GRUR International, XX(XX), 2023, 1–11

https://doi.org/10.1093/grurint/ikad098
Article

Maryna Manteghi*

Overcoming Barriers to Text and Data Mining in the Era of ChatGPT: the Proposed Data Act as a Game-Changer

Article 35 of the proposed Data Act intends to free databases comprised of machine-generated data from the constraints of *sui generis* database protection. The provision may provide some remedy for researchers facing challenges under the text and data mining (TDM) exceptions of the Directive on Copyright in the Digital Single Market. However, the wording of Art. 35 may also raise questions and create legal uncertainties for scholars, as scientific research is not the main aim of the proposed legislation. This article argues that, even after the proposed Art. 35 of the Data Act, some databases comprised of machine-generated data may still fall within the scope of the *sui generis* database protection under certain circumstances. The discussion revolves, *inter alia*, around the concept of recorded data in the context of the 'obtaining-creating dichotomy', the use of mixed databases and the notion of derived or inferred data, and the issue of researchers' access to machine-generated data. The findings of this article are intended to offer guidance to researchers using Artificial Intelligence tools, such as ChatGPT, to mine databases on how they may effectively avoid a potential infringement of the *sui generis* database right. The paper hopes to encourage new changes in the EU regulation on TDM that could create a more balanced and research-friendly framework.

I. Introduction

Text and data mining (TDM) is a significant development in research and innovation. This automated analytical technique enables users to obtain new knowledge from huge masses of raw digital data.[1] The outcome of this data analytics usually reveals new patterns, trends and insights which could be beneficial for the public. The research tool works by retrieving new information from pre-existing digital data and recombining it into targeted output.[2] Nowadays, many private and public actors employ TDM to improve their operations and meet higher standards. For instance, pharmaceutical companies use data analytics to reveal drug interactions, online movie platforms such as Netflix employ these tools to analyze ratings and predict users' preferences, and analytical journals like The Wall Street Journal leverage TDM to analyze reports and predict stock market prices.[3] Moreover, TDM has notable implications in the field of medicine. For instance, it was employed to *analyze* a significant number of scientific publications regarding the coronavirus family, helping to quickly identify potential vaccine candidates.[4] Furthermore, TDM lies at the core of Artificial Intelligence (AI) technologies.[5] The tool is employed to develop new types of information-based services and applications. One of the recent examples of AI technologies whose development and functioning depend on TDM is a cutting-edge chat called ChatGPT.

ChatGPT is a large-scale language model designed by OpenAI to generate content based on a huge number of pre-existing texts obtained from the internet within the setting of human-like conversations.[6] This AI-based software application can perform different processing tasks such as translation, summarization and creation of texts.[7] The chat has been trained on a huge amount of text data to promptly generate responses to users' requests.[8] ChatGPT was released on 30 November 2022 and it reached 100 million users within just a few months of its launch.[9] The

* Affiliated Doctoral Researcher, Faculty of Law, University of Turku, Finland.

1 For a general overview of TDM techniques and methods see Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining: Concepts and Techniques* (3rd edn, Elsevier 2012).

2 Directive on Copyright in the Digital Single Market (CDSM), art 2(2) defines TDM as 'any automated analytical technique aimed at analyzing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations.'

3 Sean Flynn and Lokesh Vyas, 'Examples of Text and Data Mining Research Using Copyrighted Materials' (*Kluwer Copyright Blog*, 6 March 2023) <https://copyrightblog.kluweriplaw.com/2023/03/06/examples-of-text-and-data-mining-research-using-copyrighted-materials/> accessed 20 March 2023.

4 Hao Lv and others, 'Application of Artificial Intelligence and Machine Learning for COVID-19 Drug Discovery and Vaccine Design' (2021) 22 Briefings in Bioinformatics 1.

5 Harry Surden, 'Machine Learning and Law: An Overview' in Roland Vogl (ed), *Research Handbook on Big Data Law* (Edward Elgar Publishing 2021) 171.

6 Jianyang Deng and Yijia Lin, 'The Benefits and Challenges of ChatGPT: An Overview' (2022) 2 FCIS 81, 82.

7 Rehan Ahmed Khan and others, 'ChatGPT-Reshaping Medical Education and Clinical Management' (2023) 39 Pak J Med Sci 605, 605.

8 Jürgen Rudolph, Samson Tan and Shannon Tan, 'ChatGPT: Bullshit Spewer or the End of Traditional Assessments in Higher Education?' (2023) 6 Journal of Applied Learning and Teaching 1, 3.

9 Dan Milmo, 'ChatGPT reaches 100 million users two months after launch' *The Guardian* (London, 2 February 2023) <https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app> accessed 31 March 2023.

chat has the potential to revolutionize all areas affected by its applications. Particularly, ChatGPT may have a significant impact on academia by helping to improve research and learning. For instance, this advanced tool could assist researchers in finding relevant literature among a huge number of scientific publications, generating patterns and correlations from such data, translating and understanding the content written in a foreign language, finding answers to specific questions and summarizing research materials.[10]

Although the use of AI technologies, such as ChatGPT, which are based on TDM, may provide lots of benefits for the public, it may also raise legal issues. In order to mine, a computer typically needs to collect and copy large volumes of digital data from various sources such as websites, social media platforms, databases and other digital repositories. Such data could be protected by copyright or database rights. Therefore, if original works or databases are copied without authorization, the act may infringe the rightholders' exclusive right to reproduction, and result in protracted expensive litigation in court. However, an infringement may also occur when TDM is performed on non-original compilations of data. In this case, the *sui generis* database right could be at risk. The *sui generis* protection functions independently of the copyright protection of databases that protects databases that, because of the selection or arrangement of their contents, constitute the author's own intellectual creation.[11] Unlike the latter, the *sui generis* database right, provided under Art. 7 of the Directive on the legal protection of databases (Database Directive),[12] may apply irrespective of whether the database or the contents of that database meet the originality requirement and could be protected by copyright.[13] The provision protects the maker of a database who has made qualitatively or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents of the database.[14]

Therefore, if researchers intend to analyze a large corpus of materials protected under the *sui generis* regime, they must either obtain permission from the rightholders through contractual agreements or rely on applicable exceptions or limitations to avoid infringement. Although the Database Directive provides some exceptions to the *sui generis* protection under Arts. 8(1) and 9(b), the provisions are neither practical nor effective in the case of TDM. Researchers could encounter legal uncertainty and challenges in complying with specific conditions and requirements stipulated under these exceptions. Two specific TDM exceptions to copyright and the *sui generis* right introduced under Arts. 3 and 4 of the Directive on Copyright in the Digital Single Market (the CDSM Directive)[15] could provide a solution for researchers in this situation. The first exception, in Art. 3, benefits research organizations and cultural heritage institutions carrying out TDM for the purpose of scientific research.[16] Other types of users are covered by Art. 4, the second exception, permitting TDM for any purpose.[17] Although these provisions provide more legal certainty compared to the situation that existed before the adoption of the CDSM Directive, the wording of these provisions has attracted heavy criticism. For instance, Geiger and Jütte argue that the exclusion of private actors from the scope of Art. 3 would deprive the EU of a crucial source of innovation and research, especially in the fields of medicine and AI.[18] Moreover, Ducato and Strowel emphasize that it could be challenging for users to prove that their research projects fall within scientific research as scientific characteristics and methods vary.[19] The exception under Art. 4 has also been subject to criticism, especially for including a so-called 'opt-out' mechanism, which allows rightholders to reserve the use of their works.[20]

Against this background, if the CDSM Directive's TDM exceptions prove to be ineffective in addressing the needs of researchers, scientists may have to explore alternative solutions to lawfully conduct TDM on *sui generis*-protected databases. For instance, the situation can be improved under the proposed Regulation on harmonised rules on fair access to and use of data (the proposed Data Act),[21] which in its current form limits the scope of *sui generis* protection to a certain extent. This legislative initiative was introduced on 23 February 2022 as a part of the European data strategy. The proposed Data Act addresses problems related to the extension of the *sui generis* right to machine-generated databases.[22] In particular, Art. 35 of this legislation clarifies that the *sui generis* protection should not apply to databases 'containing data obtained from or generated by the use of a product or a related service'.[23] The provision could provide a legal ground for scientists to lawfully conduct TDM on databases consisting of data automatically generated

10 Brady D Lund and Yi-Shun Wang, 'Chatting about ChatGPT: How May AI and GPT Impact Academia and Libraries?' (2023) 40 Library Hi Tech News 26, 27.

11 Directive 96/9 on the legal protection of databases [1996] OJ L 77/20 (Database Directive), art 3(1).

12 Directive 96/9 on the legal protection of databases [1996] OJ L 77/20.

13 Database Directive, art 7(4). See also Matthias Leistner and Lucie Antoine, 'Attention, Here Comes the EU Data Act! A Critical In-depth Analysis of the Commission's 2022 Proposal' (2022) 13 JIPITEC 339, 342.

14 Database Directive, art 7(1).

15 Directive 2019/790 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 (CDSM Directive).

16 CDSM Directive, art 3.

17 CDSM Directive, art 4.

18 Christophe Geiger and Bernd Justin Jütte, 'Conceptualizing a 'Right to Research' and Its Implications for Copyright Law: An International and European Perspective' (2022) Joint PIJIP/TLS Research Paper Series 7/2022, 1, 13 <https://digitalcommons.wcl.american.edu/cgi/viewcontent.cgi?article=1079&context=research> accessed 27 February 2023.

19 Rossana Ducato and Alain M Strowel, 'Ensuring Text and Data Mining: Remaining Issues with the EU Copyright Exceptions and Possible Ways Out' (2021) 43 E.I.P.R. 322, 333.

20 Andrew Tyner, 'The EU Copyright Directive: 'Fit for the Digital Age or Finishing It?'' (2020) 26 J. INTELL. PROP. L. 275, 281; P Bernt Hugenholtz, 'The New Copyright Directive: Text and Data Mining (Articles 3 and 4)' (*Kluwer Copyright Blog*, 24 July 2019) <http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/> accessed 12 April 2023.

21 Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act), Brussels [2022] COM (2022) 68 final (proposed Data Act).

22 Commission Staff Working Document, Impact Assessment Report Accompanying the document Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act) SWD (2022) 34 final, 6.

23 Data Act, art 35. See for similar provisions Data Governance Act, art 5(7) and the 2019 Open Data Directive, art 1(6).

by Internet of Things (IoT)[24] technologies. Even though this may not remedy the situation, researchers could benefit from more legal certainty regarding the mining of non-original databases.

This article aims to contribute to the ongoing discussions on TDM by assessing the perspectives of Art. 35 of the proposed Data Act to alleviate tension between *sui generis* database holders and researchers. This objective was motivated by the argument that the current regulation of TDM tends to prioritize the rights and interests of rightholders at the expense of TDM users, creating an imbalance that can hinder scientific research and innovation in the EU. Therefore, the structure of this paper is as follows. After the introduction (Part I), the article will explore how *sui generis* database protection could restrict TDM research in the EU. In addition, this section will examine how the exceptions provided under the Database Directive and the CDSM Directive could potentially mitigate these restrictions (Part II). Next, the article will analyze the proposed Data Act within the framework of TDM regulation. The paper will examine how Art. 35 of the proposed Data Act could impact the application of TDM in research (Part III). Finally, the article will summarize general findings and criticisms of the current legal framework on TDM, highlighting potential solutions for researchers intending to analyze *sui generis*-protected databases by using AI tools (Part IV).

II. *Sui generis* database right and TDM research: a critical analysis

1. The impact of *sui generis* database protection on TDM

Researchers may employ TDM to extract new knowledge from existing large databases.[25] For instance, ChatGPT can be used to analyze a corpus of financial data to predict stock prices, a corpus of legal cases to predict the outcomes of future cases, a corpus of legal texts to identify patterns and biases in legal language or a corpus of network traffic data to identify unusual patterns that may indicate a cyber-attack or other security threat.[26] Even if these databases fail to meet the originality requirement for copyright protection, researchers could potentially be held liable for mining them under the *sui generis* database protection.[27] The *sui generis* database right provides protection for the maker of a database, who has made a substantial investment in either the obtaining, verification or presentation of the contents of the database.[28] The investment could be financial, human or technical, and it must be substantial from a quantitative or qualitative perspective.[29] The former refers to quantifiable resources (e.g. time and money) and the latter to resources which are difficult or impossible to measure using numerical values (e.g. intellectual effort or energy).[30] The Court of Justice of the European Union (CJEU) has clarified that a 'substantial investment' should be assessed by taking into account only investments made in the creation of the database as such.[31] Therefore, the resources utilized for the creation of data which constitutes the contents of a database would be excluded from the assessment in terms of Art. 7(1) of the Database Directive.[32] This derives from the purpose of *sui generis* protection, which is to promote the establishment of storage and processing systems for existing data and not the creation of works that can be accumulated and subsequently organized in a database.[33] In other words, the protection of databases under the *sui generis* regime relates to the protection of the database as a whole, rather than protecting individual data within it.[34]

Further, the CJEU has also clarified the terms 'obtaining', 'verification' or 'presentation' of the contents of the database. The first concept addresses the users' intention 'to seek out existing independent materials and collect them in the database …'.[35] The term 'verifying' is understood as '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.'[36] And finally, the term 'presenting' refers to 'the resources used for the purpose of giving the database its function of processing information, that is to say those used for the systematic or methodical arrangement of the materials contained in that database and the organisation of their individual accessibility.'[37]

24 IoT is 'an emerging paradigm that enables the communication between electronic devices and sensors through the internet in order to facilitate our lives'; Sachin Kumar, Prayag Tiwari and Mikhail Zymbler, 'Internet of Things is a Revolutionary Approach for Future Technology Enhancement: A Review' (2019) 6 Journal of Big Data 1, 1.

25 'Database' can be defined as a 'collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means'; Database Directive, art 1(2).

26 See eg, Marny Lopez, 'ChatGPT: Different Uses and Examples' (*Devlane*, February 2023) <https://www.devlane.com/blog/chatgpt-different-uses-and-examples> accessed 20 April 2023.

27 Database Directive, art 7(4). See also Linda Kuschel and Jasmin Dolling, 'Access to Research Data and EU Copyright Law' (2022) 13 JIPITEC 247.

28 Database Directive, art 7(1).

29 See Case C-444/02 *Fixtures Marketing Ltd v Organismos Prognostikon Agnon Podosfairou* ECLI:EU:C:2004:697, para 44; Case C-338/02 *Fixtures Marketing Ltd v Svenska Spel AB* ECLI:EU:C:2004:696, para 28; Case C-46/02 *Fixtures Marketing Ltd v Oy Veikkaus Ab* ECLI:EU:C:2004:694, para 38; Database Directive recitals 7, 39 and 40; Kuschel and Dolling (n 27) 253.

30 See *Fixtures Marketing Ltd v Organismos Prognostikon Agnon Podosfairou* (n 29) para 44; *Fixtures Marketing Ltd v Svenska Spel AB* (n 29) para 28; *Fixtures Marketing Ltd v Oy Veikkaus Ab* (n 29) para 38.

31 Case C-203/02 *The British Horseracing Board Ltd and Others v William Hill Organization Ltd* ECLI:EU:C:2004:695, paras 31-49; *Fixtures Marketing Ltd v Oy Veikkaus Ab* (n 29) paras 28-42; *Fixtures Marketing Ltd v Svenska Spel AB* (n 29) paras 23-37; and *Fixtures Marketing Ltd v Organismos prognostikon agonon podosfairou AE* (n 29) paras 37-53.

32 *The British Horseracing Board Ltd and Others v William Hill Organization Ltd* (n 31) paras 31-49; *Fixtures Marketing Ltd v Oy Veikkaus Ab* (n 29) paras 28-42; *Fixtures Marketing Ltd v Svenska Spel AB* (n 29) paras 23-37 and *Fixtures Marketing Ltd v Organismos prognostikon agonon podosfairou AE* (n 29) paras 37-53.

33 Database Directive, Recital 12; *Fixtures Marketing Ltd v Oy Veikkaus Ab* (n 29) paras 33-34; *Fixtures Marketing Ltd v Svenska Spel AB* (n 29) para 24; *Fixtures Marketing Ltd v Organismos prognostikon agonon podosfairou AE (OPAP)* (n 29) para 40.

34 Kuschel and Dolling (n 27) 255.

35 *Fixtures Marketing Ltd v Svenska Spel AB* (n 29) para 24; *Fixtures Marketing Ltd v Organismos prognostikon agonon podosfairou AE (OPAP)* (n 29) para 40.

36 *Fixtures Marketing Ltd v Oy Veikkaus Ab* (n 29) para 37; *Fixtures Marketing Ltd v Svenska Spel AB* (n 29) para 27; *Fixtures Marketing Ltd v Organismos prognostikon agonon podosfairou AE (OPAP)* (n 29) para 43.

37 *Fixtures Marketing Ltd v Oy Veikkaus Ab* (n 29) para 37; *Fixtures Marketing Ltd v Svenska Spel AB* (n 29) para 27; *Fixtures Marketing Ltd v Organismos prognostikon agonon podosfairou AE (OPAP)* (n 29) para 43.

The Database Directive confers two transferable rights to *sui generis* database rightholders: the right to extraction and the right to re-utilization.[38] The former refers to the transfer of the contents of a database to another medium while the latter relates to the act of making the contents of a database available to the public.[39] The *sui generis* database rightholder is entitled to prohibit 'extraction or re-utilization of the whole or of a substantial part, evaluated qualitatively or quantitatively, of the contents of that database.'[40] The CJEU has clarified the concept of a 'substantial part' by relying on the scale and volume of the investment made in the part of a database that has been extracted or re-utilized, irrespective of whether that part constitutes a substantial portion of the general contents of such database.[41] The assessment must refer to the harm that the acts of extracting or re-utilizing cause to that investment.[42] In addition, the *sui generis* database rightholder can prohibit the use of insubstantial parts of the contents of the database when it is performed in a repeated and systematic manner.[43] Such use is infringing when it conflicts with a normal exploitation of that database, or when it unreasonably prejudices the legitimate interests of the creator of the database.[44] The CJEU clarifies that these conditions refer to unauthorized acts for the purpose of reconstituting, through the cumulative effect of acts of extraction, the entire or a substantial part of the contents of a database, which could seriously prejudice the investment made by the creator of the database.[45]

The concept of 're-utilization' is less relevant to TDM compared to the term 'extraction'. This is because the findings of TDM research, which are made publicly available through research reports, scientific articles, or conference papers, do not usually include any parts of the mined works. The right to extraction involves 'the permanent or temporary transfer of all or a substantial part of the contents of a database to another medium by any means or in any form'.[46] The CJEU has held that the concept of 'extraction' should be broadly interpreted to protect the makers of a database from the unauthorized appropriation of the results of their investment by acts of the reconstitution of that database or a substantial part of it.[47] The purpose of the transfer should be irrelevant when determining whether an act constitutes 'extraction' within the meaning of Art. 7(2)(a) of the Database Directive. Therefore, regardless of whether the act of transfer is for the purpose of creating another database in the new medium or merely analyzing the contents of the original database, as in the case of TDM, both actions may constitute the act of extraction.[48]

However, if TDM is deemed to infringe the *sui generis* right to extraction, it should be regarded as an unauthorized appropriation of mere facts and data from a database.[49] As Ducato and Strowel reasonably claim, researchers 'consult' databases for informational purposes only, and any reconstitution of a database which may occur during the copying stage of the TDM process is necessary solely for accessing incorporated data and extracting new insights.[50] The CJEU has already clarified that the *sui generis* right to extraction should not apply to the right to consult a database, which is a fundamental right that must be weighed against the interests of the database owner.[51] In this sense, Recital 46 of the Database Directive also clarifies that the right to prevent unauthorized extraction should not lead to the creation of a property right on the information itself that could limit access to data included in the database.[52] Against this background, TDM should be viewed only as an act of 'consultation' rather than the act of 'extraction' of the contents of *sui generis*-protected databases. However, as long as a research process involves copying, the risk of infringement exists. Therefore, it could reasonably be argued that the *sui generis* protection restricts TDM by intensifying the control over access to data.

Therefore, if researchers intend to *analyze* a large corpus of materials protected under the *sui generis* regime, they must either obtain permission from the rightholders through contractual agreements or rely on applicable exceptions or limitations to avoid infringement. The Database Directive provides a few exceptions to the *sui generis* right which could potentially apply to TDM. The first possible candidate is the exception for a 'lawful user' provided under Art. 8(1) of the Directive. The provision permits extraction or re-utilization of insubstantial parts of the contents of a database, evaluated qualitatively or quantitatively, for any purpose.[53] However, this

38 Database Directive, art 7(2)(a), (b).

39 Database Directive, art 7(2)(a), (b).

40 Database Directive, art 7(1).

41 *The British Horseracing Board Ltd and Others v William Hill Organization Ltd* (n 31) paras 68-71.

42 Database Directive, recital 42. See also *The British Horseracing Board Ltd and Others v William Hill Organization Ltd*. (n 31) para 69.

43 Database Directive, art 7(5). See also *The British Horseracing Board Ltd and Others v William Hill Organization Ltd* (n 31) paras 83-95 and Case C-304/07 *Directmedia Publishing GmbH v Albert-Ludwigs-Universität Freiburg* ECLI:EU:C:2008:552, paras 43-44. Case C-762/19 *SIA 'CV-Online Latvia' v SIA 'Melons'* ECLI:EU:C:2021:434, para 47.

44 Database Directive, art 7(5). See also *The British Horseracing Board Ltd and Others v William Hill Organization Ltd* (n 31) paras 83-95; *Directmedia Publishing GmbH v Albert-Ludwigs-Universität Freiburg* (n 43) paras 43-44 and *SIA 'CV-Online Latvia' v SIA 'Melons'* (n 43) para 47.

45 *The British Horseracing Board Ltd and Others v William Hill Organization Ltd* (n 31) para 89.

46 Database Directive, art 7(2)(a).

47 *Directmedia Publishing GmbH v Albert-Ludwigs-Universität Freiburg* (n 43) paras 31-33 relying on *Fixtures Marketing Ltd v Oy Veikkaus Ab* (n 29) para 35; *The British Horseracing Board Ltd and Others v William Hill Organization Ltd* (n 31) paras 32, 45, 46 and 51; *Fixtures Marketing Ltd v Svenska Spel AB* (n 29) para 25; *Fixtures Marketing Ltd v Organismos prognostikon agonon podosfairou AE (OPAP)* (n 29) para 41; Case C-545/07 *Apis-Hristovich EOOD v Lakorda AD* ECLI:EU:C:2009:132, para 40; Case C-202/12 *Innoweb BV v Wegener ICT Media BV and Wegener Mediaventions BV* ECLI:EU:C:2013:850, paras 33-34, 38.

48 See *Directmedia Publishing GmbH v Albert-Ludwigs-Universität Freiburg* (n 43) paras 39, 46, 47 (referring to Recital 44 of the Database Directive and to *The British Horseracing Board Ltd and Others v William Hill Organization Ltd* (n 31) paras 47-48).

49 Ducato and Strowel, 'Ensuring Text and Data Mining: Remaining Issues with the EU Copyright Exceptions and Possible Ways Out' (n 19) 351.

50 ibid 350.

51 See *The British Horseracing Board Ltd and Others v William Hill Organization Ltd* (n 31) para 54; *Directmedia Publishing GmbH v Albert-Ludwigs-Universität Freiburg* (n 43) paras 51-54.

52 Database Directive, recital 46. *Directmedia Publishing GmbH v Albert-Ludwigs-Universität Freiburg* (n 43) paras 51-54; *The British Horseracing Board Ltd and Others v William Hill Organization Ltd* (n 31) para 55; Ducato and Strowel, 'Ensuring Text and Data Mining: Remaining Issues with the EU Copyright Exceptions and Possible Ways Out' (n 19) 351.

53 Database Directive, art 8(1).

exception could be regarded as ineffective and worthless for TDM users as the successful training of algorithms during a mining process requires the copying of a whole or substantial part of accessed databases.[54] Moreover, the exception applies only to 'lawful users' of a database who can access the database either through contractual agreements or through relevant exceptions.[55] In this regard, the rightholders have control over the contents of a database, as they would be able to restrict or even prohibit TDM through licensing for the sake of their own interests.[56]

The requirement of a 'lawful user' is also a condition of another potential exception provided under Art. 9(b) of the Database Directive that could apply to TDM. The provision allows extraction or re-utilization of a substantial part of the contents of a database for the purpose of illustration for teaching or scientific research.[57] However, the exception consists of requirements which are difficult to meet in the case of TDM. First, the provision requires that users indicate the source of a work when they extract the contents of a database. However, it could be difficult in practice to attribute huge amounts of digital data utilized for mining. The output of TDM research does not contain any parts (samples) of mined works, but only new patterns and correlations. Therefore, it is unlikely that the maker of a processed database would obtain any significant benefits from being listed among thousands of sources. Second, Art. 9(b) requires that the extraction of a substantial part of a database is justified by a non-commercial purpose. This requirement would deprive private actors of the possibility to conduct TDM research under this exception as they usually pursue, directly or indirectly, commercial purposes.[58] The condition would also affect many research organizations conducting research in the framework of public-private partnerships, as it is in practice, difficult to distinguish between commercial and non-commercial activities within these collaborations.[59] Moreover, the optional nature of this provision has led to a fragmented or divergent implementation of the exception within the EU Member States.[60]

To sum up, even though the Database Directive provides some exceptions to the *sui generis* protection under Arts. 8(1) and 9(b), the provisions are neither practical nor effective in the case of TDM. In this regard, researchers facing legal uncertainty under these provisions may find a remedy in Arts. 3 and 4 of the CDSM Directive, which introduce specific TDM exceptions to copyright and the *sui generis* database right. The next section of this article elaborates on the potential application of such exceptions to ChatGPT-based research.

2. Applying the CDSM Directive's TDM exceptions to ChatGPT-based research

Researchers engaging in TDM research on databases, including those using AI language models, such as ChatGPT, would probably benefit from specific TDM exceptions provided under Arts. 3 and 4 of the CDSM Directive. The former provision allows research organizations and cultural heritage institutions to reproduce and extract works for the purpose of text and data mining, providing that they have lawful access to the works and the mining is for scientific research purposes.[61] The latter provision is intended to compensate for the narrow scope of Art. 3, and as a result it benefits any type of 'lawful' user and covers any purposes provided that the use of works has not been reserved by their rightholders.[62] Therefore, researchers are required to comply with certain conditions and limitations to enjoy the TDM exceptions under the CDSM Directive.

Considering real-life scenarios can provide more tangible insight into how these exceptions might function in practice and the effects they could have on researchers. As an example, we could consider TDM research that would employ ChatGPT to analyze large databases of various types of medical data to uncover new insights and patterns with the aim of improving the accuracy and efficiency of medical diagnosis and treatment. This research project would include several essential stages. Initially, researchers need to obtain databases of health-related data either from publicly available sources or through collaboration with medical institutions. Once medical data is collected, researchers need to employ TDM tools to pre-process, clean and transform it into a suitable format for data analysis. Python, a commonly used programming language, is one of the popular tools that can be employed for such tasks. Therefore, researchers could upload the obtained databases into Python through its libraries, and pre-process them in a format that would meet the input requirements of the ChatGPT model. Once the data is pre-processed, researchers need to load ChatGPT into Python by using the Hugging Face Transformers library to perform data analysis.[63] The ChatGPT model can be trained on databases for a variety of natural language processing applications including text classification, clustering, summarization, information retrieval and others. This would help researchers uncover, for instance, recurring medical conditions or symptoms, or identify patterns in the language used in the medical discharge summaries, which could improve medical diagnosis and treatment. The determination of the specific data analysis technique

54 Maryna Manteghi, 'The Insufficiency of the EU's Text and Data Mining Exceptions for Using Artificial Intelligence' (2022) 44 E.I.P.R. 657.

55 *The British Horseracing Board Ltd and Others v William Hill Organization Ltd* (n 31) para 58. See for more discussions on the concept of 'lawful user' eg, Tatiana Eleni Synodinou, 'Lawfulness for Users in European Copyright Law: Acquis and Perspectives' (2019) 10 JIPITEC 20.

56 Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko, 'The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Legal Aspects' (2018) Centre for International Intellectual Property Studies Research Paper No 2018-02, 1, 16 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3160586> accessed 20 April 2023.

57 Database Directive, art 9(b).

58 Rossana Ducato and Alain M Strowel, 'Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility"' (2019) 50 IIC 649, 661.

59 Maria Bottis and others, 'Text and Data Mining in the EU Acquis Communautaire Tinkering with TDM & Digital Legal Deposit' (2019) 2 Erasmus Law Review 190, 193.

60 Geiger, Frosio and Bulayenko (n 56) 12.

61 CDSM Directive, art 3(1).

62 CDSM Directive, art 4(1).

63 Documentation for Python's standard library, along with tutorials and guides, is available at <https://www.python.org/> accessed 12 March 2023.

depends on the type of processed data and the research purpose. ChatGPT, when integrated with Python, could help researchers create a more sophisticated and intelligent application of the medical research in question.[64]

To lawfully mine medical databases, presumably protected under the *sui generis* database regime, researchers should obtain permission from rightholders through contractual agreements or rely on the TDM exception in Art. 3 of the CDSM Directive. However, not all researchers may conduct research under this exception, even though the provision was specifically designed for researchers. Only research organizations and cultural heritage institutions can be the beneficiaries of this exception. This may include universities, research institutions, libraries, museums, or any other entity conducting scientific research on a nonprofit basis or that reinvests all the profits in its scientific research or pursuant to a public interest mission.[65] Therefore, since private actors, such as individual researchers, journalists, small and medium-sized enterprises (SMEs), and start-ups, are usually profit-oriented, TDM research could only be carried out by universities or similar institutions under this provision.

Therefore, private clinics would not be able to mine databases containing health-related data for the purpose of the research in question, without permission from rightholders. Would it be appropriate to prevent such users from conducting TDM research on databases to which they have obtained lawful access through contractual agreements or other lawful means? Why should they pay additional fees to merely analyze such databases? Both commercial and non-commercial research organizations could produce valuable, informative outputs, which are crucial for economic and technological developments.[66] Society will often benefit from profit-driven research. For instance, private clinics may develop predictive models for diagnosing certain conditions that would improve treatment outcomes, and potentially reduce healthcare costs (e.g. excluding some tests and procedures) and also the time that patients spend in the hospital.

Further, researchers relying on the exception in Art. 3 should take into account that the provision allows researchers to employ TDM tools only for the purpose of scientific research. Therefore, researchers would need to prove that their research on medical databases qualifies as scientific research. The process is complex because scientific research is characterized by different approaches, categories and styles.[67] The CDSM Directive does not provide much clarification of this concept, merely stating that it should cover 'both the natural sciences and the human sciences'.[68] The research in question would likely fall within this categorization of the fields of science, as medical research is part of both branches and would also contribute to the current state of science.[69] However, in order to qualify as scientific research, the research project in question should employ scientific methods which should satisfy many characteristics such as replicability, precision, falsifiability, objectivity, reliability, parsimony and others.[70] Therefore, the research in question would not be considered scientific if it is unable to explore medical databases using appropriate research methods.[71]

Even if research organizations prove that the research in question is scientific and their access to databases is lawful, they could still be deprived of the possibility to mine due to the possible application of technological protection measures (TPMs). Article 3(3) of the CDSM Directive allows rightholders to employ such measures only to ensure the security and integrity of their systems and databases. Such measures could, for instance, be applied to ensure that only users with lawful access to mined data can access them.[72] However, if TPMs such as IP address validation or user authentication could be viewed as necessary for controlling access to protected databases,[73] technological measures such as IP address blocking or domain name server blocking may lead to overblocking (blocking of lawful access).[74] As automated filtering technologies cannot properly distinguish between lawful and unlawful access, the 'lock-outs' may occur frequently and take lots of time to resolve.[75] This may affect the research in question and lead to undesirable consequences. If researchers exploring medical databases encounter blockages, they may have to suspend research for weeks or even months until they are reconnected to the digital repositories. The delays could require additional costs and time (e.g. researchers would need to continue receiving their salaries) and could make research outdated and irrelevant.[76] In theory, researchers wishing to conduct the research in question

64 I Gencay, 'ChatGPT and AI Merged in Data Science with Python' (*DataDrivenInvestor*, 13 April 2023) <https://medium.datadriven-investor.com/chatgpt-and-ai-merged-in-data-science-with-python-6ca8c2cb387> accessed 2 May 2023.

65 CDSM Directive, art 2(1) and (3).

66 Geiger and Jütte (n 18) 44. Ducato and Strowel, 'Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to 'Machine Legibility'' (n 58) 651.

67 For more discussions on what can be considered 'science' see Anol Bhattacherjee, 'Social Science Research: Principles, Methods, and Practices' (*Textbooks Collection* 3, University of South Florida 2012) <https://digitalcommons.usf.edu/cgi/viewcontent.cgi?article=1002&context=oa_textbooks> accessed 2 March 2023.

68 CDSM Directive, recital 12.

69 Brian Kennett, *Planning and Managing Scientific Research: A guide for the beginning researcher* (ANU Press 2014) 2.

70 Bhattacherjee (n 67).

71 Kennett (n 69) 2.

72 CDSM Directive, art 3(3) and recital 16.

73 CDSM Directive, recital 16. Case C-484/14 *Tobias Mc Fadden v Sony Music Entertainment Germany GmbH* ECLI:EU:C:2016:689, paras 94, 98-99.

74 Tatiana-Eleni Synodinou, "Intermediaries' Liability for Copyright Infringement in the EU: Evolutions and Confusions' (2015) 31 Computer Law & Security Review 57, 62.

75 Case C-360/10 *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV* ECLI:EU:C:2012:85, paras 50 and 52; Case C-70/10 *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)* ECLI:EU:C:2011:771, para 50; Case C-401/19 *Republic of Poland v European Parliament and Council of the European Union* ECLI:EU:C:2022:297, para 86; *Ahmet Yıldırım v Turkey* App No 3111/10 (ECtHR, 18 December 2012), paras 66-68; *Kharitonov v Russia* App No 10795/14 (ECtHR, 23 June 2020), paras 38-40; See also Julia Reda, Joschka Selinger and Michael Servatius, 'Article 17 of the Directive on Copyright in the Digital Single Market: A Fundamental Rights Assessment' 26-27 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3732223> accessed 1 May 2023.

76 LIBER, 'Europe's TDM Exception for Research: Will It Be Undermined By Technical Blocking From Publishers?' (*Libereurope*, 10 March 2020) <https://libereurope.eu/article/tdm-technical-protection-measures/> accessed 30 April 2023. *Republic of Poland v European Parliament and Council of the European Union* (n 75) para 60.

may rely on another TDM exception in Art. 4 of the CDSM Directive, which was designed to benefit all users seeking to engage in mining. However, they may benefit from this exception unless rightholders reserve the use of their works, either through contractual agreements or through the terms and conditions of their websites, or by implementing TPMs.[77] As Hugenholtz argues, the 'opt-out' mechanism of Art. 4 gives rightholders excessive power to limit or completely prohibit TDM.[78] This may impede research and innovation in different fields, including healthcare and medicine. Moreover, even if rightholders do not reserve the use of their works in an appropriate manner, they may still prevent TDM through licensing terms since, unlike Art. 3, this exception can be overridden by a contract.[79]

Against this background, both research organizations and private clinics seeking to mine databases that contain health-related data could face challenges and legal uncertainty when relying on the CDSM Directive's TDM exceptions. The passive and unclear nature of these provisions could potentially reduce the research power of the EU. In this regard, there is a need to expand the research TDM exception to commercial research, or even introduce a wider TDM exception which would allow TDM for any purpose without a rightholders 'opt-out' mechanism, to improve the situation. A broad TDM exception would ensure greater access to data by researchers, journalists, commercial companies, AI developers and other stakeholders. The improved legal framework for TDM would stimulate new discoveries and advances in various fields that could contribute to economic and technological development, and thus foster social progress and wider innovation. However, broadening the existing TDM exceptions may raise concerns that copyright and database owners could become less motivated to produce new content since they would have less control over their works. The concerns appear to be groundless, as rightholders would still rely on the 'lawful access' requirement, and thus be able to receive adequate compensation for the use of their works through licensing, subscriptions or other lawful means. Moreover, copyright and database owners would still have the opportunity to apply TPMs to ensure the security and integrity of their systems and databases. The said safeguards, therefore, would help rightholders obtain financial gain and also protect their content. Furthermore, the introduction of a wider TDM exception would encourage many investors and business companies to set up and run their businesses within the digital single market (DSM). As a result, the expanding application of advanced analytical tools such as TDM and ChatGPT would significantly increase the demand for data access. This would have to encourage content creators to produce vast amounts of original and quality data as they would still gain licensing revenue for allowing access to such content.

To sum up, the vague and narrow nature of the CDSM Directive's TDM exceptions could force users to consider alternative solutions to avoid potential infringement claims. One of these solutions could be Art. 35 of the proposed Data Act, which clarifies the scope of application of the *sui generis* right provided for by the Database Directive. In particular, the provision excludes certain types of databases from the protection under this regime. Therefore, the next chapter will explore the extent to which Art. 35 of the proposed regulation can remedy the situation in question.

III. The proposed data act in the context of TDM Regulation

1. Background

The proposed Data Act[80] is one of the essential elements of the European strategy for data[81] together with the Digital Markets Act,[82] the Data Governance Act,[83] the Digital Services Act[84] and the AI Act.[85] The regulation aims to unlock huge amounts of digital data which is concentrated in the hands of relatively few actors, and ensure fairness in access to and use of this data.[86] This would be achieved by facilitating access to and use of data by consumers and businesses including clarifying the Database Directive,[87] enabling public sector bodies to use data held by enterprises in exceptional situations,[88] facilitating switching between cloud and edge services,[89] establishing safeguards against unlawful data transfer without notification by cloud service providers and developing interoperability standards for data.[90] The proposed Data Act introduces the right of users to access and use data generated by the use of products or related services,[91] the right to share such data with third parties,[92] and an obligation

77 CDSM Directive, art 4(3) and recital 18.

78 Bernt Hugenholtz, 'The New Copyright Directive: Text and Data Mining (Articles 3 and 4)' (*Kluwer Copyright Blog*, 24 July 2019) <http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/> accessed 12 April 2023.

79 CDSM Directive, art 7(1).

80 Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act), Brussels [2022] COM (2022) 68 final (proposed Data Act).

81 Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions A European Strategy for Data, COM/2020/66 final.

82 Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) (Text with EEA relevance) [2022] OJ L 265, 1-66.

83 Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act) (Text with EEA relevance) [2022] OJ L 152, 1-44.

84 Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance) [2022] OJ L 277, 1-102.

85 Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM/2021/206 final.

86 Proposed Data Act. See also Estelle Derclaye and Martin Husovec, 'Why the *sui generis* database clause in the Data Act is counter-productive and how to improve it?' 1 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4052390> accessed 12 May 2023.

87 Proposed Data Act, cc I-IV and X.

88 Proposed Data Act, c V.

89 Proposed Data Act, c VI.

90 Proposed Data Act, c VIII.

91 Proposed Data Act, c IV.

92 Proposed Data Act, art 5.

for private companies to make data available to public sector bodies to satisfy exceptional needs,[93] including the prevention of a public emergency.[94] Access to data and the ability to use it lie at the core of the EU data policy for promoting innovation and fostering economic growth in the time of increasing use of digital technologies and the internet.[95]

The proposal comes with a special emphasis on broadening access to and use of data collected by sensors and machines based on cutting-edge technologies, such as IoT.[96] The regulation reviews the Database Directive as a part of its broader policy initiative to ensure that the Directive remains relevant in the data-driven society, without impeding access to and use of data within the EU.[97] Particularly, the proposed Data Act aims to get rid of legal uncertainty regarding the relation of the *sui generis* right to machine-generated data.[98] However, only one provision, namely Art. 35, addresses this issue.[99] The provision provides that such protection 'does not apply to databases containing data obtained from or generated by the use of a product or a related service'.[100] The proposed Data Act defines a 'product' as 'a tangible, movable item, including where incorporated in an immovable item, that obtains, generates or collects, data concerning its use or environment, and that is able to communicate data via a publicly available electronic communications service and whose primary function is not the storing and processing of data.'[101] Further, the proposed regulation describes the term 'related service' as 'a digital service, including software, which is incorporated in or inter-connected with a product in such a way that its absence would prevent the product from performing one of its functions'.[102] The proposed Data Act includes only one Recital that relates to Art. 35, namely Recital 84 which further emphasizes that the 'regulation should clarify that the *sui generis* right does not apply to such databases as the requirements for protection would not be fulfilled.'[103]

Therefore, the proposed regulation aims, *inter alia*, to clarify what is not protected under the *sui generis* regime without expressly amending the Database Directive.[104] Article 35 of the proposed Data Act is welcomed as it aims to reduce the availability of excessive IP protection over at least some types of databases.[105] The legal certainty regarding the use of machine-generated data could stimulate innovation and promote research activities within the DSM.[106] Against this background, the following section of this article will examine the application of Art. 35 of the proposed Data Act to TDM, covering not only the positive aspects but also potential limitations or restrictions related to the use of this provision in the context of TDM research.

2. The application of Art. 35 to TDM research

Researchers using ChatGPT to conduct TDM research on medical databases are likely to process different types of databases comprised of data generated by humans,[107] machines,[108] or a combination of both.[109] Article 35 provides a safe harbor for users needing to access, use or share only databases made of data generated by the use of a product or related service (IoT data).[110] This would allow researchers to freely access and *analyze* databases which consist of data generated by, for instance, wearable devices for healthcare monitoring such as fitness trackers, smartwatches, blood glucose monitors, smart technology clothing and others. This would enable researchers to identify fresh insights and patterns to develop new treatments, improve diagnosis, and optimize preventive care.[111] Even though the proposed Data Act intends to free machine-generated data from the constraints of *sui generis* protection, some aspects of the proposed regulation may require further clarification to ensure a greater and fairer flow of data in all sectors, including research and related fields.[112]

Although Art. 35 intends to provide more clarity on the protection of machine-generated raw data, the

93 Proposed Data Act, art 14.

94 Proposed Data Act, art 15.

95 European Commission, 'Study to Support an Impact Assessment for the Review of the Database Directive Final Report' 1 <https://copenhageneconomics.com/wp-content/uploads/2022/02/study-to-support-an-impact-assessment-for-the-review-of-the-database-directive.pdf> accessed 12 March 2023.

96 Giuseppe Colangelo, 'European Proposal for a Data Act-A First Assessment' (20 July 2022) CERRE Evaluation Paper 2022, 1, 6 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4199565> accessed 10 April 2023; European Commission, 'Study to Support an Impact Assessment for the Review of the Database Directive Final Report' (n 95) 1.

97 ibid 1 and 2.

98 ibid 2.

99 Proposed Data Act, art 35.

100 Proposed Data Act, art 35.

101 Proposed Data Act, art 2(2).

102 Proposed Data Act, art 2(2).

103 Proposed Data Act, recital 84.

104 Guido Noto La Diega and Estelle Derclaye, 'Opening Up Big Data for Sustainability: What Role for Database Rights in the Fourth Industrial Revolution?' in Ole-Andreas Rognstad, Taina Pihlajarinne and Jukka Mähönen (eds), *Promoting Sustainable Innovation and the Circular Economy: Legal and Economic Aspects* (Routledge 2022) 26; Derclaye and Husovec (n 86) 2; Inge Graef and Martin Husovec, 'Seven Things to Improve in the Data Act' 1, 4 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4051793> accessed 12 March 2023.

105 European Commission, 'Impact Assessment Report' Accompanying the document Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act), SWD (2022) 34 final, 15; Graef and Husovec (n 104) 4. In this sense, see also Federal Supreme Court (BGH), 25 March 2010, I ZR 47/08 – *Autobahnmaut*.

106 European Commission, 'Impact Assessment Report' Accompanying the document Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act), SWD (2022) 34 final, 16, 70 and 139.

107 For instance, an electronic health record is a database containing patients' health information (eg, diagnoses, allergies, prescribed medication etc,), which is created and maintained by medical personnel.

108 For instance, wearable health devices (eg, fitness trackers, smartwatches, smart inhalers, surgical robots and others).

109 For instance, biobanks are collections of biological samples, which may consist of human-generated data (eg, patient's medical history) and machine-generated data (eg, genetic sequencing).

110 Proposed Data Act, arts 4-5 and art 35.

111 Roberta De Michele and Marco Furini, 'IoT Healthcare: Benefits, Issues and Challenges' (GoodTechs '19: EAI International Conference on Smart Objects and Technologies for Social Good, Valencia, September 2019).

112 Estelle Derclaye and others, 'Opinion of the European Copyright Society on selected aspects of the proposed Data Act' (*European Copyright Society*, 12 May 2022) 5 <https://europeancopyrightsociety-dotorg.files.wordpress.com/2022/05/opinion-of-the-ecs-on-selected-aspects-of-the-data-act-1.pdf> accessed 20 March 2023.

wording of this provision may still raise questions and create uncertainties. The provision excludes databases made of machine-generated data from the *sui generis* database protection to safeguard the rights of users to access, use and share such data specified in Arts. 4 and 5 of the proposed Data Act.[113] The question is whether such exclusion is absolute or only applicable if it interferes with outlined users' rights. A possible clarification could be found in Recital 84 of the proposed Data Act, which indicates that databases comprised of data obtained by connected products should not be protected under Art. 7 of the Database Directive 'where such databases do not qualify for the *sui generis* right' as 'the requirements for protection would not be fulfilled.'[114] These two excerpts from Recital 84 have caused some controversy among scholars. Some claim that the former element of this provision indicates that databases of machine-generated data could still fall within the scope of this right under certain circumstances.[115] Others see this wording as a 'confirmatory statement' claiming that the latter element of Recital 84 suggests that databases of machine-generated data were not eligible for protection under the *sui generis* regime from the outset.[116] However, both perspectives fail to fully elucidate the scope of the *sui generis* right in relation to databases outlined in Art. 35 of the proposed legislation.

As Senftleben observed, the proposed Data Act suggests, *inter alia*, that the 'mere putting into operation of an automated process that constantly creates data is not sufficient to acquire *sui generis* database rights'.[117] But what for example if the installation of sensors or software in wearable health monitoring devices requires a substantial investment? Can it be considered an investment in obtaining data rather than in its creation and thus protected under the *sui generis* database regime? The answer to these questions revolves around the 'obtaining-creating dichotomy', which is an essential element in the CJEU's legal test on the scope of the *sui generis* right.[118] As mentioned above, the CJEU interpreted the term 'obtaining' as referring to resources used to search existing materials and collect them in the database, excluding resources used for creating data itself.[119] However, it could be difficult to draw a clear line between the 'collection' and the mere 'creation' of data in machine-generated databases.[120] For instance, data generated by wearable health monitoring devices should be viewed as 'collected' rather than 'created' since such data already exists. Researchers who obtain measurements, such as sugar levels, heart rate, blood pressure and body temperature, do not create the data itself but rather create a record of that data through the use of monitoring devices.[121]

However, the concept of recorded data[122] raises intense debates among scholars in the context of the 'obtaining-creating dichotomy'. Some academics claim that this data is created because it constitutes 'representations of natural phenomena, not the phenomena themselves,'[123] but others insist that it is both created and obtained since the two activities cannot be separated when recording data that already exists in nature.[124] However, certain scholars, including the author of this article, assume that recorded data refers to the process of obtaining only.[125] In this regard, data generated by wearable health monitoring devices should be considered as obtained rather than created because such information exists regardless of its recording.[126] Health trackers only process pre-existing data to make it 'intelligible' and accessible.[127] In this sense, the *sui generis* database protection could certainly extend to some scenarios involving, for instance, the installation of sensors or expensive software in wearable health monitoring devices as such actions could be considered as a separable relevant investment in obtaining data rather than in its creation.[128] Moreover, there could be a lot of intelligence in the form of algorithms in the sensors that could already select and arrange the data. Therefore, under the proposed Data Act, the presumption that researchers may automatically acquire the right to access databases

113 Proposed Data Act, art 35.

114 Proposed Data Act, recital 84.

115 In this sense see eg, Derclaye and others (n 112) 3; Graef and Husovec (n 104) 4; Derclaye and Husovec (n 86) 2.

116 See for instance Martin Senftleben, 'Study on EU Copyright and Related Rights and Access to and Reuse of Data' (European Commission. Research and Innovation. Independent Expert Report 2022) 50 <https://pure.uva.nl/ws/files/85957820/KI0822205ENN.en.pdf%20p.50> accessed 20 March 2023.

117 ibid 49.

118 Noto La Diega and Derclaye (n 104) 26; Colangelo (n 96) 20; European Commission, 'Study to Support an Impact Assessment for the Review of the Database Directive Final Report' (n 95). See also *Fixtures Marketing Ltd v Organismos prognostikon agonon podosfairou AE (OPAP)* (n 29), *Fixtures Marketing Ltd v Svenska Spel AB* (n 29), *Fixtures Marketing Ltd v Oy Veikkaus AB* (n 29) and *The British Horseracing Board Ltd and Others v William Hill Organization Ltd* (n 31).

119 *Fixtures Marketing Ltd v Svenska Spel AB* (n 29) para 24; *Fixtures Marketing Ltd v Organismos prognostikon agonon podosfairou AE (OPAP)* (n 29) para 40; *The British Horseracing Board Ltd and Others v William Hill Organization Ltd* (n 31) para 80.

120 Matthias Leistner and Lucie Antoine, 'IPR and the Use of Open Data and Data Sharing Initiatives by Public and Private Actors' (Study commissioned by the European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs at the request of the Committee on Legal Affairs 2022) 50 <https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2022)732266> accessed 12 April 2023; Commission Staff Working Document, 'Evaluation of Directive 96/9/EC on the legal protection of databases', SWD (2018) 146 final, 36; Noto La Diega and Derclaye (n 104) 15; *Innoweb v Wegener* (n 47) para 39.

121 In this sense see *Innoweb v Wegener* (n 47) paras 39 and 57.

122 Estelle Derclaye defines recorded data as 'data [that] occur in nature or in time and are generally recorded by instruments of measure in order for them to be intelligible by man. Examples include results of sport competitions, meteorological, astronomical data and genomic data' (Estelle Derclaye, *The legal Protection of Databases: a Comparative Analysis* (Edward Elgar Publishing 2008) 99).

123 Mark J Davison and P Bernt Hugenholtz, 'Football Fixtures, Horse Races and Spin-offs: the ECJ Domesticates the Database Right' (2005) 3 E.I.P.R. 113, 118.

124 Derclaye (n 122) 99.

125 See for instance Matthias Leistner, 'Big Data and the EU Database Directive 96/9/EC: Current Law and Potential Reform' in Sebastian Lohsse, Reiner Schulze and Dirk Staudenmyer (eds), *Trading Data in the Digital Economy: Legal Concepts and Tools* (Nomos 2017) 27, 28.

126 In this sense, see for instance Case C-490/14 *Freistaat Bayern v Verlag Esterbauer GmbH* ECLI:EU: C:2015:735; *Autobahnmaut* (n 105).

127 See for a similar conclusion *Freistaat Bayern v Verlag Esterbauer GmbH* (n 126).

128 In this sense, see Derclaye and Husovec (n 86) 2; Graef and Husovec (n 104) 4; Robbert Fisher and others, 'Study in Support of the Evaluation of Directive 96/9/EC on the Legal Protection of Databases' Brussels: European Commission 2018, 1, 30-31 <http://publications.europa.eu/resource/cellar/244f227a-597d-11e8-ab41-01aa75ed71a1.0001.01/DOC_1> accessed 12 March 2023.

made of machine-generated data may not be absolute in all cases. Such databases could still fall within the scope of the *sui generis* database protection under certain circumstances. Against this background, it would be appropriate to refine the wording of Art. 35 to clarify that databases made of machine-generated data could never enjoy protection under the *sui generis* regime. This minor legislative revision, if implemented, would result in greater data openness.

Another point of uncertainty may emerge when researchers use mixed databases involving data falling within the scope of the proposed Data Act and so-called derived or inferred data excluded from it.[129] Recital 14 clarifies that the latter type of data when lawfully held should not be considered within the scope of the proposed legislation.[130] This would imply that researchers intending to perform TDM on databases containing information (e.g. statistical data such as average heart rate during different activities) derived from data collected by health monitoring devices (e.g. heart rate collected by a fitness tracker) would have to acquire lawful access to it through licensing or other lawful means. In other words, such databases may be covered by the *sui generis* protection. However, it is unclear how researchers would recognize which data is subject to the proposed Data Act and which is not.[131] Moreover, the exclusion of derived or inferred data from the scope of Art. 35 is not well-justified. Recital 14 indicates that to qualify as machine-generated, data should 'represent the digitalization of user actions and events' and be 'valuable to the user and support innovation and the development of digital and other services protecting the environment, health and the circular economy'.[132] But what if derived or inferred data meets these requirements? What if such data constitutes valuable information and insights that may contribute to research? The inclusion of inferential data in the scope of the proposed Data Act would enhance data availability and its integrity for research purposes.

Further, it should be highlighted that the exclusion of databases made of machine-generated data from the *sui generis* protection in Art. 35 of the proposed Data Act would not automatically guarantee researchers the right to access and use such databases. Database holders may still employ TPMs (e.g. password-protected login to a website with data repositories) or establish contractual agreements to control such activities (e.g. block or restrict access to data collections).[133] Although some provisions of the proposed regulation clearly indicate that TPMs or contracts should not be applicable to the detriment of the user's rights and obligations, such provisions refer only to third parties' rights and data sharing agreements.[134] Moreover, the relevance of these limitations to the *sui generis* database right remains unclear. To prevent a risk of data appropriation by means of contractual or technological mechanisms at the expense of users' rights, Art. 35 should be constructed in a way that would clarify that the provision cannot be overridden by a contract or TPMs. This could ensure better access and utilization of data.

Moreover, the main aim of the proposed Data Act is not focused on facilitating or ensuring access to data for the purpose of scientific research.[135] The proposed legislation intends to facilitate the accessibility of machine-generated data by users, trade and business persons and, where there is an exceptional need to access such data, by public sector bodies.[136] Therefore, researchers should affiliate themselves with either of these groups to benefit from data access specified under this new legislation. In this regard, researchers may benefit from the provisions allowing users to share machine-generated data with third parties as 'third party' also covers research organizations or not-for-profit organizations.[137] The data holders should make available such data to these beneficiaries 'without undue delay, free of charge to the user, of the same quality as is available to the data holder'.[138] Moreover, researchers may potentially rely on Art. 14(1) of the proposed Data Act which obliges data holders, in cases of exceptional need, to make machine-generated data available to public sector bodies.[139] In this regard, Recital 56 further clarifies that 'research-performing organizations and research-funding organizations could also be organized as public sector bodies or bodies governed by public law.'[140] Therefore, publicly funded medical research institutions could be able to benefit from this provision when carrying out TDM on databases comprised of data generated by health monitoring devices. Moreover, they are also entitled to share this data with 'individuals or organizations in view of carrying out scientific research'[141] providing that such actors 'act either on a not-for-profit basis or in the context of a public-interest mission recognized by the State.'[142] This implies that independent individual researchers and private research institutions would not be able to acquire even indirect access, based on a collaboration with public actors, to databases made of machine-generated data. Furthermore, the provisions of the proposed Data Act allowing public bodies to access such data and share it with other actors could be only applicable in the case of exceptional needs (e.g. public emergency situations).[143]

To sum up, the proposed Data Act does not prioritize the needs of researchers, resulting in a lack of 'robust access and use guarantees for scientific research' under this legislation.[144] The scope of data access outlined in Art. 14(1) and 21(1) of the proposed Data Act is narrow and thus offers limited benefits to researchers conducting

129 Noto La Diega and Derclaye (n 104) 28.
130 Proposed Data Act, recital 14.
131 Noto La Diega and Derclaye (n 104) 28.
132 Proposed Data Act, recital 14.
133 Derclaye and others (n 112) 5; Noto La Diega and Derclaye (n 104) 28.
134 Proposed Data Act, arts 11(1) and 12(2).
135 Proposed Data Act, art 1.
136 Proposed Data Act, art 1.
137 Proposed Data Act, art 51(1) and recital 29.
138 Proposed Data Act, art 51(1).
139 Proposed Data Act, art 14(1).
140 Proposed Data Act, recital 56.
141 Proposed Data Act, art 21(1).
142 Proposed Data Act, recital 68 and art 21(2).
143 Proposed Data Act, arts 14(1), 15 and 21(1) and recitals 56 and 62.
144 Derclaye and others (n 112) 6.

scientific research based on TDM tools. Against this background, Art. 35 should be amended to enable efficient and broad access to and use of machine-generated raw data collections for research purposes. In this regard, it is crucial to include researchers among beneficiaries of this provision and address their specific needs in the framework of the proposed regulation. This would facilitate analytical work based on data-driven research activities and ensure a research-friendly regime in a time of the growing relevance of data.

IV. Conclusion

Researchers engaging in TDM research on databases, including those using AI language models such as ChatGPT, might not find safe harbors from *sui generis* database infringement under the CDSM Directive's TDM exceptions. The construction of these provisions perfectly demonstrates how strong copyright and database protection, on the one hand, and the limited exceptions, on the other hand, could affect research in the EU. Therefore, researchers may need to seek alternative solutions to avoid potential infringement claims. One of these solutions could be Art. 35 of the proposed Data Act, which intends to free databases comprised of machine-generated data from the constraints of *sui generis* protection. However, researchers may also encounter legal uncertainty when relying on this provision as many aspects remain unclear. The 'refined' wording of Art. 35 and Recital 84 of the proposed Data Act could address many problematic issues discussed in detail in the last chapter of this article. In particular, it needs to be clarified that the databases in question could never fulfil the requirements for *sui generis* database protection. Moreover, it seems reasonable to include derived or inferred data in the scope of the proposed Data Act, as it may constitute valuable information and insights, and also refine Art. 35 in a manner that would clarify that the provision cannot be overridden by a contract or TPMs at the expense of users' rights. And finally, there is a need to improve the wording of Art. 35, or even introduce new provisions, to explicitly address researchers' specific needs in the framework of the proposed regulation. This would facilitate analytical work based on data-driven research activities and ensure a research-friendly regime in a time of the growing relevance of data. To sum up, all the proposed solutions, if implemented, would definitely stimulate cutting-edge scientific research based on TDM and advanced AI tools, such as ChatGPT, within the DSM. This could strengthen the research and innovative power of the EU at a global level.