholder's legitimate use of their work. The requirement for the transformative nature of the use has been evolving in recent years to include certain cases of direct reproduction. Fair use decisions have established[33] that reproducing copyrighted works as one step in

---

[28] In the EU, prior to the adoption of the CDSM Directive, the slogan of the free TDM initiative was „The right to read is the right to mine".

[29] According to Peter Murray-Rust, representative of the ContentMine initiative, professor of molecular informatics at the University of Cambridge and one of the pioneers of open access, 'The Right to Read is the Right to Mine. Anyone who has lawful access to read the literature with their eyes should be able to do so with a machine. We want to make this right a reality and enable everyone to perform research using humanity's accumulated scientific knowledge.' See Joseph, H. (2015). The Right to Read is the Right to Mine... <https://sparcopen.org/news/2015/the-right-to-read-is-the-right-to-mine/>.

[30] See Dusollier, S. (2016) *Scoping Study on Copyright and Related Rights and the Public Domain*. World Intellectual Property Organisation Publication.

[31] See *inter alia* Carroll, M. (2019) Copyright and the Progress of Science: Why Text and Data Mining Is Lawful'. *UC Davis Law Review* 53: 893.

[32] According to the U.S. Copyright Act (17 U.S.C. § 107), „Notwithstanding the provisions of sections 17 U.S.C. § 106 and 17 U.S.C. § 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include: (i) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; (ii) the nature of the copyrighted work; (iii) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and (iv) the effect of the use upon the potential market for or value of the copyrighted work. The fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors.".

[33] See *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014) and *Authors Guild, Inc. v. Google, Inc.* 721 F.3d 132 (2d Cir. 2015).

the process of „knowledge discovery" through text data mining was transformative[34] and thus have permitted copying that is necessary to use the information embedded in the copied works for „non-expressive purposes" - that is, not to supplant the works themselves but to generate insights *about* the works[35]. The last criterion, namely - „the effect of the use upon the potential market for or value of the copyrighted work", guarantees the absence of the so-called „market substitution" to the detriment of the rightsholder. According to some authors, this market substitution must be „substantial", i.e., unauthorised use should cause „cognizable market harm" to qualify as an infringement[36]. It should be safe to say that, barring the currently controversial cases of use for the purpose of generative AI, data mining, and especially mining for the purpose of investigative journalism should fall within the scope of fair use.

In the context of a highly formalised copyright protection in the European Union, however, certain acts involved in the process of extracting information from data, text, images etc. could formally constitute acts of use within the meaning of copyright. Accordingly, all these acts would formally require the permission of the author or, as the case may be, another rightsholder. Whether or not there is a risk of potential copyright infringement while mining will depend on the particular methods and tools used. In some cases, mining would not involve acts within the rightsholder's domain and therefore would not require the rightsholder's authorization. Thus, the unauthorised use cannot constitute an infringement. This is the case whenever mining uses tools that provide for minimal copying of a few words or the so-called „crawling" and processing pieces of information[37].

In many cases, however, the processing of large datasets for the purpose of extracting patterns and information would involve temporary or permanent reproduction - a type of use of data and content that is generally within the rightsholder's domain, provided, of course, that the content used is eligible for copyright protection. As a matter of principal, such protection is granted over works that are original, in the sense that they are „the author's own intellectual creation"[38] and the expression of their „creative ability in an original manner by making free and creative choices" giving the work a „personal touch"[39]. Copyright protection could be also granted to databases when the selection and arrangement of the database constitutes the author's own intellectual creation.

---

[34] The case law is documented in Sag, M., (2018) The new legal landscape for text mining and machine learning. *J. Copyright Soc'y USA*, 66, p.291.

[35] See Yildirim, Van Houweling Shaffer, Lazarova and Vézina (n 6).

[36] Sun, H. (2021). Creating a Public Interest Principle for the Adjudication of Fair Use and Fair Dealing Cases. *The Cambridge Handbook of Copyright Limitations and Exceptions*, p. 233. Cambridge: Cambridge University Press. doi:10.1017/9781108671101.019.

[37] Communia Association (2020) Guidelines for Implementation of the DSM Directive <https://www.communia-association.org/2019/12/02/guidelines-implementation-dsm-directive/>.

[38] See e.g. Judgment of the Court of Justice of the European Union in Case C-5/08, Infopaq International A/S v Danske Dagblades Forening [2009] ECLI:EU:C:2009:465, and Judgment of the Court of Justice of the European Union in Case C-145/10, Eva-Maria Painer v Standard VerlagsGmbH and Others [2011] ECLI:EU:C:2011:798.

[39] Judgment of the Court of Justice of the European Union in Case C-604/10, Football Dataco Ltd et al. vs. Yahoo UK Ltd [2012] ECLI:EU:C:2012:115.

However, journalists should also be aware of the existence of related (also called neighbouring) rights for which the requirement of originality does not apply. On the contrary, in some cases, copyright and related rights protection can overlap and create layers of IP protection over the same material. This is the case with the press publishers' rights, introduced under Art.15 of the CDSM Directive, where journalistic publications can be both the subject of copyright protection and protection over „press publications" unbound by the concept of originality"[40].

This can be also the case for databases, which can be protected by both copyright and a *sui generis* right - a specific right for their „makers", which is similar to a producer's right and exists independently of the possible copyrighted status of both the database and its content. A database can qualify for the neighbouring-like right's protection whenever a qualitatively and/or quantitatively substantial investment has been made in either the obtaining, the verification or the presentation of the contents of said database[41]. In *Ryanair v. PR Aviation*[42], a case concerning screen scraping, the CJEU analysed the issue in terms of both copyright and database protection requirements, concluding that computer-generated airline schedules did not meet neither copyright's originality threshold nor the substantial investment requirement under the *sui generis* database right. The right protects the investment in the collection of data into the database but not the creation of data as a by-product of another economic activity. In *British Horseracing Board Ltd v William Hill*[43] the CJEU stated that „The expression 'investment in [the] verification [...] of the contents' of a database in art 7(1) of Directive 96/9 must be understood to refer to the resources used, with a view to ensuring the reliability of the information contained in that database, to monitor the accuracy of the materials collected when the database was created and during its operation. The resources used for verification during the stage of creation of materials which are subsequently collected in a database do not fall within that definition." However, commentators observe remaining uncertainty over the distinction between creation and obtaining data in the context of machine generated data[44]. In relation to live information from football matches (goals, times, scorers), the Court of Appeal in the UK found that investments necessary to record such data should be viewed as investments in obtaining the data and therefore the *sui generis* right should apply.[45]

---

[40] Lazarova, A. (2021). Re-use the news: between the EU press publishers' right's addressees and the informatory exceptions' beneficiaries. *Journal of Intellectual Property Law & Practice*, 16(3) 236.

[41] See art 7 of Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, OJ L 77, 27.3.1996, p. 20-28.

[42] Judgment of the Court (Second Chamber) of 15 January 2015 Ryanair Ltd v. PR Aviation BV, case C30/14 [2015] ECLI:EU:C:2015:10.

[43] Judgment of the Court of Justice of the European Union in Case C-203/02, British Horseracing Board Ltd v William Hill [2004] ECLI:EU:C:2004:695.

[44] See Escribano, B. and Fontanals, S. (2022) The Data Act: new EU rules for data sharing <https://www.ey.com/en_es/law/the-data-act-new-eu-rules-for-data-sharing>. The authors point out that, for example, if sensors are set up to measure meteorological data, that data could be said to be collected. But on the other hand, data internally generated by, for example, a machine in a manufacturing plant recording its own performance, could be said to be created. The distinction can, in some circumstances, be a fine one.

[45] Judgment of the Court of Justice of the European Union in Case C-604/10, Football Dataco Ltd et al. vs. Yahoo UK Ltd [2012] ECLI:EU:C:2012:115.

Copyright Exceptions and Limitations

It should be noticed that, not unlike privacy regulations, protection under both copyright and neighbouring rights is not absolute. There are certain cases where unauthorised use of protected content is permissible by law in the public interest - these carve-outs of the rightsholder's monopoly are called permitted or free uses, user rights or copyright exceptions and limitations.

On the EU level, there are several legacy exceptions that can be potentially used for the purposes of text and data mining (TDM) in journalistic investigations. Firstly, mining could fall under Art. 5.1.1. of the InfoSoc Directive[46], when the activity implies temporary acts of reproduction, which are transient or incidental and an integral and essential part of a technological process and whose sole purpose is to enable a transmission in a network between third parties by an intermediary, or a lawful use, and which have no independent economic significance. This possible application of said exception is expressly mentioned in Recital 18 of the CDSM Directive. Secondly, in some cases Art. 5.3.a. of the InfoSoc Directive can allow for mining - when the activity is performed for research and non-commercial purposes. Thirdly, mining can be covered by Art. 5.2.b of the Directive, where it is affected by physical persons for personal use. This exception can possibly be combined with the application of Art. 5.3.n., which allows libraries to make protected subject matter available to individual members of the public for research or private study. Lastly, under Art. 6.2.b of the Databases Directive, users can reproduce temporarily or permanently, translate, adapt, arrange, distribute and communicate, display or perform to the public, where other exceptions to copyright which are traditionally authorized under national law are involved.

In 2019, recognizing the need for a more consistent approach, the EU co-legislators introduced in the CDSM Directive two provisions dedicated to text and data mining specifically.

Art. 3 of the Directive provides a mandatory exception allowing research organisations and cultural heritage institutions to make reproductions and extractions, in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access. The exception under Art. 3 cannot be overridden by contract or by the so-called technical protection measures (TPMs)[47]. Nevertheless, Art. 2 of the CDSM Directive defines research organisations narrowly. Although subject to debate prior to the adoption of the directive, in view of the final wording of the provisions, investigative journalism would most certainly fall outside the scope of beneficiaries to the exception. However, individual journalists could possibly be able to benefit of this opportunity whenever mining would be carried out through the collections of libraries.

Furthermore, Art. 4 of the CDSM Directive introduces an exception concerning both commercial and non-commercial uses by any users. This exception is thus also

---

[46] Directive 2001/29/EC of the European Parliament and of the Council on the harmonisation of certain aspects of copyright and related rights in the information society [2001] OJ L167/01.

[47] On the topic of contractual and technological override of exceptions, see Lazarova, A., (2022). Contractual override of copyright exceptions. *Contemporary Law*, 4/2021.

available to organisations mining for journalistic purposes. Its application, however, can be blocked unilaterally by the rightsholder by way of an express reservation of rights. Under Recital 18 of the CDSM Directive, the rightsholders should reserve the rights to make reproductions and extractions for text and data mining „in an appropriate manner." In the case of content that has been made publicly available online, it should only be considered appropriate to reserve those rights using machine-readable means. Notwithstanding the reservation regime, the application of this exception under Art. 4 is safeguarded against override by TPMs.

It is important to note, that both CDSM exceptions (Arts. 3 and 4) contain a requirement for the beneficiary to have *lawful access* to the respective materials as a prerequisite for the permitted use. This condition can impede mining in cases like the *Panama papers*[48], thus can be considered a hinderance to investigative journalism in and of itself. In addition, the concept of „lawful use"[49] and „lawful source"[50] in the EU *acquis* is a complicated one. It requires, in order for the use under an exception to be lawful, that the subject matter was made available with the consent of the rightsholder. It should be noted that there is no express legal definition for „lawful access" in the CDSM Directive, but according to Recital 14 „lawful access should also cover access to content that is freely available online."

Conclusion

When it comes to the automated processing of data and content, data mining can constitute a regulated activity depending on the sectorial legislation governing the handling of the relevant type of data or content. Firstly, the data used to derive patterns and information from, could be personal data. In such a case, its processing would be subject to privacy concerns. At the EU level, the General Data Protection Regulation (GDPR) would be applied to such uses. Furthermore, at the EU level a trend can be observed for the collection and processing of non-personal data to become more heavily regulated areas. Last but not least, mining creative materials, databases, software, etc., can be subject to copyright and related rights protection. In all these instances investigative journalists and research organisations, although often

---

[48] According to some commentators in the US, TDM research conducted on infringing sources, such as Sci-Hub, is still lawful because the research provides transformative benefits without causing harm to the markets that matter. See Carroll (n 31).

[49] According to Recital 33 of the InfoSoc Directive, „A use should be considered lawful where it is authorised by the rightholder or not restricted by law."

[50] The „lawful source" concept was introduced by the CJEU. See Judgment of the Court (Second Chamber) of 26 April 2017 in the case C-527/15, Stichting Brein (Filmspeler) [2017] EU:C:2017:300, where the Court says that the use of hyperlinks to websites - that are freely accessible to the public - on which copyright-protected works have been made available without the consent of the right holders - is unlawful. See also Judgment of the Court (Fourth Chamber) of 10 April 2014 in the case C435/12, ACI Adam BV v. Stichting de Thuiskopie, Stichting Onderhandelingen Thuiskopie vergoeding [2014] ECLI:EU:C:2014:254. In § 38 the Court says that „national legislation, such as that at issue in the main proceedings, which does not draw a distinction according to whether the source from which a reproduction for private use is made is lawful or unlawful, may infringe certain conditions laid down by Article 5(5) of Directive 2001/29."

beneficiaries to certain exemptions of the heaviest obligations imposed on general actors in the respective sector, will overall need to account for the compliance with the relevant normative system.

Bibliography

- Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases (OJ L 77, 27.3.1996).
- Directive 2001/29/EC of the European Parliament and of the Council on the harmonisation of certain aspects of copyright and related rights in the information society [2001] OJ L167/01.
- Directive 2006/116/EC of the European Parliament and of the Council of 12 December 2006 on the term of protection of copyright and certain related rights (OJ 2006 L 372).
- Directive 2012/28/EU of the European Parliament and of the Council of 25 October 2012 on certain permitted uses of orphan works Text with EEA relevance, OJ L 299, 27.10.2012.
- 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, OJ L-130/92 of 17 May 2019.
- Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast), OJ L 172, 26.6.2019.
- Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) OJ L 119, 4.5.2016, p. 1-88.
- Regulation (EU) 2018/1807 on a framework for the free flow of non-personal data in the European Union (Regulation on free flow of non-personal data).
- Proposal for a Regulation on European data governance (Data Governance Act) COM(2020) 767 final, 2020/0340(COD).
- Proposal for a Regulation on harmonised rules on fair access to and use of data (Data Act) COM(2022) 68 final, 2022/0047(COD).
- Proposal for a Regulation laying down measures for a high level of public sector interoperability across the Union (Interoperable Europe Act), COM(2022) 720 final, 2022/0379 (COD).
- Judgment of the Court of Justice of the European Union in Case C-203/02, British Horseracing Board Ltd v William Hill [2004] ECLI:EU:C:2004:695.
- Judgment of the Court of Justice of the European Union in Case C-5/08, Infopaq International A/S v Danske Dagblades Forening [2009] ECLI:EU:C:2009:465.
- Judgment of the Court of Justice of the European Union in Case C-145/10, Eva-Maria Painer v Standard VerlagsGmbH and Others [2011] ECLI:EU:C:2011:798
- Judgment of the Court of Justice of the European Union in Case C-604/10, Football Dataco Ltd et al. vs. Yahoo UK Ltd [2012] ECLI:EU:C:2012:115.
- Judgment of the Court (Fourth Chamber) of 10 April 2014 in the case C435/12, ACI Adam BV v. Stichting de Thuiskopie, Stichting Onderhandelingen Thuiskopie vergoeding [2014] ECLI:EU:C:2014:254.

- Judgment of the Court (Second Chamber) of 15 January 2015 Ryanair Ltd v. PR Aviation BV, case C30/14 [2015] ECLI:EU:C:2015:10.
- Judgment of the Court (Second Chamber) of 26 April 2017 in the case C-527/15, Stichting Brein (Filmspeler) [2017] EU:C:2017:300
- Bulgarian Personal Data Protection Law, amendment published in SG No. 17 of 2019.
- Authors Guild, Inc. v. Google Inc., No. 13-4829-cv (2d Cir. Oct. 16, 2015).
- Authors Guild, Inc. v. HathiTrust, 755 F.3d 87 (2d Cir. 2014).
- Décision quand au fond 13/2022 du 27 janvier 2022 de la Chambre Contentieuse de l'Autorité de protection des données, N° de dossier: DOS-2018-04433. Disponible sur: https://www.autoriteprotectiondonnees.be/publications/decision-quant-au-fond-n-13-2022.pdf.
- Polish Personal Data Protection Office (UODO) v. Bisnode, ZSPR.421.3.2018 (2019). Available at: https://uodo.gov.pl/decyzje/ZSPR.421.3.2018.
- Bramer, M. (2007) *Principles of data mining* (Vol. 180, p. 2). London: Springer.
- Carroll, M. (2019) Copyright and the Progress of Science: Why Text and Data Mining Is Lawful'. *UC Davis Law Review* 53: 893.
- Communia Association (2020) Guidelines for Implementation of the DSM Directive <https://www.communia-association.org/2019/12/02/guidelines-implementation-dsm-directive/>.
- Dusollier, S. (2016) *Scoping Study on Copyright and Related Rights and the Public Domain.* World Intellectual Property Organisation Publication.
- Escribano, B. and Fontanals, S. (2022) The Data Act: new EU rules for data sharing <https://www.ey.com/en_es/law/the-data-act-new-eu-rules-for-data-sharing>.
- Joseph, H. (2015) The Right to Read is the Right to Mine... <https://sparcopen.org/news/2015/the-right-to-read-is-the-right-to-mine/>.
- Kotenidis E, Veglis A. (2021) Algorithmic Journalism-Current Applications and Future Perspectives. *Journalism and Media* 2(2), p. 244 <https://doi.org/10.3390/journalmedia2020014>.
- Lazarova, A., (2022). Contractual override of copyright exceptions. *Contemporary Law*, 4/2021.
- Lazarova, A. (2021). Re-use the news: between the EU press publishers' right's addressees and the informatory exceptions' beneficiaries. *Journal of Intellectual Property Law & Practice*, 16(3) 236.
- Rutgers Bootcamps (2022) What Is Data Mining? A Beginner's Guide <https://bootcamp.rutgers.edu/blog/what-is-data-mining/>.
- Sag, M., (2018) The new legal landscape for text mining and machine learning. *J. Copyright Soc'y USA*, 66, p.291.
- Sun, H. (2021). Creating a Public Interest Principle for the Adjudication of Fair Use and Fair Dealing Cases. *The Cambridge Handbook of Copyright Limitations and Exceptions*, p. 233. Cambridge: Cambridge University Press. doi:10.1017/9781108671101.019.
- Trapova, A. and Mezei, P. (2022) Robojournalism - A Copyright Study on the Use of Artificial Intelligence in the European News Industry. *GRUR International*, *71*(7), p. 589.
- Ufarte-Ruiz, M.J., Murcia-Verdú, F.J. and Túñez-López, J.M. (2023) Use of artificial intelligence in synthetic media: first newsrooms without journalists. *Profesional de la información*, *32*(2).

- Yildirim, E., Van Houweling Shaffer, M., Lazarova, A. and Vézina, B. (2023) *Freedom to Share: How Government's Data Sharing Policies Concerning Publicly Available Data Impact Academic Research and Journalism in the Public Interest*. Creative Commons Medium Blog <https://medium.com/creative-commons-we-like-to-share/freedom-to-share-how-governments-data-sharing-policies-concerning-publicly-available-data-impact-d09cb736aebf>.
- Zhuhadar, L. and Ciampa, M. (2019) Leveraging learning innovations in cognitive computing with massive data sets: Using the offshore Panama papers leak to discover patterns. *Computers in Human Behavior*, *92*, p. 507.