TEXT AND DATA MINING EXCEPTION - TECHNOLOGY INTO OUR LIVES

Monica LUPAȘCU[*]

Abstract

Since 2016, when the Copyright Reform Directive (Directive in the Digital Single Market) had been proposed for adoption, two major versions of the text were under examination and negotiation in the European Parliament. The comparative study of those two legal texts proposed allowed a more comprehensive understanding of the provisions' scope as well as of the different public interests that sustained the recently introduced exceptions. This paper focuses on interpretation of the proposed provisions of text and data mining exceptions by mostly explaining the technical concepts involved. The text and data mining exception has a very important place between the proposed legal texts since its corresponding provision addresses some new type of uses that should be understood in a technological context and have the potential to affect our lives.

Keywords: *data-mining, copyright exceptions, Copyright Reform, public interest, technology, Big Data,*

1. Introduction. The need to know the technical implications of data-mining in the context of the new copyright changes

This paper aims at presenting the content of the text and data mining exception from a comparative perspective, showing the provision proposed in 2016[1] in parallel with the one adopted[2], to highlight some of the main changes made within its content, namely: (i) the ones regarding the expansion of the sphere of beneficiaries, (ii) modification of the automatic analysis sphere, following which data mining will function over "works and other subject matter", (iii) expanding the analysis scope, by highlighting the fact that the information generated through the process of data mining will not need to be limited to "trends, patterns, and correlations".

Without doubt, these changes would be relevant to be studied from the perspective of the interests that generated them, but before we can study the extent to which an exception can or cannot support public interests, as it is normally and naturally determined by the concept of "exception" itself and in relation to the position that the exception has as a norm, we will focus on the importance of the technical details involved, since this paper incorporates a lot of technical explanations that will translate the relevant terminology.

In addition to the importance that the technical explanations will demonstrate in relation to the study of some of the abovementioned modifications, these will also facilitate the understanding of the exception as a whole and the value it has in the general legislative framework, and not only as a part of what was called "Copyright Reform" at a European level. To identify just one example, as we will show, "data mining" is not limited only to pattern extraction or, even if it will be understood exclusively in relation to pattern recognition[3], the implications of such an assessment brings the process itself closer to artificial intelligence, and this will certainly have its say in appreciating the way in which the expansion of the beneficiary sphere can be understood, indeed as supporting public interests or, on the contrary, certain private interests.

On the contrary, omitting to study the exception by relating to as profound and correct of a study of the technical details involved in data mining, risks limiting the interpretation, the application sphere being incomplete or even completely misunderstood. The example of the erroneous translation of the Romanian version is a small one, yet edifying, the exception of the text and data mining being translated as "extraction of text and data", a title that contradicts not only what the usage within the industry defines as text and data mining, but even the very provisions of the Directive which define the activity itself[4]. The similarities with extraction of text and data, even if they exist from a semantic point of view, do not explain the activity of data mining, the extraction at all, though it exists, concerns a completely different object, with an aim to generate completely different information to the data on which the process of data mining is performed (namely, certain patterns, correlations, etc.). Without advancing too far with the details of the following

[*] PhD Candidate – Faculty of Law, University Nicolae Titulescu (email:monica.lupascu@cyberlaw.ro)

[1] Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on Copyright in the Digital Single Market - https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2016:0593:FIN

[2] http://www.europarl.europa.eu/sides/getDoc.do?type=TA&reference=P8-TA-2019-0231&language=EN&ring=A8-2018-0245#BKMD-16

[3] Pattern recognition - an introduction to data mining https://www.dataiq.co.uk/articles/articles/marpattern-recognition-introduction-data-mining

[4] art.2 paragraph 1, point 2 of the proposed Directive:

"text and data mining means any automated analytical technique aming to analyse text and data in digital form in order to generate information such as patterns, trends and correlations" recital 8 of the proposed Directive:

"those techniques allow researchers to process large amounts of information and to gain new knowledge and discover new trends" chapter, we only mention that the wrongful interpretation of the process as "data extraction" takes the reader towards completely different activities, such as data analysis[5], there being some significant differences[6] between that and data mining.

In any case, this paper aims at studying the exception from the perspective of the technology involved, without insisting over some personal opinions, which, even if they could be inferred from subsequent interpretations, do not represent an objective in itself, their role being rather informative, the study being presented as a start for a future and more complex approach of the same subject not only from the standpoint of the new Copyright Directive, but also of the Directive regarding databases, the new Regulation[7] regarding the protection of personal data, as well as of the Commission's communications in the field of Artificial Intelligence[8].

2. Technical details about text and data mining

What does "**text and data mining**" actually mean? The Romanian version of the Directive's proposed text, mentions exclusively **the extraction of text and data**, which would sound, at least for a connoisseur of copyright, quite similar to the terminology that identifies the permission to use short extracts from works. This would not be hard to understand, because "mining" identifies, indeed, the action of "extracting", "removing", "taking over", in a broad sense – "separation of a substance from a compound". And yet, even if perceived exclusively as an individual term, neither "mining", nor its activity of "text and data mining", is not resumed to just that, a complete interpretation leading to concepts such as "exploitation", "transformation", and even "the release of some ideas or conclusions from a set of facts."

In fact, the technical details of data "mining", highlight, clearly, the fact that the main purpose of this activity is not the data itself, the process of extraction, although it operates on them, regards, in fact, **the patterns and corresponding knowledge** of this data, and not the data itself[9].

Specialized literature tends to equate data the mining of data to its exploitation, to further highlight the analytical and transformative process that is the foundation of this activity. The difference between the simple analysis of data and the exploitation of data (data mining), also mentioned in the introductory chapter, is that "data analysis" is used to test models and hypotheses on the data set in question, for example, analyzing the efficacy of a marketing campaign, regardless of the quantity of data, however, data mining (also known as "data exploitation") uses models of automated learning and statistics to discover **clandestine or hidden models** in a high volume of data. The analysis process and the purpose of each one is, therefore, different, the **main task**[10] **being the semi-automatic and automatic analysis of large quantities of data**[11] **in order to extract previously unknown, potentially useful patterns from databases**[12]**, such as data records (cluster analysis), unusual records (anomaly detection), and connections between data (association rule mining, sequential pattern mining).**

Although a definition can be identified, including relating to what the European law has defined as being this process, "text and data mining" represents an operation fairly hard to appreciate in relation to other technological processes that work in the pattern sphere, of generating new knowledge or predictabilities, being often presented as an expression used interchangeably to also define "pattern recognition", "knowledge discovery" in databases (KDD), the abovementioned "data analysis", "artificial intelligence", and even the entire field of "data science". The differences between data mining and each of the aforementioned processes exist and, nevertheless, using data mining to designate other technological phenomena is not exactly a mistake, because such operations merge with each other, most often, being even difficult to delimit. Without a doubt there are works that have established differentiating elements, useful for both the industry, as well as for the study of the phenomena in question and implications, for example, it is stated that pattern recognition, although aiming at the same purpose as data mining, represents, along with machine learning, a

[5] "The difference between data analysis and data mining is that data analysis is used to test models and hypotheses on the dataset, e.g., analyzing the effectiveness of a marketing campaign, regardless of the amount of data; in contrast, data mining uses machine-learning and statistical models to uncover clandestine or hidden patterns in a large volume of data." - Olson, D. L. (2007). Data mining in business services. Service Business, 1(3), 181-193. doi:10.1007/s11628-006-0014-7

[6] https://www.educba.com/data-mining-vs-data-analysis/

[7] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data - https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1532348683434&uri=CELEX:02016R0679-20160504

[8] Artificial Intelligence for Europe - COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE EUROPEAN COUNCIL, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS - https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe

[9] It should also be mentioned that, in the general sense, text information is integrated in the data category, being, in other words, a subset of it. Although the legal definition identifies, in the designation of this type of analysis, both "text" (text information), as well as "data", the explanations in this paper will regard, without any distinction, the mining of data, which practically includes the operations of the first category.

[10] https://en.wikipedia.org/wiki/Data_mining.

[11] Big Data

[12] It not being yet clear for most people that "text and data mining" is not just text extractions but, primarily, computerized analysis of information, and that "databases" is not limited to MySQL, but involves any collection of information, user-generated content platforms, for example, being based on such databases, collections with information generated by the activity of the users of such platforms.

supervised method, as opposed to data mining that is presented as being one of the unsupervised[13] ones.

In this context, it must be mentioned that, although data mining represents and activity known and practiced for approximately 20 years with profound implications including in regards to the lives of each and every one of us, and pattern extraction, although previously done manually, is hundreds of years old[14], its regulation and the regulation of technologies associated with this type of data analysis, artificial intelligence and automated data processing, only now knows a sustained approach[15], including by attempting to develop ethics policies[16] in the field of artificial intelligence. For example, the definition of artificial intelligence as it was set out through the Communication[17] of the European Commission, presents such an AI system[18] as being that which analyses the environment[19] and acts, with some autonomy, to reach certain goals[20], which would bring the concept closer to those unsupervised methods of data mining, mentioned above, perceived as being some of the analysis methods with the purpose of extracting patterns and correlations, or other types of knowledge. Of course, there is no equality sign between the two "phenomena", artificial intelligence not limiting itself to pattern extraction, and being able to autonomously act (but, in other words – unsupervised, in an independent manner in relation to other systems, or even with the human admin) to reach a larger, wider palette of purposes, nevertheless, the data mining algorithms are frequently used by AI systems, an operation which in itself represents, sometimes, the main component of such a system. On the other hand, as it's been identified through recent studies[21], the knowledge process of it is specific to both methods, and AI techniques can further augment the ability of existing data mining systems to represent, acquire, and process various types of **knowledge and patterns** that can be integrated into many large, advanced applications, such as computational biology, Web mining, and fraud detection.

But that "knowledge" is more than an abstract concept and the need to understand (to assimilate that knowledge) was transposed into computerized systems for a certain reason. "The traditional method of turning data into knowledge relies on manual analysis and interpretation. (…) The specialists then provide a report detailing the analysis. Be it science, marketing, finance, health care, retail, or any other field, the classical approach to data analysis relies fundamentally on one or more analysts becoming intimately familiar with the data and serving as an interface between the data and the users and products. For these (and many other) applications, this form of manual probing of a data set is slow, expensive, and highly subjective. In fact, as data volumes grow dramatically, this type of manual data analysis is becoming completely impractical in many domains. The need to scale up human analysis capabilities to handling the large number of bytes that we can collect is both economic and scientific. Because computers have enabled humans to gather more data than we can digest, it is only natural to turn to computational techniques to help us unearth meaningful patterns and structures from the massive volumes of data. Hence, KDD is an attempt to address a problem that the digital information era made a fact of life for all of us: **data overload**[22]."

But, perceived in abstract, mining and the data on which it operates, may seem like algorithms willing to perform only on a series of 0s and 1s. From a certain perspective this simplification wouldn't be wrong either, but despite that, we would be far from identifying real mining examples. Indeed, the process itself would not mean anything if we wouldn't be able to appreciate **the environment** in which these types of algorithms function, more precisely, **the data that is now in abundance and that is known, effectively, as the World Wide Web**, an enormous collection of information – the largest database that gathers information (or in which information gathers) in text, video, and audio formats, transposing **works that are protected by copyright or not, personal data, metadata**. This association with databases is not a coincidence, disposing all of the information available online, publicly, and through cloud and intranet systems, being subordinated to databases in which **all of this information is organized** in such a way as to be administrated or accessible to the public in a certain form. The electronic data collections that define the concept of database according to Directive 96/9/EC regarding databases[23], are not limited to those MySQL formats that professionals in the field recognize as

[13] although there are several unsupervised data mining techniques (or predictive) which are appropriate when you have a specific target value you'd like to predict about your data. The targets can have two or more possible outcomes, or even be a continuous numeric value.
https://cloudtweaks.com/2014/09/use-supervised-unsupervised-data-mining/ - the article identifies three supervised data mining techniques classification, regression and anomaly detection those unsupervised being – clustering, association and feature extraction.

[14] Bayes Theorem (of the 1700s) and the regression analysis of the 1800s.

[15] https://ec.europa.eu/digital-single-market/en/artificial-intelligence

[16]https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[17]https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe

[18] *Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.*

[19] The environment being represented, in fact, by the data that the AI system acts upon.

[20] *"analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals"*

[21] https://pdfs.semanticscholar.org/d21a/faeffa895c0a641a5aa64248d2401db5f572.pdf

[22] https://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

[23] https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A31996L0009

such, being available for all desktop and mobile applications. To give an even more simple example, even if we are used to interact with the interface of a website, in fact, the interaction is made with the collection of information belonging to that website, the access of which, in reality, is called "database access."

So, aside from the private collections[24] of research organisms and institutions, or the collections identified as such, aside from the specialized format, such as MySQL, all online information is accessible within some databases[25], whether we're looking at simple websites, or large user-generated data platforms, or e-commerce websites, news and media outlets, the only difference lies only in the size and content of the database or in the method through which it can be accessed.

Simplification of the multitude of information mentioned previously and which is found in any types of databases, public or private, is part of the mining process, the knowledge it aims towards meaning, practically, the order of some information disposed randomly and provided from different sources, to identify patterns and correlations. The essence of the data -mining is „the Pattern" - that "Consistent and recurring characteristic[26] or "trait" that helps in the identification of a phenomenon or problem, and serves as an indicator or model for predicting future behavior." Futurist and entrepreneur Ray Kurzweil considers pattern recognition so important that in his recent book, *How to Create A Mind*, he argued that "**pattern recognition and intelligence are essentially the same thing.** Expertise, in essence, is the familiarity of patterns of a specific field." A related concept is that of cause and effect[27]. "We expect meaning in the patterns we see because, in a random universe, it takes energy to create order. So when we see a particular pattern, we expect that through investigation we can identify the force that caused it. **That's how we learn new things**." Due to its predictable characteristics[28], patterns are also used to predict some phenomena and behaviors, and reports can be made regarding the extent to which the present databases will be capable of generating others, similar or identical. In fact, in the study "From Data Mining to Knowledge Discovery in Databases", the author stated that „*data* are a set of facts (for example, cases in a database), and *pattern*[29] is an expression in some language describing a subset of the data or a model applicable to the subset." But the relationship between data and patterns, and the fact that the latter transpose real rules for data, doesn't wholly explain what data mining is, because the result being sought isn't limited to restoring some order in chaos or a certain structure in a set of unstructured data set, but, more than that. As it's explained above, through pattern recognition it's attempted to identify causes of that other information that generated the data itself. These examples also explain the transformative process that data is subjected to, following mining there being discovered some hidden patterns and correlations, non-evident through mere analysis, highlighting them creating a completely new perspective over initial data.

From concretely identifying the elements specific to data mining, namely - databases and the type of information that is being operated on, the essential piece of the puzzle that would be missing for an integral perception of the concept, would be the concrete examples of data mining.

If we were to transpose in actuality the hypothesis in which we should understand a part of all that is stored at the level of the Internet, in a type of social media platform (an example of database in which users are the ones generating the information that end up being stored in the database of the platform in question), we should first organize the chaos of comments, photos, likes, discussion groups, and location tracking, group them according to preferences and observe what kind of correlations there are between them. These would be just a part of the possibilities that would make way for other types of information – the knowledge being sought through data mining. The example of the social media platform (Facebook, Instagram) is not coincidental, other very important examples being that of e-commerce platforms (Amazon, eMag), and especially those of large communities – like Flickr, Github, Gitlab, which constitute immense data resources generated by users, the latter two being extremely relevant to open-source development. Businesses that incorporate these services, which generate information (data) or collects it, either through user activity, or by making available some methods of interaction with products or services, are, in fact, the best examples for this paper, because in each of these cases, the providers make use of data mining methods, for all kind of reasons including those that can translate

"For the purposes of this Directive, 'database' shall mean a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means."

[24] The term 'database' (...) includes a method or system of some sort for the retrieval of each of its constituent materials. A fixture list for a football league such as that at issue in the case in the main proceedings constitutes a database within the meaning of Article 1(2) of Directive 96/9. The expression 'investment in … the obtaining … of the contents' of a database in Article 7(1) of Directive 96/9 must be understood to refer to the resources used to seek out existing independent materials and collect them in the database.
https://ipcuria.eu/case?reference=C-444/02

[25] More information regarding databases is available in the paper: DATABASES AND THE SUI-GENERIS RIGHT – PROTECTION OUTSIDE THE ORIGINALITY. THE DISREGARD OF THE PUBLIC DOMAIN
https://www.academia.edu/36276259/DATABASES_AND_THE_SUI-GENERIS_RIGHT_-_PROTECTION_OUTSIDE_THE_ORIGINALITY._THE_DISREGARD_OF_THE_PUBLIC_DOMAIN

[26] http://www.businessdictionary.com/definition/pattern.html

[27] https://www.forbes.com/sites/gregsatell/2015/05/01/the-science-of-patterns/#4816ced71900

[28] A pattern is a form or model proposed for imitation. The elements of a pattern repeat in a predictable manner.

[29] https://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

into improving user experience or subordination to various marketing strategies.

Below are examples that also show specific data mining methods used to obtain certain results[30].

"**Cluster Analysis** is a data mining technique that is useful in marketing to segment the database and, for example, send a promotion to the right target for that product or service (young people, mothers, pensioners, etc.). Regression analysis, another data mining technique enables us to study changes, habits, customer satisfaction levels and other factors linked to criteria such as advertising campaign budget, or similar costs. Classification analysis is the data mining technique that enables recognizing the patterns (recurring schemes) inside a database thus allowing us to detect spam and improve your marketing strategy performance. To eliminate any database inconsistencies or anomalies at source, a special data mining technique is used called anomaly detection. To avoid using databases infected by intruders (individual values added by hackers, or even viruses that duplicate the data) it is sufficient to search for the intruders, a data mining technique that decontaminates the database and guarantees greater security for the entire system. Association rule learning is used for all product sale activities, especially when large volumes are concerned. Neural networks is one of the latest data mining applications whereby the means you use for marketing operations, i.e. the computer managing your database, "learns" to identify a certain pattern containing elements with precise relationships with each other. The outcome of this learning is the recognition and storing of patterns that will be useful, perhaps not immediately, but in the future to decide whether and how to pursue a goal. The same neural network can also help to recognize the composition of the product or service target more precisely. The last, essential data mining technique, or better said application, is data warehousing. We are now in the sphere of customer (and not only) profiling, especially regarding Big Data processing. Data warehousing means simplifying your database, extracting the most interesting data about your customers, simplifying the creation of detailed reports and much more besides."

3. How does the European law define data mining

The definition given by the Directive in its initial phase[31] (the DSM version proposed in 2016), identified mining of text and data as being **any automated analytical technique that aims at the analysis of text and data in digital form in order to generate information such as patterns, trends and correlations**. This definition has added, in the final form (meaning the version adopted on 26 March, 2019), details regarding the object over which the mining activity will be operated, replacing "text and data" with "**works and other subject matter**" to make even more evident the fact that not only copyrighted information are subject to mining, but also other works (including other "data" not classified as "works" that we can identify as being personal data plus data that is unprotected/unprotectable – still being, for various reasons, part of the public domain).

For a more precise highlight of the modifications from 2016 to 2019 over the exception from the directive, find below a comparative presentation of the text of art. 2, paragraph (2), as it's presented in both versions of the law:

Article 2 – paragraph 2 point 2 - text proposed by Commission	article 2 paragraph 1 point 2 – adopted text
"text and data mining means any automated analytical technique aiming to analyse text and data in digital form in order to generate information such as patterns, trends and correlations;"	"text and data mining means any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations."

The change made in the adopted text also generated a clear increase in the scope of the mining activity, which no longer limits to patterns, trends and correlations. It's much clearer the fact that the extraction doesn't concern the data itself that the mining will be performed on, **but the previously unknown and potentially useful information of the data stored in databases**[1]. This useful information represents, in essence, **a first component of the knowledge sought to be obtained** as a result of this analysis, which may consist of **patterns**, **trends** and **correlations**, but **not only**, obtained through the application of some techniques such as **classification or characterization**. It's not by accident, as shown above, that the data mining can be confused with the whole process of knowledge discovery, since the data extraction, although just one stage, represents an essential component of the process of "Knowledge Discovery in Databases" (or "KDD"). Going through such a system allows the extraction of the essence from the evaluated information, highlighting the connections between data and even the methods through which some of the data is capable of generating similar ones, as part of a process whose goal is predictability.

The legislation supports a part of these technical explanations that identify data mining as part of a data

[30] https://www.egon.com/blog/666-techniques-data-mining-marketing

[31] https://eur-lex.europa.eu/legal-content/RO/TXT/HTML/?uri=CELEX:52016PC0593&from=EN

[1] The text is also supported by the first part of recital 8 ***"Text and data mining makes the processing of*** large amounts of information ***with a view*** to ***gaining*** new knowledge and ***discovering*** new trends possible."

discovery and transformation process, within the recitals of the proposed Directive[2], a relevant role in this context belonging to recital 8 and the following.

In the context of the explanations that the legislator offers regarding the relevance of the new technologies for the automate analysis of electronic data, in the initial form of recital 8[3], it was appreciated that these new types of technologies would allow researchers to process high quantities of data to obtain new knowledge and discover new trends. The researchers are no longer mentioned in the new adopted version of the paragraph, mentioning only that the text and data mining activity allows for the read and analysis of a high quantity of digitally stored data with the purpose of gaining new knowledge and discovering new trends.

The lack of a limitation in the new version of the law creates, in fact, an expansion of the sphere of beneficiaries, **the activity being no longer associated with a certain subject group**. This is also supported by the modification process of recital 5, shifting **from research to innovation**[4], implying, in fact, an expansion of the sphere of beneficiaries.

Another important aspect of the provision is the fact that both versions of recital 8 mention the term of "**process**" or "**processing**", the first mention of its kind that brings **copyright reform closer to the other European initiative regarding privacy**[5] – as data mining and the analysis involved, regardless of the form in which it is perceived and the data on which it operates, **is, above all, data processing**[6], the new Regulation applicable in the field even treating in detail a form of data mining applicable to personal data, namely – **online profiling**[7], defined as being "*any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular* ***to analyse or predict aspects***[8] *concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.*"

An intermediate form of the Directive (from September 2018)[9] which, even not adopted, is important to consider because it brought information to complete the definition of mining as part of a more ample process of interference on data. **The text of** recital 8 a) **provided limitations in exercising the right of text and data mining,** the "**right of access**" to information (lawful access) being mentioned as a condition that needs to pre-empt the activity itself to ensure what the norms called "**content normalization**", meaning the method through which the content's format is changed or it is **extracted** from a database in a **format** that allows the mining of data. The text of this recital 8 a) explained that the process of data mining in itself **does NOT present relevance in the field of copyright**, but, rather the action through which **content is accessed, as well as the procedure through which a certain information is normalized to allow automatic analysis**, so long as this process involves database extraction. So, the exceptions of the text and data mining need to be understood as referring/being applicable (to) **processes relevant to copyright needed for mining**, considered to be, at least according to the paragraph in question, not mining in itself, but "**the right of access" and the right to**

[2] Recitals 5-18 have been identified as being relevant for the new exception of text and data mining introduced through the proposed directive regarding copyright on the digital single market.

[3] New technologies enable the automated computational analysis of information in digital form, such as text, sounds, images or data, generally known as text and data mining. Those technologies allow researchers to process large amounts of information to gain new knowledge and discover new trends. Whilst text and data mining technologies are prevalent across the digital economy, there is widespread acknowledgment that text and data mining can in particular benefit the research community and in so doing encourage innovation. However, in the Union, research organisations such as universities and research institutes are confronted with legal uncertainty as to the extent to which they can perform text and data mining of content. In certain instances, text and data mining may involve acts protected by copyright and/or by the sui generis database right, notably the reproduction of works or other subject-matter and/or the extraction of contents from a database. Where there is no exception or limitation which applies, an authorisation to undertake such acts would be required from rightholders. Text and data mining may also be carried out in relation to mere facts or data which are not protected by copyright and in such instances no authorisation would be required.

[4] If in the initial form of this paragraph, only the "research" (meaning the activity of examination and profound analysis in a certain field made by certain identified entities) was appreciated as representing a public interest solid enough to enforce the implementation of an exception for ensuring a state of equilibrium with the rightsholders, in its revised form, "innovation" joins in as a distinct type of activity, which, although it supposes a smaller scope of actions, being limited by the creation of an improvement, of an added value, can be exercised by any entity, no longer representing a specific of certain entities.

[5] https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1532348683434&uri=CELEX:02016R0679-20160504

[6] Art.4(2) of 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".

[7] several other references explain how data mining techniques are used for profiling users of online services: "Using Data Mining Methods to Build Customer Profiles" by Gediminas Adomavicius and Alexander Tuzhilin New York University - https://pdfs.semanticscholar.org/b298/e06b9ee4b3056c68a023035f228527a891a2.pdf

"Customer Profiling and Segmentation using Data Mining Techniques" by Prof. Tejal Upadhyay, Assistant Professor, Nirma University Atma Vidhani, Student, Nirma University

Vishal Dadhich, Student, Nirma University - http://csjournals.com/IJCSC/PDF7-2/10.%20Tejpal.pdf

Data Mining and Internet Profiling: Emerging Regulatory and Technological Approaches by Ira S. Rubinstein, Ronald D. Lee, & Paul M. Schwartz- https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1116728

[8] Follow the technical explanations from the previous chapter to be able to identify data mining as being a method of identification of patterns to predict their repeatability.

[9] https://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2018-0337+0+DOC+XML+V0//EN

reproduction in a format that allows automatic analysis – “*but the* ***process of accessing*** *and* ***the process by which information is normalised to enable its automated computational analysis****, insofar as this process* ***involves extraction from a database or reproductions****.”* This “**normalization process**” is no longer taken over in its adopted form except as an independent mention within recital 8, but its explanations, as they were expressed in the intermediate form of the directive, can remain valid, because they transpose that which is essential to be studied in the context of data mining operated on data protected by copyright, namely the fact that exceptions work to create a **right of access** and **adaptation**, operations which are implicit to **reproduction and other transformation specific to data mining**.

In any case, even if the form adopted in 2019 no longer takes the full **process description**, the explanations are available, as they describe correct technical explanations, with strict regards to an essential aspect of mining, namely **the data access**, in our opinion implicit, but which, as we will see, will be regulated in a different way within the new articles introduced in the adopted version of the directive, namely art. 3 and 4, texts which represent, in fact, primary provisions in the context of this paper, regulating the new exceptions of text and data mining, through which transposes the permissions to **perform acts of reproduction and extraction for the purpose of data mining.**

Article 3 – text proposed by Commission “Text and data mining Member States shall provide for an exception to the rights provided for in Article 2 of Directive 2001/29/EC, Articles 5(a) and 7(1) of Directive 96/9/EC and Article 11(1) of this Directive for reproductions and extractions made by research organisations in order to carry out text and data mining of works or other subject-matter to which they have lawful access for the purposes of scientific research.”	Article 3 – adopted text “Text and data mining for the purposes of scientific research Member States shall provide for an exception to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, and Article 15(1) of this Directive for reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access.”
	Article 4 Exception or limitation for text and data mining 1. Member States shall provide for an exception or limitation to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, Article 4(1)(a) and (b) of Directive 2009/24/EC and Article 15(1) of this Directive for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining. 2. Reproductions and extractions made pursuant to paragraph 1 may be retained for as long as is necessary for the purposes of text and data mining. 3. The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine readable means in the case of content made publicly available online. 4. This Article shall not affect the application of Article 3 of this Directive.

Aside from the expansion of the sphere of beneficiaries, the first paragraph of art. 3 states that the subjects indicated as beneficiaries of the exception are conditioned by the existence of a “**right of access”** to works and materials over which the computerized analysis will be performed, which contradicts, from a certain viewpoint, what would be specific to mining, from a technical standpoint.

The term of **legal access** imposed as a preemptive condition in this situation, seems to contradict the very activity aimed at being legalized, as well as the concept of exception itself, which should, at least theoretically, transpose **granting of some freedoms, some rights, in essence, of reproduction or extraction**. But these freedoms (or rights) cannot be exercised, in fact, except by accessing the works, for what value would granting rights of reproduction to other subjects, if the access would remain conditioned on the rightsholder's permission? It would be, most probably equivalent with granting a freedom only theoretically, as in practice it would still depend on the rightsholder's will, whereas the existence of an exception in the field of copyright consists of precisely the freeing of the public from the necessity of authorization.

Moreover, if in the case of reproduction, the possibility could be admitted that the "access" represents an earlier stage, therefore different, and which can manifest in time, previously to the act of reproduction itself, in the case of **extraction, access could be confused with the action of extraction itself**, and it cannot be admitted that, in fact, an extraction could be made without the actual access to the database.

The "legal access" phrase cannot be conceived outside of the concept of consent, **legal access** being, above all, the access that is granted, which is allowed, unrestricted, consented or accepted by those entitled to grant and coordinate it, meaning the rightsholders. The **permission to access** the work is, in a way, synonymous with what we could call **authorization** but does the **lawfulness of access** depend entirely on consent?

Or, in other words, **are these rightsholders the only ones entitled to intervene within the action of accessing works to grant the legal attribute, to legalize it**? The answer seems to be affirmative in regards to works already attributed (or that are the object of protection) but what happens in the case of those outside the scope of protection or works that, although belong to certain rightsholders, can be used/modified, in certain limits and circumstances by other individuals specifically identified as part of certain exceptions?

It's possible that the answer to these questions depends largely to the way in which **the institution of copyright exceptions and limitations is interpreted, as involving rights of access or not**. For more details, refer to the author's paper - Public Domain Protection. Uses and Reuses of Public Domain Works[1], in which it's appreciated that the same so-called freedoms derived from the applicability of copyright exceptions and limitation, could not be exercised without the existence of an **implicit right of access**, because the lack of it and the uncertainties in regards to its existence question even the validity of the exception since it is impossible to imagine the public having the possibility to reproduce texts for private purposes without first accessing those works. To call these possibilities (freedoms) "rights", whether we're talking about the **right to reproduce** for private purposes or **the right of access** (implicit, prior, necessary, obligatory) to works for the purpose of reproduction (in the same private purpose), depends on the perception of copyright as a whole, to the extent to which the public interest is or isn't appreciated as being valuable in its relationship with the interests of rightsholders.

Coming back to the text and data mining exception, a correct interpretation of what "legal access" means depends the validity of the exception itself, as a whole, as the rights conferred through derogation itself could not even be exercised in the case in which they are conditioned by an access that could be legal only to the extent that the rightsholders would decide it. **Admission of this situation would be equivalent with a right of reproduction awarded only to the extent to which the rightsholders consent to it** but this would have been achievable anyway, even without any regulation in regards to it, because rightsholders can grant authorization for any kind of use, including for data mining activities.

Until the emergence of the new directive[2], a legal definition of what "lawful access" means does not exist, however. Benjamin Ferrand[3] said that **lawful access** is not limited only to obtaining a permission, but rather to obtaining **the permission that is needed for the intended use**, a first conclusion that can be drawn from this opinion being that of the existence of **an equivalence between right of access and right of use**, as the access detached from the granted permission cannot be admitted. In this vision, the access is not general, but regards a certain type of usage.

Despite this, and especially, in the context of the new provisions in the field of copyright, we cannot place an equal sign between "**access" and "use"**, especially considering the fact that, at the level of the current legislation the term of "access" doesn't replace but, on the contrary, is presented in addition to the term of "usage". An example is the Directive regarding databases itself, which, in Melanie Dulong de Roney's[4] opinion, grants **the maker the right of "access" and "reuse".** We will render below a selection of dispositions from the aforementioned Directive, in which the terms of access and use or reuse are mentioned, especially, with regards to the authorized/lawful user.

*"Whereas, nevertheless, once the rightholder has chosen to make available a copy of the database to a user, whether by an on-line service or by other means of distribution, **that lawful user must be able to access and use the database for the purposes and in the way set out in the agreement** with the right-holder, even if such access and use necessitate performance of otherwise restricted acts;"*

"Whereas the term 'database' should be understood to include literary, artistic, musical or other collections of works or collections of other material such as texts, sound, images, numbers, facts, and data; whereas it should cover collections of independent works, data or other materials which are

[1] https://www.academia.edu/22943385/Public_Domain_Protection._Uses_And_Reuses_of_Public_Domain_Works

[2] See recital 14, which can be interpreted as regulating lawful access.

[3] https://www.copyrightuser.org/understand/rights-permissions/legal-access/

[4] https://halshs.archives-ouvertes.fr/halshs-01572132/document

systematically or methodically ***arranged and can be individually accessed;"***

"art.1. (1) For the purposes of this Directive, 'database' shall mean a collection of independent works, data or other materials arranged in a systematic or methodical way and ***individually accessible by electronic or other means****."*

"art.6 (1) The performance by the lawful user of a database or of a copy thereof of any of the acts listed in Article 5 which is necessary ***for the purposes of access to the contents of the databases and normal use of the contents by the lawful*** *user shall not require the authorization of the author of the database."*

Indeed, each of the above texts seems to identify differently and distinctly "access" over "use", making the term of access to be perceived more as identifying **the action of visualization** preemptive to any form or usage. Despite this, we must consider that the **general** definition of access does not identify visualization or, at least, not the superficial form through which a user becomes aware of the content, but diving deeper, more profoundly, entering. We will remember this definition and we will observe that this context of databases allows the interpretation according to which, most likely, the meaning taken into account by the legislator is that access involves entering the database, similarly to the permissions which require passwords for accessing certain platforms, networks, systems.

To access, therefore, does not only mean to view and is, probably, an action that also depends on the intention of the database maker, who controls entry at the level of the data collection he owns. However, that is, even in the context in which they are accepted as being different actions, **the idea of a common permission must not be rejected**, which is, exactly as the law says – the access and use (being performed in the same time) (of a certain type). This is explained because, in practice, it's harder to accept a situation in which **permission is granted only for access, but not for a method of usage**. A different example, though, would be the one in which the lawfulness of the access depends on the rightsholder previously making available the work for the public, in which case a lawful access could be observed in the case of all information made available to the public from the rightsholder's initiative.

Without a doubt, there are also arguments for which a lawful access can only be authorized expressly by the rightsholder, the permission not being deduced from the simple public release of the work. In these two latter cases, in which, therefore, access must be authorized previously and distinctly from any other form of use, one could ask the question, what will be the situation of public information? Would these need express explanations regarding access and use for the purpose of in-depth analysis (automated processing), for example (text and data mining)? To answer this question, we can look to the new text of art. 4, interpreted in corroboration with recitals 10[5] and 14[6], newly introduced with the adopted directive, both texts transposing a view of the legislator in regards to the **works made available online**. And because reference to free license works – identified as being open source or creative commons, were mentioned within recital 10, a concrete example of the software development communities would be relevant to study in this context. Gitlab[7] or GitHub[8], for example, represent two of the platforms that make available software works in a collaborative system, being subordinate to open-source licenses, each of them having an immense database to which access is offered through sign-up. Undoubtedly, the interest for the mining of such databases is quite high and mainly due to the popularity among software developers, these communities attracted developers from all over the world, who brought important contributions to the industry that lead to many developments including in the field of artificial intelligence. The specific nature of these works lies, however, in the fact that they are freely licensed, most of the applicable licenses allowing reuse of those specific works[9].

Referral to art. 4 in this context is justified by the fact that, in essence, this article **transcribes a data mining permission in a context that is not subordinate to research**[10], permitted, practically, to all subjects of right, but conditioned by the same **lawful**

[5] (10) Union law provides *for* certain exceptions and limitations covering uses for scientific research purposes which may apply to acts of text and data mining. However, those exceptions and limitations are optional and not fully adapted to the use of technologies in scientific research. Moreover, where researchers have lawful access to content, for example through subscriptions to publications or open access licences, the terms of the licences could exclude text and data mining. As research is increasingly carried out with the assistance of digital technology, there is a risk that the Union's competitive position as a research area will suffer, unless steps are taken to address the legal uncertainty concerning text and data mining.

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

[7] https://about.gitlab.com

[8] https://github.com

[9] However, an essential attribute of free licenses is the fact that reuse is conditioned on making the work available with the same open source terms. There are variations, of course, there being several types of similar licenses.

[10] Art. 3 and 4 transcribe, in fact, the two exceptions that will function for data mining operations, one that will have a purpose dedicated exclusively to **research field**, deduced also from the title of art. 3 and from the restricted beneficiary sphere in the field of research organizations and cultural heritage institutions, and the other – art. 4, **which isn't constrained by such a beneficiary sphere**, the absence of an express

access mentioned in art. 3. An additional mention, which also marks an essential difference between the exception provided by art. 3, is pt. 3[11] of this article 4, in which it's mentioned that the permission in question will function **only if the right was not expressly reserved by rightsholders**. This mention allows the interpretation that, in the absence of an express mention that would exclude data mining from the authorizations accepted by rightsholders, the beneficiaries of the exception will be allowed to perform data mining activities on the data in question. As opposed to the art. 4 situation, in the context of mining performed by **researchers** (art.3), the mentions in recital 10 allow the interpretation according to which, **even in the situation in which the terms of licensing will exclude mining, the permissions granted by virtue of the exception will be able to function**, this also being, most probably, the reasoning that justified the adoption of a mandatory exception for the field of research.

For an overview on the way in which mining will function according to the two articles from the Directive, we consider the text of recital 14, which, although debuts with a referral to the research sphere, can be applicable in both situations, as the definition it incorporates is no longer aimed at a specific field of beneficiaries. Another argument in support of general applicability is the fact that research organizations were mentioned only with the title of example or in contexts which represent an alternative to situations of (non-differentiated) access to content based on free licenses.

"(14) ***Lawful access*** *should be understood as covering* ***access to content based on an open access policy*** *or through* ***contractual arrangements*** *between rightholders and research organisations or cultural heritage institutions, such as* ***subscriptions****, or through other* ***lawful means****. For instance, in cases of subscriptions taken by research organisations or cultural heritage institutions, the persons attached thereto and covered by those subscriptions should be deemed to have lawful access.* ***Lawful access should also cover access to content that is freely available online****."*

By virtue of the above text, a text made publicly available or to which the access is based on free license, involves, clearly, **a context in which access is appreciated as being lawful**, **allowing mining,** even in the context of art. 4, that is even in other purposes other than those of research, if we consider the provisions within the Directive. As the text of recital (14) expresses, access is interpreted as not being able to be differentiated from "viewing", with the mentions previously stated in this paper, a paper made available without the restriction of an imposed password, **being considered freely accessible and, most important, freely to be mined.**

We come back to the example of the aforementioned big collaborative development platforms to mention that this interpretation can only be to the detriment of these communities, the results of which could be accessed lawfully by any entity, by virtue of the above mentioned provisions. The only way to make this exception inapplicable would be, as per the text of art. 4, pt. 3, providing some special exceptions to forbid data mining, but this would only contradict the spirit by virtue of which these communities were created.

4. Conclusion

The issue of data mining is far from being resolved through this paper. Moreover, as it was mentioned at its beginning, this was not the purpose, not only due to the fairly complex technical details, but especially to the fact that the text and data mining permissions are newly regulated, there being no history of neither jurisprudence nor doctrine to support the method of approach and nuanced interpretations. In addition, the implications of the operation in itself, are multiple because, as could be seen in the chapter on technical details, operations included a lot of specific techniques, any one of which being able to operate for a variety of purposes and on a diversity of data. In this context, an aspect that could not be integrated in this paper and is necessary to be approached in other studies, is the information on which mining would be performed, the existing differences between them giving rise to new possibilities of interpretation.

Another aspect which wasn't analyzed in its entirety is that of the sphere of beneficiaries, this subject being one that could be discussed in a separate paper, as beneficiaries are to be appreciated including in correlation with the types of data on which mining could be performed. This is one very important aspect since the appropriate identification of the beneficiaries has relevance for the study of the public interest that justified the adoption of the new exceptions in the European legislation on copyright.

Despite the shortcomings generated, mainly, by the extent of the subject, this study primarily clarified essential technical details needed to understand any subsequent interpretation on the legal text of the exceptions. Moreover, the study of lawful access in this context, brings new light on works presented/made available online, especially those under free licenses.

mention being understood as a possibility awarded to any subject of law of becoming a beneficiary of it. - Ubi lex non distinguit, nec nos distinguere debemus = where the **law does not** distinguish, **nor** the interpreter must distinguish.

[11] The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine readable means in the case of content made publicly available online.

References

- Proposal for a Directive of European Parliament and the Council on Copyright in the Single Digital Market;
- Adopted text of the Directive on Copyright in the Single Digital Market;
- An introduction to data mining and other techniques for advanced analytics – Journal of Direct, Data and Digital Marketing Practice, Vol 12, act-dec 2010;
- Artificial Intelligence for Europe – Communication from the Commission to the European Parliament;
- Data Mining v. Data Analytics – an online course by Corporate Bridge Consultancy Ldt;
- Use of supervised and unsupervised data-mining by Keith Cawley;
- Data Mining – an AI perspective by Xindong Wu;
- From Data Mining to Knowledge Discovery in Databases by Usama Fayyad, Gregory Piatetsky, Padhraic Smyth;
- How to Create a Mind by Ray Kurzweil;
- Legal Access – by Benjamin Ferrand;
- The Legal and Policy Framework for scientific data sharing, mining and reuse.