UNIVERSITY OF LAPLAND

FACULTY OF LAW

INTERNATIONAL AND COMPARATIVE LAW

OTMEVAL0033 Master Thesis

RYKOV VIKTOR

Big Data and IP Law: Risk Assessment and Fostering Data-Driven Transactions

Autumn 2018

CONTENTS

INTRODUCTION

Big Data is one of the newest concepts in the field of data analysis. In general, results of data processing may have multiple possible applications. Due to this reason phenomenon of Big Data proved to be useful in numerous fields of economic activities. In general, the Big Data phenomenon was brought into existence with the development of computers‘ technological capacity to analyse extremely large datasets. From the legal perspective, Big Data proves to be revolutionary across various fields of law: competition law, data protection, law on confidentiality, copyright law, database law – to name a few. Quite obviously, fruitful research cannot be conducted within all affected areas of law simultaneously. Hence, intellectual property law is primarily in focus of the present research, and, more specifically, I would scrutinise mechanisms of confidentiality, copyright and database protection.

So, to contribute scientific dialog and provide some guidance for businesses, this paper attempts to shed the light on the Big Data phenomenon from the standpoint of IP law. While numerous legal implications related to IP arise with respect to the Big Data exploitation, there is little of consistent research approaching Big Data and data mining from the IP law perspective. This conflicts with the fact that Big Data is at heart of numerous scientific and business projects. Hence, the objective of this study is to analyse interactions between IP law and Big Data, providing some guidance for those interested in extracting economic value associated with Big Data. Apart from revealing legal risks connected with the Big Data exploitation, this study searches for best ways to incentivise investments in data-driven innovation.

Taking step forward, it is important to note that along with the term ‗Big Data‘ I frequently employ the term ‗data mining‘ on the pages of this paper. While more insights on the terminology would be provided later, it is necessary to mention on this stage that ‗data mining‘ can be simply defined as the knowledge discovery process performed on digital datasets with application of automated algorithms and humans‘ skills and judgement. I provide this core definition here, in the introduction, because otherwise it is troublesome to draw a clear illustration of the research basic elements.

Boundaries of the Study and its Structure

As I already stated, this research devoted to IP law, and, more specifically, to law on confidentiality, copyright and database law. Thus, such related fields as privacy and competition law fall outside the scope of this study. Nevertheless, as Big Data is an interdisciplinary concept, sometimes comments regarding privacy or competition law are unavoidable.

The whole study is introduced from the perspective of data mining entities, meaning market players which invest their economic resources in exploiting Big Data. Their interests are

analysed as opposing to the interests of database producers and other holders of IP rights enshrined in Big Data. Hence, when I address legal risks associated with the exploitation of Big Data I almost exclusively refer to risks of a data mining entities‘ business. Furthermore, apart from highlighting legal risks, I attempt to provide a piece of advice on the business strategy for data mining entities.

Following commentaries on the structure of the thesis should serve as the clarification of the very essence of my research. Accordingly, it would be easier to comprehend ideas behind research questions. The paper is logically structured as follows. There are five paragraphs. Paragraphs are divided in sections which, in turn, are divided in subsections. Conclusion reveals whether posed research questions were answered and suggests directions for further scientific studies.

First paragraph highlights technical sides of the Big Data phenomenon, its so-called V-characteristics. Furthermore, some illustrations of Big Data projects are introduced in the very beginning of the study. Technical side of the issue is important, as we should realise factual relations before introducing relevant legal concepts. Subsequently, first paragraph explains terminology. Big Data and data mining, being key terms, are explained in detail. Moreover, such related terms as data analysis and information analysis are also clarified.

Paragraphs through second to fourth provide a thorough analysis of the applicable legislation. All three paragraphs provide a comparative study of EU and US laws. Analysis starts from the introduction of general laws on database and copyright. Database law is strongly connected with the Big Data phenomenon, because when market players use Big Data they factually exploit tens and hundreds of databases. Hence, it is highly important to examine EU database law. Database law similar to the European is almost lacking in the US. I am stating ~~–a~~lmost", because the hot news doctrine exists within the US legal order, and respective part of the thesis would demonstrate that this doctrine can be viewed as a modicum of the EU database protection. Another subject of scrutiny is copyright law, because Big Data usually includes vast amounts of texts, sounds, video and other copyrighted materials. The EU copyright law is examined in detail, while the US copyright law is scrutinised mainly in the part related to copyrights in selection and arrangement of databases. Fourth paragraph, mainly mirroring the structure of second paragraph, is devoted to exceptions and limitations under database and copyright laws. This paragraph answers a question whether data mining can benefit from general exceptions and limitations under EU and US law.

Besides scrutinising general copyright and database law, second paragraph highlights problems related to such concepts as ownership and access to data. Examination of these concepts helps to view Big Data strategically. In other words, if the analysis of database and

copyright law, as such, is focused on the general risk assessment, examining concepts of ownership in and access to data would provide specific guidance on the risk evaluation, which can be used by businesses.

Third paragraph is devoted exclusively to law on confidentiality. I believe that confidentiality, being of a distinctive nature from other areas of IP, can be better examined separately from database and copyright law. Latter two reflect, in the nutshell, statutory stipulations, while confidentiality is largely based on contractual arrangements. Due to this reason third paragraph also contains fewer risk assessment considerations and more advice on the business strategy. In particular, I assess how data mining entities and data producers can benefit from law on confidentiality in the negotiation process.

Fifth paragraph is concerned about legislative and alternative solutions, specifically addressing Big Data implications. For instance, the UK and Japan already implemented data mining exceptions into their IP laws. I consider that these exceptions can be better examined in the separate part of the study, as they stem from the recently implemented unique legislation. Furthermore, the EU proposal, following approach of the UK legislators, is currently pending before EU legislative bodies and is also examined in fifth paragraph.

Remaining part of fifth paragraph discusses scientific and alternative social initiatives. Scientific initiatives complement the picture and give deeper understanding of the preceding analysis of legislative sources. Alternative social initiatives rest upon the idea that our society can frequently fix a lack of legislative solutions or existence of extreme solutions through adaptation of a ‗soft law'. Creative Commons and Free Open Software movements are taken as valuable examples. Ideas and principles guarded by these movements are applied to the Big Data phenomenon. In the nutshell, while first four paragraphs deal mainly with the issues of risk assessment and business strategy, fifth paragraph highlights additional objective: unpopular legislative solutions can be balanced through social initiatives. Conclusion provides whether research questions are answered and suggests directions for the future scientific studies.

Methodology of the Research

One of traditional methods of the research in jurisprudence is method of legal dogmatic. Following terms related to this type of research method are usually perceived by the scientific society as synonyms: ‗legal doctrine, black Letter law, formalism, doctrinalism and legal dogmatic research'[1]. Accordingly, method of legal dogmatic can be defined ‗as research that aims to give a systematic exposition of the principles, rules and concepts governing a particular

[1] See, p. 5, Smits, Jan M., *What is Legal Doctrine? On the Aims and Methods of Legal-Dogmatic Research* (September 1, 2015). Maastricht European Private Law Institute Working Paper No. 2015/06. Available at SSRN: https://ssrn.com/abstract=2644088.

legal field or institution and analyses the relationship between these principles, rules and concepts with a view to solving unclarities and gaps in the existing law'[2].

This definition of legal dogmatic strictly governs approach to the methodology which I follow throughout my research. I will clarify principles and concepts behind Big Data, and then I will examine in detail rules of particular legal fields – law on confidentiality, copyright and database law. Upon clarifying relationship between Big Data and relevant legal rules and principles, I will provide suggestions how to eliminate defined gaps and uncertainties. Method of legal dogmatic, therefore, is highly valuable for my research. More specifically, I will use method of legal dogmatic to weigh existing case law and scholarly opinions related to Big Data against applicable legal provisions. It helps to emphasise interrelatedness of legal norms and circumstances of economic interactions.

Another important research method applied in this study is comparative. With the exception of the first paragraph, I will use comparative method throughout the whole paper. In second and third paragraphs I compare EU and US law to define which legal order better suits Big Data transactions. In fourth paragraph I will briefly examine different legislative approaches to the concept of confidentiality. In fifth paragraph I will compare legislative and alternative approaches to Big Data implications.

Additional methods of this study are economic and empirical. Economic method helps to address an issue, whether granting IP rights in specific cases are excessive from the standpoint of social costs and benefits. Moreover, economic method reminds a reader that the purpose of law is to support, not to hinder, economic growth of society. Empirical method is employed due to the shortage of case law and relevant scientific studies. Hence, while addressing some research questions, I either draw hypothetical illustrations or bring case-study examples from sources of reference.

Research Questions

Risk assessment of data mining and examination of incentives for data-driven transactions are at heart of this research, as follows from its title. Introduction of the structure has revealed that legal risks would be evaluated largely on the basis of analysing applicable laws. In turn, incentives for businesses can be established if most dangerous risks are determined and mitigated. Incentives for businesses are approached through assessment of possible business strategies in Big Data settings. Accordingly, to realise objectives of risk assessment and business strategy I pose following research questions:

[2] Ibid.

1) Which relationships define Big Data from practical and technical standpoints?
2) How key terms related to the Big Data phenomenon can be defined?
3) Which risks of Big Data exploitation are associated with copyright and database laws?
4) Whether and how Big Data projects may benefit from using confidentiality clauses?
5) Whether general limitations and exceptions under IP laws cover data mining activities?
6) Whether data mining activities infringe copyrights in specific literary and artistic works?
7) Which modern initiatives, legislative and alternative, are available to incentivise investments in Big Data?
8) Whether and how data mining entities could mitigate risks under IP laws currently in force?

These and also some other questions posed in the course of the discussion would serve as useful orienteer for this study.

1. DEFINING BIG DATA: SCIENTIFIC SPECULATIONS AND LEGISLATIVE APPROACHES

Big Data as any novel concept has many various definitions but conventional one. At the same time, a certain consensus exists in the literature with respect to the Big Data crucial characteristics. Furthermore, preparatory legislative documents in the EU and recent legislative changes, introduced in the UK and Japan, have already coined definitions determining first features of the Big Data legal framework. Accordingly, attempting to define Big Data from the legal standpoint, this paragraph is structured as follows. First section reveals main characteristics of Big Data elaborated in the scientific literature and provides examples of Big Data projects. Second section deals with the term ‗data mining', analysing relevant scientific and legislative approaches and examining neighbouring concepts.

1.1. Big Data Framework: 4V-Characterisitcs and Insights of Big Data Projects

There is no generally accepted definition of Big Data. Scholars employ this collocation to depict various concepts and to solve problems in distinct areas of scientific knowledge. Here are very few definitions elaborated so far by scholars.

‗Big Data is data that exceeds the processing capacity of conventional database systems', notes Edd Dumbill O'Reilly[3]. It is troublesome for the conventional systems to process Big Data, mainly, due to its so called ‗3V'-characteristics which were firstly coined by Gartner Inc.: ‗Big data, in general, is defined as high volume, high velocity, and high variety assets that demand cost effective, innovative forms of information processing'[4]. These characteristics – volume, velocity and variety – refer respectively to the volume of data, speed of its processing and to the various forms in which data are presented. Moreover, in the recent works scholars coined fourth V-characteristic, veracity of data, which refers to trustworthiness of Big Data[5].

In general, V-characteristics depict Big Data technical sides; although at the same time these characteristics signal existence of legal implications. For instance, it is quite clear, that rights of numerous natural and legal persons intertwine in Big Data due to volume and variety of datasets concerned.

In addition to existing legal implications some papers raise ethical concerns with respect to implementing Big Data practices. Professor E. Adar opines that ‗the confounding of academic and industrial practice of Big Data under one name [...] makes it difficult to hold the position of being for academic Big Data practices, but against aspects of corporate practice [...]. This particular feature will hopefully become irrelevant as we move away from the Big Data moniker and adopt more specific ways of discriminating between techniques, applications, and values'[6].

It follows from Professor E. Adar's notion that currently the Big Data definition comprises distinctive techniques, applications and values, and this is why academic and industrial practices can be contradictory. Commercial application of Big Data is concerned about advertising, insurance and credit rating, search of employees, to name a few. Scientific application of Big Data is relevant for development of new pharmaceutical products, urban planning and innovative research, though not exclusively, and businesses may be also involved. Although this paper is not specifically concerned about science-versus-commerce dichotomy of Big Data practices[7], existing contradictions demonstrate that the Big Data phenomenon should be approached cautiously, and terminology employed should be precisely defined.

Nevertheless, even under pressure of ethical and privacy concerns, successful Big Data projects do exist. To provide an example of the data-driven decision making in the aviation industry, one Big Data project has successfully realised the objective of eliminating a gap

[3] Dumbill, Edd. *What is big data?* Available at: http://radar.oreilly.com/2012/01/whatisbigdata.html.
[4] *The Big Data Explosion: Maximizing information value while minimizing risk*. (2013) Information Management, Volume 42, (2), p s2.
[5] See, p. 347, Rubinfeld L., Daniel; Gal S., Michal. *Access Barriers to Big Data*. 59 Ariz. L. Rev. 339, 382 (2017).
[6] P. 766, Eytan Adar, The Two Cultures and Big Data Research, 10 ISJLP 765, 782 (2015).
[7] See, for more information on Big Data and ethics: boyd, danah and Crawford, Kate, Six Provocations for Big Data (September 21, 2011). A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 2011. Available at SSRN: https://ssrn.com/abstract=1926431.

between estimated and actual arrival times of aircrafts in the airport. Combining and examining data about weather conditions, flight schedule and proprietary data provided by an aviation company, the PASSUR Aerospace was able to erase a gap between estimated and actual arrival times, saving arguably millions of dollars for the companies managing airport services[8]. Quite obviously, this monetary loss arising from the gaps is directly associated with corresponding social costs, and successful realisation of this Big Data project removed these costs.

Another illustrative example is the Big Data project organised by Telenor, a global mobile operator, collaborating with the Harvard T.H. Chan School of Public Health, Oxford University, the U.S. Centre for Disease Control, and the University of Peshawar. The project's objective was to find the cure against epidemic disease, dengue fever. It was realised through analysing anonymised call data from more than 30 million users of Telenor mobile services in Pakistan. Large volume of data at the disposal of collaborating parties afforded „to accurately map the geographic spread and timing of the epidemic"[9].

These examples serve as the valuable illustration why Big Data is translated in V-characteristics – volume, velocity, variety and veracity – and why the Big Data discussion is relevant from the pure perspective of social costs and benefits. To summarise the analysis so far, for the purposes of the subsequent discussion I define Big Data as the term referring to large sets of data fixed in digital form which cannot be processed by conventional computing tools and which utilisation, therefore, requires more substantial financial and laborious investment for extracting commercial and/or social value.

1.2. Big Data Definitions: From Buzzwords to Legal Concepts

The Big Data phenomenon not only refers to large digital sets of data but embraces wider range of concepts. In the previous section I have already employed the term „Big Data practices", and similar to it would be the term „data mining practices". Data mining, computational analysis and data analysis – those are the concepts constituting Big Data practices and synonymous one to another to the certain extent. Before clarifying the difference between provided terms, let us start with examining terms „Big Data" and „data mining" as methods of data science.

According to the literature, Big Data is perceived not only as datasets with V-characteristics but is also considered as a method of, or tool for, analysing data. Here are some examples. Frank Fagan used „quantitative machine learning and "big data" processing techniques" in order to analyse the 2,100 court decisions on the topic of the successor liability

[8] See, for more details: McAfee, Andrew: Brynjolfsson, Erik. *Big Data: The Management Revolution*. October 2012. Harvard Business Review. Available at: https://hbr.org/2012/10/big-data-the-management-revolution.
[9] See, for more details: https://www.gsma.com/mobilefordevelopment/programme/digital-identity/big-demand-for-big-data-new-telenor-study-on-dengue-fever-in-pakistan. Accessed on 08.02.2018.

under the US law[10]. Authors of another paper consider Big Data scientific toolset useful for the automatic discovery of prior art in published patents[11].

Professor M. Mattioli, revealing insights of Big Data business practices, opines: ‗The term, [Big Data], refers to a new method of empirical inquiry'[12]. The author further explains at length that data mining combines employment of advanced algorithms with human skill and judgement, when it comes to arrangement of extracted data. While collaboration of machine and human analysis techniques would be examined in the third paragraph of this paper, it is necessary to emphasise: Big Data is viewed within the literature as both object of data analysis and method of such analysis.

The reasonable question would be whether such terminological co-dependency is justified. When Big Data is defined as a pool of datasets there is little of the conceptual misrepresentation. Big Data in this case simply refers to all the information out there fixed in a digital form and suitable for processing and analysis. However, if Big Data is defined as a methodological tool for analysing a large number of datasets, the scope of this definition is not clear. In my view, from the methodological standpoint, use of Big Data for reference to both mined datasets and methods of data mining appears to be misleading.

As an alternative, to define methods employed in Big Data projects, I suggest to use more precise well-established terms mentioned above: data mining (a), computational analysis and data analysis (b).Question which still arises with respect to the scope of mentioned definitions is following: does it refer to the employment of a mathematical algorithm which mines datasets, or does it go further and includes statistic and social science practices employed by humans dealing with the results of the algorithm use? I attempt to address this question below.

a) Data Mining

To the date, both legislative and scientific definitions of data mining were introduced. ‗Among economists, data mining often refers to statistical tests run on quantitative data without proper theoretical preparation'[13], - says one research. Another researcher provides more illustrative explanation, stating that analysts employ ‗multiple data sorting techniques [...] such

[10] See, Frank Fagan, From Policy Confusion to Doctrinal Clarity: Successor Liability from the Perspective of Big Data, 9 Va. L. & Bus. Rev. 391, 456 (2015).
[11] See, Amir H. Khoury; Ron Bekkerman, Automatic Discovery of Prior Art: Big Data to the Rescue of the Patent System, 16 J. Marshall Rev. Intell. Prop. L. [i], 65 (2016).
[12] P. 539. Michael Mattioli, Disclosing Big Data, 99 Minn. L. Rev. 535, 584 (2014).
[13] p. 2, Handke, Christian and Guibault, Lucie and Vallbé, Joan-Josep. *Is Europe Falling Behind in Data Mining? Copyright's Impact on Data Mining in Academic Research* (June 7, 2015). Available at: https://ssrn.com/abstract=2608513.

as clustering, classification, and sequence analysis‘ to reveal ‗previously unseen patterns and relationships from large datasets and derive a business value from these‘[14].

Legal scholar G. Tzanis provides a clear-cut methodological explanation of data mining. He defines data mining as ‗the main step in the process of knowledge discovery in databases‘[15]. The whole process, according to G. Tzanis, includes pre-processing (data selection, cleansing and transformation), data mining itself which converts pre-processed data into pattern models and, finally, post-processing (evaluation, interpretation) which results in the knowledge. However, G. Tzanis further notes that the term ‗data mining‘ is frequently used to address the whole knowledge discovery process[16].

The authors of the EU study on text and data mining define ‗data mining‘ as the ‗automated processing of digital materials, which may include texts, data, sounds, images or other elements, or a combination of these, in order to uncover new knowledge or insights‘[17]. This study has examined legal and economic background of text and data mining, on which basis the European Commission subsequently drafted the Proposal for a Directive on copyright in the Digital Single Market (hereinafter – the Proposal). This Proposal enshrines following definition: ‗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‘[18].

In my point of view, the definition provided in the EU study is a better option than this provided in the Proposal. Drafters of the latter have decided to include two elements, text and data, in the term, while any text is simply one possible form of data. Hence, the collocation ‗text and data mining‘ is logically inconsistent. Furthermore, the Proposal lists types of information which should result from data mining – patterns, trends and correlations. It is not clear from the wording of the definition whether the list is exhaustive or not. At the same time, authors of the EU study use broader reference to ‗new knowledge or insights‘. In the nutshell, the definition in the Proposal can create more implications with respect to its possible interpretation than the definition provided in the EU study. In any case, as the Proposal should go through the whole legislative process, one can still hope that more consistent wording would be adopted.

[14] Brooks et al (2017). *Artificial Intelligence vs. Machine Learning vs. Data Mining 101 – What's the Big Difference?* (Guavus Blog). Available at: http://guavus.com/artificial-intelligence-vs-machine-learning-vs-data-mining-101-whats-big-difference/.
[15] p. 5, Tzanis, George. *Biological and Medical Big Data Mining* (January 2014). Available at: https://www.researchgate.net/publication/261958613.
[16] Ibid.
[17] P. 17, European Union. De Wolf & Partners. *Study of the legal framework of text and data mining (TDM).* 2014.
[18] Text of the Proposal is available at: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52016PC0593.

b) Data Analysis and Computational Analysis

Data analysis is the concept more general in comparison with data mining, because the latter, strictly speaking, refers to the single stage of the knowledge discovery process. Mainly due to this reason authors of the EU study on text and data mining argue that the term ‗data analysis' is more suitable for the legislative implementation[19]. The following statement also supports their vision: ‗The challenge of better law-making is that of enacting provisions at a sufficient level of generality to make copyright principles applicable to unforeseen situations'[20].

For instance, authors of the EU study argue that the action ‗to mine' refers only to extraction of data, while analysis overarches such operations as ‗crawl, process, compare, copy, analyse, retrieve, interpret, search, sort, parse, remove'[21]. While it might be more suitable for the prospective legislation to incorporate the term ‗data analysis' rather than ‗data mining', I would use throughout this paper primarily the latter, due to the reason that it is usually relied upon in the scientific sources of this research.

Computational analysis is the term introduced in the UK copyright law to provide data mining exception for non-commercial research. This exception would be examined in further detail later, and here only comments concerning the term itself are necessary. According to the Oxford dictionary ‗to compute' means ‗reckon or calculate (a figure or amount)'[22]. Thus, the term ‗computational analysis' can be rephrased as data analysis by way of calculation. This means that a word ‗computational', strictly speaking, makes data analysis narrower: not all operations of the knowledge discovery process on datasets can be called calculations. Therefore, terms ‗data mining' and ‗data analysis' are better placed to speculate between various ways of the knowledge discovery.

Summarising the first paragraph, I believe that existing scientific and legislative interpretations of terms related to Big Data indicate that both scientists and law-makers are moving from buzzwords to legal concepts. In the upcoming discussion I would refer to Big Data as to large sets of data fixed in digital form which cannot be processed by conventional computing tools and to data mining as to the whole process of knowledge discovery on digital datasets. Terms ‗Big Data practices' and ‗data mining practices/activities' will be used as synonyms of ‗data mining'.

---

[19] See, p. 8 et seq. European Union. De Wolf & Partners. *Study of the legal framework of text and data mining (TDM).* 2014.

[20] P. 63. Borghi, M.; Karapapa, S. *Copyright and Mass Digitization: a Cross-Jurisdictional Perspective*, Oxford University Press, 2013.

[21] P. 9. European Union. De Wolf & Partners. *Study of the legal framework of text and data mining (TDM).* 2014.

[22] https://en.oxforddictionaries.com/definition/compute. Accessed on 12.02.2018.

2. BIG DATA LEGAL FRAMEWORK: LEGAL RISKS ASSOCIATED WITH THE USE OF DATASETS IN BIG DATA PROJECTS

After definitions relevant for the Big Data discussion were sorted out, it is necessary to introduce applicable legislation. As I mentioned, I approach Big Data focusing on such areas of IP law as copyright, database law and confidentiality. Accordingly, this paragraph is devoted to EU and US IP laws and structured as follows. First section introduces database and copyright laws under the EU legal regime. Second section discusses US copyright law and the US hot news doctrine which establishes limited ownership over some categories of factual information. Third section would elaborate on issues of ownership in data and access to data, as these concepts are highly controversial in the digital reality.

It is also necessary to note that this paragraph pursues the objective of the risk assessment under intellectual property laws from the perspective of market players involved in Big Data projects. Confidentiality, in turn, represents a unique body of law, and in the context of Big Data it is more feasible to speak about business strategies related to confidentiality rather than about the risk assessment. Due to these reasons, confidentiality would be dealt with separately, in the next paragraph.

2.1. Big Data Risk Assessment under the EU Legal Order

When important definitions are clarified it is feasible to start the analysis of the existing legal framework predetermining concerns of Big Data projects. The overall volume of Big Data, which grows continuously every minute and every second of digital world‘s life, still falls into very few categories of currently available IP protection. Under the EU legislation the EU database protection regime is highly relevant for the Big Data discussion. This regime was established under the Database Directive[23] introducing a unique mixture of copyright and *sui generis* protection for databases. The EU copyright law is also relevant for the Big Data discussion, as data mining activities frequently involve copying large volumes of copyrighted materials, such as articles, videos and sounds.

Accordingly, this section is structured as follows: first subsection discusses European database law, and second subsection examines general copyright law, specifically the InfoSoc Directive[24], assessing risks which could arise for data mining entities involved in Big Data projects.

[23] Hereinafter „Database Directive‘ refers to „Directive No. 96/9/EC of the European Parliament and of the Council, of 11 March 1996 on the legal protection of databases‘.
[24] Hereinafter „InfoSoc Directive‘ refers to „Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society‘.

2.1.1. Rights Enshrined in Big Data under the EU Database Law

Pursuant to Article 1(2) Database 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. As long as a dataset at issue falls in the scope of this definition, a database owner would enjoy either copyright or *sui generis* protection under the Directive. It is also possible that both copyright and *sui generis* regimes may overlap in the single database[25].

It is quite evident that in most instances datasets in question would qualify for the protection under the Database Directive. This is so, because data mining, in any case, may be performed only on materials which ‗arranged in a systematic or methodological way'[26]. However, legal certainty is implicated by the fact that single dataset in question is usually a composition of tens and hundreds of interrelated databases.

To provide simple illustration, let us assume that we need to assess IP rights persisting in a dataset[27] protected under the EU legal regime. Let us say, we deal with a dataset comprised from 100 databases. 80 of these databases are accessible as the part of Open Data projects or, otherwise, available for everyone's free use. Other 10 databases are not available due to confidentiality restrictions imposed by the owner of databases. However, remaining 10 databases are available but protected under the Database Directive. In practice, it may be problematic further distinguish between databases which protected by copyright and those which protected by *sui generis* regime. Therefore, I will consider these two mechanisms separately, addressing two main points: which rights relevant for the Big Data discussion are enshrined in databases at issue and to which extent it raises risks of a copyright infringement.

a) Copyright Protection of Databases

Article 3(1) of the Directive states that databases which, by reason of the selection or arrangement of their contents, constitute the author's own intellectual creation shall be protected as such by copyright. Article 3(2) further stipulates that the copyright protection of database does not cover its content and shall be without prejudice to any rights persisting in the materials comprising the database in question.

Furthermore, Article 5 of the Database Directive states that the author of a database shall have the exclusive right to carry out or to authorise:

---

[25] Article 7(4) Database Directive.

[26] Situations, when data mining concerns non-systemised datasets, e.g. raw data, are behind the scope of this paragraph. The use of non-systemised data, arguably, is predetermined by confidentiality arrangements.

[27] Oxford Dictionary defines ‗data set' as ‗a collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer'. Hereinafter I employ the term ‗dataset' referring to the collection of separate databases.

(a) temporary or permanent reproduction by any means and in any form, in whole or in part;

(b) translation, adaptation, arrangement and any other alteration; […]

(e) any reproduction, distribution, communication, display or performance to the public of the results of the acts referred to in (b).

These restricted acts are relevant for the Big Data discussion, while other acts listed in Article 5 do not implicate Big Data projects. The question arises: whether data mining practices shall be considered as alteration of selection and arrangement of a database within the meaning of Articles 3 and 5 of the Database Directive? I would argue that the answer is negative. While to the date, there is no straightforward clarification from the European Court of Justice (‗the ECJ' – hereinafter) which could help to answer the question, French court in *Dictionnaire Permanent des Conventiones Collectives* analysed issues concerning interpretation of Article 5 of the Directive.

The case was at length examined by P. Virtanen in his monograph devoted to the EU Database IP Law. Particularly, P. Virtanen stated that ‗copying parts of the content of a copyright database without copying the ‗structure' of the data, does not *ex facie* amount to copyright infringement, thus rendering the protection rather thin'[28]. Likewise, French court in *Dictionnarie* observed: ‗The ‗Dictionnaire' was a copyright database on account of its original presentation and the grouping of its headings. The respondent had yet copied to its own publication merely the contents [...] and no any original elements from the claimant's database'[29].

These interpretations of the copyright regime under the Directive, raised by the court and supported by the scholar, may justify an assumption that asserting copyright in a database against data mining entities is practically impossible: data mining is mainly concerned about data themselves and not about their selection and arrangement protected by copyright. Although there could be an act of temporary or permanent data copying before performing data mining activities, it would have been extremely hard to prove a copyright infringement.

For instance, B. Michaux opines: ‗The part of the database must be recognizable as such in the alleged [copyright] infringement. It can happen that the derived work shows differences with the original work so that the similarities are not identifiable anymore. More precisely, from the moment the differences are so that the global impression is not the same, there is no infringement, but independent creation'[30]. This statement describes exactly the situation of data mining: when numerous datasets are mined, and extracted information is combined and arranged

[28] p. 147, Virtanen, P. *Evolution, practice and theory of European database IP law.* Lappeenranta: Lappeenrannan Teknillinen Yliopisto, 2008.

[29] Tribunal de Grande Instance Lyon, 28th December 1998. *Dictionnaire Permanent des Conventiones Collectives.* English version of the case is sited as provided on p. 147, Ibid.

[30] P. 119. Michaux, M. *Droit des bases de données*. Kluwer, 2005. English translation is sited from p. 35, Ibid.

in the independent manner, the probability that copyright in the original selection and arrangement of a database in question would manifest itself in the results of data mining is extremely small.

Presumably, the speculative situation of the infringement, although exceptionally hard to prove, still could exist. If copyrighted selection and arrangement of a database somehow might be manifestly reflected in the results of data mining, and if simultaneously mined data might be aligned with copied part of a database in question, there is a copyright infringement. In other words, theoretical case could be:

- data miner employs 100 databases, as suggested above, selection and arrangement in 10 of which is protected by copyright;
- in the result of data mining, newly designed database manifestly coincides with selection and arrangement of a single database from the 10 protected;
- the contents of the infringed database also appear in a new database.

Having this said, one could argue that this delineated use of a single database, when there were 100 databases processed, analysed and restructured, should not be held for a copyright infringement. Might one not be acquainted with the Database Directive, he or she could argue that the portion reused without authorisation is extremely small. Nevertheless, under the Database Directive this argument could protect only reuse of a database for scientific research, with the vague requirement to use a copyrighted work ‗to the extent justified by the non-commercial purpose to be achieved'[31]. Hence, whenever small was the portion reused for commercial purposes, a question of a copyright infringement could arise if illustrative criteria suggested above are met.

The main conclusion to draw, so far, is that Big Data assets partially are protected under the copyright regime of the Database Directive. This protection does not contain necessary balancing exemptions which could foster commercial reuse of Big Data. Although with respect to the quantity an infringement seems to be immaterial, plain interpretation of the Database Directive raises legal risks for entities interested in the commercial exploitation of Big Data.

b) Sui Generis Protection of Databases in the EU

While an assertion that Big Data practices might infringe copyright in databases under the EU law seems to be speculative one, this is not the case with the *sui generis* regime under the Database Directive. As I would further demonstrate, *sui generis* right provides more aggressive protection scheme for database owners, which, accordingly, is more dangerous from the perspective of data mining entities.

[31] Article 6(2)(b) Database Directive.

According to Article 7(1) of the Database Directive the maker of a database which shows that there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents shall have a right to prevent extraction and/or re-utilisation of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database.

This is not an objective of my research to clarify the legal test which is to be complied with in order to trigger protection under Article 7(1). Therefore, I start my analysis from the assumption that a database in question is protected under the *sui generis* regime. Anyhow, managing large sets of data collected in the modern digital environment does require a substantial investment either qualitatively or quantitatively or both. Keeping this in mind, we should focus on restricted acts which are to be exclusively authorised by the database maker under Article 7 of the Directive: extraction and re-utilisation.

aa) Extraction as a Restricted Act under the Sui Generis Regime

Under Article 7(2) ‗extraction' shall mean 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. This definition is so inclusively construed that it leaves almost no room for the doubt that data mining practices include extraction in the sense of Article 7(2). It was noted by several scholars that ‗data analytics or data mining will often involve the wholesale copying of information or databases'[32]. Indeed, what data miners frequently do: they extract either a portion or an entire database and transmit it to their own medium or medium which they technically or contractually control.

Most likely, data miners would not transfer data on the permanent basis, because a single database is interesting for data mining entities only as a small fraction of a larger dataset. At the same time, temporary transfer of data does occur, and, consequently, the whole extraction process shall be authorised by a holder of the *sui generis* right.

If it might happen that under specific circumstances data mining practices do not count for an ‗extraction' within the meaning of Article 7(2) of the Database Directive, it is further necessary to examine second restricted act under the EU *sui generis* regime.

bb) Re-utilisation as a Restricted Act under the Sui Generis Regime

Under the Database Directive a term ‗re-utilisation' shall mean any form of making available to the public all or a substantial part of the contents of a database by the distribution of copies, by renting, by on-line or other forms of transmission. This term has two problematic elements, which should be analysed to assess interplay between rights of a database maker and

[32] P. 217, Ursic, Helena; Custers, Bart. *Legal Barriers and Enablers to Big Data Reuse*. 2 Eur. Data Prot. L. Rev. 209, 221 (2016). See also references provided therein.

of entities practising data mining. These two elements are (1) the action of making available to the public and (2) the definition of a substantial part of the contents.

Firstly, what can be considered as the making available to the public in the Big Data framework? There are so many variations of Big Data practices that, depending on the circumstances of the case, making a database available can be done in numerous ways. Article 7(2) of the Database Directive, to the general satisfaction of the EU database makers does not provide a closed list concerning forms of making available to the public. Therefore, as long as an entire content or a substantial part of a database in question can be accessed by the general public by any means, there is a copyright infringement under the *sui generis* regime.

Secondly, it is problematic to determine when making available to the public of a part of a database should be considered substantial? The legal clarification of the *sui generis* right makes analysis even more complicated, as, pursuant to Article 7(1), an author has an exclusive right to prevent re-utilisation „of a substantial part, evaluated qualitatively and/or quantitatively". Fortunately, the ECJ clarified wording of Article 7(1) in *The British Horseracing Board and Others*[33] (hereinafter – *Horseracing*). Following binding interpretations of the ECJ in *Horseracing* are relevant here:

- the expression „substantial part, evaluated quantitatively" refers to the volume of data extracted from the database and/or re-utilised and must be assessed in relation to the total volume of the contents of the database;

- the expression „substantial part, evaluated qualitatively" refers to the scale of the investment in the obtaining, verification or presentation of the contents extracted and/or re-utilised, regardless of whether the contents at issue represent a quantitatively substantial part of the general contents of the protected database[34].

Interpretation of the ECJ related to the quantitative criterion of illegally extracted/re-utilised part of a database seems to be somehow self-evident. This is not entirely helpful for the risk-aware decision-making to know that quantitatively a substantial part depends on the whole volume concerned. However, one could not reasonably expect from the ECJ more straight-forward guidelines in this respect: the quantitative criterion by its very nature is circumstantial.

In turn, ECJ's interpretation of the qualitative criterion is practically more helpful. The qualitative substantiality of a part of a database is linked with the corresponding investment of a database maker. Hence, when it comes to the quality of illegally extracted/re-utilised part of a database, this gives to a database maker an idea how to fulfil its burden of proof: to submit the evidence of the investment. At the same time, the ECJ clarifies that the qualitative criterion

[33] *The British Horseracing Board and Others.* The ECJ. C-203/02. 09.11.2004.
[34] Ibid, Ruling 3.

should be considered in isolation from the quantitative. This also makes life of a database maker much easier, as long as it can demonstrate substantial investment in creating even quantitatively insubstantial parts of a database in question.

From the perspective of entities involved in Big Data projects both actual wording of Article 7(2)(b) of the Database Directive and its interpretation in *Horseracing* are problematic. While it is less likely that data mining entities could extract and re-utilise substantial parts of a database in the sense of the volume, the risk of an infringement substantiates itself when it comes to the qualitative criterion. Indeed, data mining entities, while processing huge amounts of data, most likely, have no clue whether some insubstantial parts of database's contents correspond to a substantial investment of a database maker. This would mean that the safest preventive strategy for data mining entities is to negotiate a licence from a right holder. Although the safest, this option can be not cost-efficient. The case could be that it is complicated to define a right holder, costly to approach it or that it simply has no interest in granting a licence.

Before summing up the discussion about the *sui generis* regime of the European database law, let us reveal another possible legal argument in favour of Big Data practices, as per *Horseracing*. The ECJ has ruled that the Database Directive prohibits unauthorised acts of extraction or re-utilisation the cumulative effect of which is to reconstitute and/or make available to the public the whole or a substantial part of the contents of that database *and thereby seriously prejudice the investment by the maker*[35].

At the first glance, it seems that this statement adds nothing new to the one discussed above. However, the ECJ introduces a peculiar requirement of a serious prejudice to the investment. One may ask why the ECJ needs to introduce this wording if it was already established, what constitutes substantial extraction or re-utilisation? A link with the investment concerned was also clarified. If one is searching for the argument in favour of data mining practices, reasoning may be as follows.

It is internationally recognised principle of legal interpretation that the provision shall be kept effective rather than null and void[36]. Hence, the additional phrase in the statement of the ECJ should bear distinctive meaning. Consequently, causing a serious prejudice to the investment is distinct from unauthorised actions of extraction and re-utilisation themselves. What is the difference? I would argue that a database maker should prove separately that it was hindered from deriving benefits of its investment due to unauthorised acts. It could be problematic for a database maker when a data mining entity is not its rival.

[35] Ruling 4, The ECJ. *The British Horseracing Board and Others.* C-203/02. 09.11.2004. Emphasis added.
[36] See, p. 683 Born, Gary. *International commercial arbitration, Volume I-II.* Kluwer Law International, Alphen aan den Rijn, 2009*;* See also, p. 460*,* Kröll, Stefan et al. *Comparative International Commercial Arbitration.* Kluwer Law International, 2003.

Overall argument based on the interpretation of 4th ECJ ruling in *Horseracing* seems to be intricate one. My purpose here, therefore, is not to argue that a legal protection of Big Data practices in Europe exists but instead to highlight how legally uncertain situation is.

All in all, European regime of database protection is highly controversial for Big Data innovations. Scientific utilisation of Big Data might be arguably exempted from liability under the Database Directive, but commercial practices have no chance. Anyhow, exceptions and limitations under the European database law would be examined in detail later in this paper.

2.1.2. Data Mining and Reproduction Right under the InfoSoc Directive

Costs of negotiating a licence from every single author of hundreds and thousands of literary and artistic works[37] would be prohibitive for data mining entities. What is more, it is less likely that data mining practices prejudice effective economic exploitation of a work by a single author concerned[38]. However, authors' right to exclude others from the use of their works is absolute and does not require any economic justification. Bearing this in mind, let us consider how the general European copyright law enshrined in the InfoSoc Directive affects Big Data projects.

Article 2 of the InfoSoc Directive stipulates that member states shall provide for copyright holders the exclusive right to authorise or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part of protected literary and artistic works. Recital 21 of the Directive farther clarifies that the reproduction right shall be construed broadly to ensure legal certainty within the internal market. Therefore, the wording of the Directive implies that data mining should not be performed without an explicit authorisation by the copyright holder, insofar as data mining entities directly or indirectly, temporary or permanently reproduce, i.e. copy, the work.

The authors of the EU study on text and data mining, analysing various commentaries, make conclusion that data mining _generally involves an act of copying' of data being processed for the analytical purposes[39]. However, the authors further opine that in cases when software only _crawls' through texts or other information sources, counting, for example, repetition of the word _digital', no copying within the meaning of the InfoSoc Directive takes place[40].

I would argue that the whole idea of rendering data mining restricted act, arguably, conflicts with objective values behind copyright law. The following statement of the ECJ helps elaborate further on this:

---

[37] Hereinafter, employing the term _work', I refer to any copyrighted subject matter under the InfoSoc Directive.
[38] This notion might not be final for Big Data scrapping techniques performed on news media databases. See for that matter, Section 4.2 of the paper analysing the US fair use doctrine.
[39] P. 31, European Union. De Wolf & Partners. *Study of the legal framework of text and data mining (TDM).* 2014.
[40] Ibid.

> An act occurring during a data capture process, which consists of storing an extract of a protected work comprising 11 words and printing out that extract, is such as to come within the concept of reproduction in part within the meaning of Article 2 of [the InfoSoc Directive], if the elements thus reproduced are the expression of the intellectual creation of their author[41].

Not the notion that even the modest amount of words can be protected by copyright is interesting here, but a bold reminder that copyright concerns ‗the expression of the intellectual creation' of an author. Thus, the very basic dichotomy of the copyright law – idea v. expression – shall be heavily relied on by advocates of Big Data innovations.

Indeed, in this subsection I discuss copyright in artistic and literary works, which should be clearly distinguished from both investment in a database creation and copyright in selection and arrangement, protected under the EU Database Directive. This is true that data mining as such frequently involves the wholesale copying of data and currently this act would constitute an infringement of copyright under the InfoSoc Directive. However, when construing the law, one should keep in mind the economic purpose behind it. The purpose of copyright is protection of the original expression and focus of data mining is capturing facts. To put it simply, the EU legislator does not need to exempt data mining entities from liability, but only needs to explicitly state that their activities are behind the reach of copyright holders' monopoly.

There is little to add with respect to reproduction right and data mining on this stage of the discussion. Thus, I would revisit the ‗copyright vs. Big Data' problem later, on pages of this paper devoted to the examination of general copyright limitations and specific data mining exceptions.

2.2. Deriving Benefits of Big Data in the US: Copyright Law and the Hot News Doctrine

The US legal practice is intertwined with economic analysis. Furthermore, although one could argue that the US legal order is presently much closer to the continental statutory system than to the common law tradition of court precedents, US laws still rely heavily on the legislative discretion of judges. The subsequent analysis in this section would support this statement.

There is no database law in the US akin to one enacted within the EU. However, compilations and collections of works are protected by the US copyright law, granting, thus, similar protection to the selection and arrangement of databases' contents. To provide comparative analysis with the EU legal order, this section is structured as follows. The first part introduces the US copyright law relevant to the Big Data discussion, the second part represents case study concerning the hot news doctrine. This doctrine, as the study would demonstrate, has become relevant to Big Data projects under the US legal order.

[41] The ECJ, *Infopaq International A/S v Danske Dagblades Forening*, 16 July 2009, Case C-5/08.

2.2.1. Big Data and the US Copyright Law

Title 17 of the United States Code enshrines uniform copyright law applicable across all the United States. §§ 101, 103 U.S.C. interpreted together establish legal framework for protection of databases. § 101 U.S.C. provides a following definition relevant to the database legal regime:

> A ‗compilation' is a work formed by the collection and assembling of pre-existing materials or of data that are selected, coordinated, or arranged in such a way that the resulting work as a whole constitutes an original work of authorship. The term ‗compilation' includes collective works.

§ 103 U.S.C., in turn, elaborates on this definition stating that ‗the copyright in a compilation extends only to the material contributed by the author of such work, as distinguished from the pre-existing material employed in the work, and does not imply any exclusive right in the pre-existing material'.

Interestingly, the US legislator does not use the word ‗database' and simply refers to a compilation in more general sense. To establish the relevance of cited articles to the database protection one needs to make an inquiry in the US case law. While it could be the subject of a separate study to trace back all relevant developments of the US case law with respect to the database legal protection, below I would briefly examine cases establishing modern legal trends.

The judgement of the US Supreme Court in *Feist* rendered in 1991 has become a touchstone of the modern US database protection. In *Feist* the US Supreme Court heavily criticized the ‗sweat of the brow' doctrine. The case law preceding the ruling in *Feist* has established this doctrine – the alternative name is ‗industrious collection' – with the main objective that the one's hard work shall be rewarded by copyright law. For instance, in *Jeweler's Circular Publishing Co*, adjudicated back in 1922, the Court of Appeal for Second Circuit stated: ‗The man who goes through the streets of a town and puts down the names of each of the inhabitants, with their occupations and their street number, acquires material of which he is the author'[42]. Revising erroneous reasoning of preceding cases, the US Supreme Court maintains in *Feist* that the US copyright law ‗leave no doubt that originality, not ‗sweat of the brow', is the touchstone of copyright protection in directories and other fact-based works'[43].

In the nutshell, judgement in *Feist* demonstrates that, contrary to the EU legislative approach, the US rejects protection of the investment in the work's creation under IP law. Instead, the US legislation recognises copyright only in selection and arrangement of a database. I believe that such approach is justified by the strict adherence of the US Congress to the wording of the US Constitution, which has empowered the Congress ‗to promote the

[42] *Jeweler's Circular Pub. Co. v. Keystone Pub. Co*., 281 F. 83, 88 (C.A.2 1922).
[43] *Feist Publications, Inc. v. Rural Telephone Service Co.*, Inc., 111 S.Ct. 1282, 1295, 499 U.S. 340, 359–60 (U.S.Kan.,1991).

Progress of Science and *useful Arts*, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries'[44]. Indeed, while the investment in the work's creation could cause production of a useful subject matter, this investment alone would never overcome the originality threshold.

Following citations can support such interpretation of the US Constitution. The US scholar Xuqiong Wu back in 2002 has stated that any attempts to model the US legislation in the way similar to the EU Database Directive _may not survive the scrutiny under the US Constitution'[45]. Furthermore, Marshall Leaffer in his article published in 2016 criticises any lobbying for the introduction of the database *sui generis* right to the US legal order. He has stated: _Very little empirical research demonstrating that lack of protection of non-original databases has undermined optimal incentives for their creation is ambiguous at best. […] So far the benefits of database protection are exceeded by their costs. The United States not long ago avoided a European style sui generis law through legislative gridlock. Three cheers for legislative gridlock'[46].

The general introduction of the US copyright law with respect to the database protection already demonstrates that the US legal order is more suitable for Big Data projects. There are no risks under US IP law akin to those under the EU *sui generis* database regime. Risks associated with copyright in selection and arrangement would coincide with risks scrutinised earlier, with the reservation that these risks are always dependent on the threshold of the originality in the jurisdiction at issue.

While the general introduction of the US database law eliminates many concerns for data mining entities, when it comes to the exploitation of databases protected under the US law, issues of copyright persisting in artistic and literary works as such still are to be dealt with.

§ 102 U.S.C. states that copyright protection subsists in original works of authorship fixed in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced, or otherwise communicated. Further the paragraph cites categories of works which enjoy copyright protection. However, as previously in the discussion concerning the EU copyright law, not the general provisions establishing basis for the copyright protection are interesting for the present discussion, but rather legal mechanisms envisaging boundaries of the protection. In the US the main mechanism limiting copyright is the fair use doctrine, and it would be in focus of my research in fourth paragraph of the paper.

---

[44] *Emphasis added*. Section 8, Clause 8. Patents and Copyrights, U.S.C.A. CONST Art. I § 8, cl. 8.

[45] P. 587. Xuqiong Wu, *E.C. Data Base Directive*. 17 Berkeley Tech. L.J. 571 (2002). Available at: http://scholarship.law.berkeley.edu/btlj/vol17/iss1/33.

[46] Pp. 860-861. Leaffer, Marshall. *Database Protection in the United States is Alive and Well: Comments on Davison*. 57 Cas. W. Res. L. Rev. 855 (2007).
Available at: http://scholarlycommons.law.case.edu/caselrev/vol57/iss4/10.

At this stage another peculiar legal mechanism is to be dealt with: the US hot news doctrine. While selection and arrangement of a database can benefit from the copyright protection under the US legal order, such protection – similarly with copyright under the EU database law – would be rather thin. More interestingly, the US hot news doctrine can be viewed as a modicum of the European *sui generis* right in databases. This mechanism, strictly speaking, does not fit nicely into IP law but is highly relevant for the Big Data discussion. Consequently, the hot news doctrine is discussed in the next subsection.

2.2.2. Big Data and the US Hot News Doctrine: Battle over Ownership in Numbers

This subsection concerns utilisation of quantitative data: ideas, facts, numbers as opposed to original expression enshrined in literary and artistic works. Quantitative data at issue is referred to within the US as „hot news", due to its instant economic value. US courts found it fair that entities investing in gathering and/or generating some specific types of information can enjoy limited monopoly over re-utilisation of this information. In other words, the hot news doctrine to the certain extent serves as a substitute for the EU *sui generis* right, when it comes to the protection of news media compilations. It is worth mentioning that the hot news doctrine was developed in the US already in the beginning of the 20th century, much earlier than the EU database law.

a) The Hot News Doctrine and Data Mining

The hot news doctrine under the US law belongs to realm of unfair competition law and represents „state law claim that protects the ownership of discrete facts for a short period of time after publication"[47]. Although, at the first glance, the doctrine lies outside the boundaries of the present research, existing scientific debate over doctrine's future in the digital reality makes it highly relevant for the current discussion.

The hot news doctrine was introduced by the Supreme Court in *International News Service v. Associated Press*[48] already in 1918. The Court granted news providers limited property rights in published „hot news". No generally applicable time frame was defined during which the right can be exercised. However, the Court stated that the right is to be enjoyed „only to the extent necessary to prevent that competitor from reaping the fruits"[49].

As authors of the article „*From Hot News to Hot Data…*" have noted, the hot news doctrine was consequently invoked by plaintiffs for protection of financial data. As highlighted in the article, there is a serious debate over modern developments of the doctrine. I would refer

[47] P. 305, Ekstrand, Victoria Smith; Roush, Chirstopher. *From Hot News to Hot Data: The Rise of Fintech, the Ownership of Big Data, and the Future of the Hot News Doctrine*, 35 Cardozo Arts & Ent. L.J. 303, 340 (2017).
[48] *International News Service v. Associated Press*, 248 U.S. 215 (1918).
[49] Ibid, at 241.

to this discussion later in my research, but now let us examine the existing US legal test under the doctrine. The Court of Appeal for Second Circuit in *National Basketball Association v. Motorola, Inc.* stipulated new test under the hot news doctrine which includes following steps:

(i) a plaintiff generates or gathers information at a cost;

(ii) the information is time-sensitive;

(iii) a defendant's use of the information constitutes free-riding on the plaintiff's efforts;

(iv) the defendant is in direct competition with a product or service offered by the plaintiffs;

(v) the ability of other parties to free-ride on the efforts of the plaintiff or others would so reduce the incentive to produce the product or service that its existence or quality would be substantially threatened[50].

This test seems to be quite straightforward, aiming to protect entities, which gather or generate information at a cost, exclusively from their direct competitors. It is worth mentioning that the Court of Appeal has not considered value of the information as a prerequisite of protection. Indeed, it goes without saying that data in question by default should be deemed valuable if competitors attempt to reuse it shortly after publication.

Following observations can be made if one considers the legal standard under the hot news doctrine from the data mining perspective. Data mining entities should not compete directly with producers of hot information which they gather. If they do compete directly with information producers, they would need to prove that other steps of the test above are not manifested in their actions. I believe that fifth step is the most problematic one in the Big Data settings. This step goes beyond the classic consideration of substitution products. Indeed, it should be easier for a plaintiff to prove that an online database, for example, just constitutes a substitution market for the product or service existing in the material world. Nevertheless, according to *National Basketball Association v. Motorola, Inc.*, a plaintiff – a database producer in Big Data settings – shall also demonstrate that either product‘s existence or quality is endangered.

One could still argue that in many situations online databases generated as a result of data mining practices create services of a higher quality than those, provided by traditional service-providers outside of the Big Data framework. In any case, in the situation of hot news‘ reuse, allocation of burden of proof is less favourable for data mining entities than in the situation giving raise to the fair use defence, as I will demonstrate later on pages of this paper. The US case law concerning the hot use doctrine in many instances is more reliable for hot information producers[51]. Such approach seems to be reasonable. The rationale of the hot news doctrine is to

[50] *Nat'l Basketball Ass'n v. Motorola, Inc.*, 105 F.3d 841 (2d Cir. 1997).
[51] See pp. 306-316, Ekstrand, Victoria Smith; Roush, Chirstopher. *From Hot News to Hot Data.*

protect instant value of the information against unfair competition. Honest business practices of data mining entities, anyhow, should be conducted in a way which does not raise issues of unfair competition. Bearing this in mind, I would argue that the hot news doctrine does not complicate existence of Big Data projects as long as such projects do not deprive information producers from the instant value of their ‗hot' information.

b) The US Hot News Doctrine and the EU Database Directive

The EU Database Directive has established the *sui generis* regime protecting investments in either the obtaining, verification or presentation of database's contents. Article 10 of the Database Directive stipulates that the term of protection of the database is 15 years from the date of either its creation or its making available to the public. Moreover, this term can be renewed pursuant substantial qualitative or quantitative changes of a database. Thus, the EU database producers can enjoy almost perpetual protection against misappropriation of databases' contents if they timely represent changes into either structure of a database or into its contents.

The hot news doctrine can be viewed as a modicum of the EU *sui generis* protection model. It grants short-term protection for initially published information, as opposed to the almost perpetual term under the Database Directive. The US doctrine protects only against misappropriation of narrowly defined categories of data, as opposed to the overarching approach of the EU *sui generis* regime. Indeed, the article ‗*From Hot News to Hot Data*…' provides that some commentators ‗seek solutions [suggesting federalising] the doctrine, as the European Union has essentially done'[52], in support of the statement that the EU Directive is an extreme version of the US legal model protecting facts and ideas against misappropriation in narrowly defined circumstances.

This brief comparison clarifies once again that the US legislation views exceptionally wide monopoly of database producers to be dangerous for the market. In my opinion, this is the wise approach perfectly suiting needs of the digital era.

The article which I cited above discusses at length possible developments of the hot news doctrine. Its authors opine that in Big Data settings the hot news doctrine evolved in the hot data doctrine. I believe that the doctrine in its existing shape does not underpin Big Data developments. It might only be argued that fifth step under *National Basketball Association v. Motorola, Inc.* needs clarification to establish more legal certainty. At the same time, authors Ekstrand V. and Roush C. provide ironical notion which points in the direction of costs and benefits considerations:

[52] Ibid, p. 316.

> Given that Wall Street has the money to fight, the battle over who will own the discrete numbers that fuel financial markets now and in the future is likely to be driven by the legacy firms. The public is not likely to be included in that conversation, but the *public interest in that conversation is significant*[53].

The notion „public interest", from the economic standpoint, can be understood as the rationale of social costs and benefits. This subsection has demonstrated how courts accounted public interest with respect to the hot news doctrine. Quite evidently, this rationale should be also applied when balancing social costs and benefits of Big Data projects.

2.3. Ownership in Data and Access to Data: Implications for Big Data Projects

An issue of the ownership in data is a topic of a controversial discussion in scientific circles. Consequently, there is no conventional approach to the issue. For instance, authors of the article about assessment of legal risks in Big Data and the cloud coherently employ the term with quotation marks[54], implying that, in authors' opinion, the ownership in data as a legal concept does not exist. Such approach can be seen as a reasonable one due to the very nature of „ownership" as a legal concept. In the nutshell, the ownership title, or the property right, entitles an owner to exclude others from deriving benefits connected with the use of his/her property. Historically, this entitlement was rooted in the rivalrous nature of tangible goods. Quite recently, due to development of the information society, people started to consider information, knowledge and data as valuable assets, which, at the same time, do not possess rivalrous characteristics of tangible goods. Hence, non-rivalrous nature of data could serve as one possible ground to reject existence of ownership rights in data. However, controversies or, one could even say, struggles with relations between ownership and data in the literature require further clarifications.

This section is structured as follows. The ownership in data is analysed as a general legal concept in the first subsection. Then in the second subsection ownership in data is more narrowly defined from the focus of IP law. Finally, the third subsection analyses an alternative, or related, concept of the access to data which can serve as a useful risk assessment tool for data mining entities.

2.3.1. General Concept of Ownership in Data

When dealing with Big Data one should keep in mind that Big Data projects are rarely linked to the single jurisdiction but rather involve either collaboration between the parties established in different countries or gathering of data which are generated within various jurisdictions. This raises the question whether searching a unified legal definition of the

[53] Ibid, p. 339 (emphasis added).
[54] See, Corrales M.; Djemame K. *A Brokering Framework for Assessing Legal Risks in Big Data and the Cloud*. 2017.

ownership in data is a feasible way to foster Big Data transactions? Bundle of rights and obligations stemming from the concept of ownership is not harmonised across various jurisdictions, due to the simple fact that ownership is one of the most ancient economic and legal concepts developed by different legal cultures[55]. Therefore, any scientific endeavours to elaborate potentially suitable definition of the ownership in data would encounter numerous obstacles.

Furthermore, even if one might develop a coherent concept of the ownership in data from the standpoint of international private law, it still could take an eternity for politicians and legislators to implement such revolutionary concept in the harmonized fashion, because such implementation would require an exceptionally high threshold of a political consensus. Thus, I believe that one should treat cautiously any attempts to regulate economic relationship between producers of data and ‗their' data applying a general notion of ownership.

Let us consider how ideas of the ownership in data manifest themselves. At least, following categories of data clearly reflect issues of the ownership: (1) personal data, (2) raw data which nevertheless promises a hidden value and (3) commercially valuable confidential data.

First, although issues of processing personal data are beyound the reach of my research, it is worth mentioning that scholars and legislators recognise how personal data acquire proprietary characteristics, or monetary value, in the framework of digital transactions[56]. Indeed, scientific papers discuss that personal data should be considered as means of payment for some ‗free' digital services, and no one could reasonably challenge the fact that ownership determines legal connection between a person and his/her means of payment.

Second category, raw data, usually mediates all processes of digital communication. With respect to this category the following statement is of an interest: ‗Legal systems generally differentiate between raw data and databases. Raw data refers to basic, unprocessed data, such as internet traffic. Generally, raw data, including private data, are not seen as owned by anyone'[57]. In my opinion, the notion that raw data include personal data is not wholly consistent. However, what is more important: whether raw data indeed are not owned by anyone? Quite evidently, no legal title could be derived from legislation, to be attached to those, who, for instance, generate traffic as one of the forms of raw data. Nevertheless, this does not imply that any person upon recognising commercial value of raw data still would abandon it for the public domain. To the

[55] See, e.g. on history of common law property rights, p. 6 et seq. Siegan, Bernard H. *Property Rights: From Magna Carta to the Fourteenth Amendment*. Social Philosophy and Policy Foundation, Transaction Publishers. 2001.
[56] For the discussion of personal data as new currency see e.g., p. 5, Chirita, Anca D., *The Rise of Big Data and the Loss of Privacy*. June 15, 2016. Available at SSRN: https://ssrn.com/abstract=2795992.
[57] P. 362, Rubinfeld, Daniel L.; Gal, Michal S. *Access Barriers to Big Data*. 59 Ariz. L. Rev. 339, 382 (2017).

contrary, in absence of statutory provisions private persons envisage contractual terms to fill caveats. Thus, it would be reasonable for data producers (controllers) to impose the confidentiality regime on sets of raw data. Unfortunately, contractual provisions do not eliminate legal uncertainty concerning official definition of raw data. Whether such data are analogous to facts and ideas and, therefore, falling within the public domain, or, to the contrary, should it be deemed as the analogue of commercially valuable information, enjoying trade secrecy protection? These questions have no clear answer, once again leaving the term –ownership in data" not clarified.

Third, commercially valuable confidential data, although already were in the focus before, deserve additional commentaries in relation to the concept of ownership. Legislation on confidentiality lacks any substantial harmonisation, contrary to many other areas of IP protection which were harmonised by the Bern Convention and other international treaties and conventions[58]. Accordingly, distinctive legal doctrines exist in the framework of confidentiality/trade secrecy. The US case law has established that trade secrets are proprietary in nature. For instance, in *Ruckelshaus* v. *Monsanto Co.* the US Supreme Court states: ‗The right to exclude others is generally one of the most essential sticks in the bundle of rights that are commonly characterized as property. With respect to a trade secret, the right to exclude others is central to the very definition of the property interest'[59]. In turn, Marius Petroiu in his master thesis observes that there are more remedies available for the trade secret holder under the proprietary doctrine than under contractual or breach of confidentiality doctrines[60]. Two latter prevail, for example, in the United Kingdom and Canada.

In the nutshell, existence of the proprietary interest in the commercially valuable information does not automatically justify ownership in data as such. In any case, the confidentiality regime provides more modest protection than that enjoyed under the classic ownership theory. Let alone that the confidentiality protection is not unconditional but exists only in the case of trade secret holder's strict adherence to the specific conduct: preservation of a secret, imposition of protective measures, etc. Furthermore, the fact that some developed legal orders do not treat trade secrets as having proprietary nature makes feasible the following statement: if there is no proprietary, ownership-like, protection for confidential information, it would be even less reasonable to grant such protection for data, in general.

---

[58] E.g., the WIPO Copyright Treaty, TRIPS. Comprehensive list of international treaties and conventions on IP is available at: http://www.wipo.int/export/sites/www/about-ip/en/iprm/pdf/ch5.pdf.

[59] *Ruckelshaus v. Monsanto Co.,* 104 S.Ct. 2862, 2877, 467 U.S. 986, 1011 (U.S.,1984). For the comprehensive overview of the US case law, see: pp. 24-28, Marius, Petroiu. *Forms of trade secret protection: a comparative analysis of the United States, Canada, the European Union and Romani*. Master thesis. 2005.

[60] P. 28. *Forms of trade secret protection*. Master thesis. 2005.

All in all, I believe that introduction of the ‗ownership in data' concept into legislation is not only impractical but also dangerous. This is so, because a property right vests on its holder a thick bundle of entitlements, initially designed to establish relationship between a person and tangible goods. Numerous reservations were made when the notion of ownership was borrowed by law-makers to deal with issues of intellectual property. Even more cautious approach was taken – the confidentiality regime across the globe or the hot news doctrine in the US – to deal with appropriation of the information. Logical conclusion to draw from these observations is that protection of proprietary rights in data cannot be stronger than this granted to owners of the secret information. Hence, from my point of view, one should treat the notion of ownership in data sceptically.

2.3.2. Ownership in Data from the Focus of IP Law

Previous subsection has scrutinised ownership in data as such, focusing on some main categories of data: private, raw and confidential. To finalise the scrutiny it is necessary to focus on the concept of ownership purely from the perspective of IP law. Accordingly, this subsection briefly addresses three most relevant scenarios under IP law: (a) ownership in data under database law, (b) under the US hot news doctrine and (c) ownership in data under confidentiality.

Let us revisit the notion of the Database Directive which provides exclusive *sui generis* right for EU database producers. According to Article 7(1) of the Database Directive the maker of a database which shows that there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents *shall have a right to prevent extraction and/or re-utilisation of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database*.

Now, applying the concept of ownership to Article 7(1) points out that ownership in data does exist. However, one should remember that ownership in data under *sui generis* right arises not upon creation of data but upon creation of a database – shall we not focus on all other important conditions under Article 7(1) of the Database Directive.

Grounds of protection are distinct for ‗hot data' under the US hot news doctrine. I am not going to reproduce all conditions of the doctrine upon fulfilment of which the ownership is granted. Relevant conditions can be revisited above[61]. On this stage it is important to note that ownership in ‗hot data' arises upon actual creation of data, protecting ‗hot data' producers against its instant misappropriation.

One should acknowledge that ownership granted under both EU database law and the US hot news doctrine is narrowly defined in comparison with the general concept of ownership. One

[61] See, supra Subsection 2.2.2.

may imagine such process of defining level of necessary protection as if data went through a bottleneck, which was designed to limit proprietary rights in data. We gather vast amounts of data and measure it against the ownership concept. In majority of cases granting of ownership in data would be denied[62]. Nevertheless, conditional ownership exists if data falls into one of following categories: (a) data as part of EU databases protected by sui generis right; (b) hot data under the US hot news doctrine; (c) confidential data.

In all three scenarios ownership is much more limited than its general notion. This is important to remember. The ownership in data as such – by way of comparison with the ownership in land – does not exist. The ownership in data is conditioned upon many requirements which shall be fulfilled by data producers. And even after these requirements were fulfilled, the protection granted is rather limited in comparison with the ownership, in general, or with the ownership in traditional categories of IP – e.g., copyrights, patents, trade marks. To point out another implication, ownership in data is not a unified concept, and specific entitlements of a data producer depend on the category of data – contents of EU database, hot or confidential information.

2.3.3. Access to Data: Issues of Risk Assessment

When the concept of ownership in data was clarified from the perspectives of law, in general, and IP law, specifically, we can evaluate another important concept: the access to data. This concept appears to be more suitable for the risk assessment in the course of Big Data transactions than the controversial notion of the ownership in data, and this subsection would demonstrate, why.

The basic idea behind the word „access" itself is not legal, but rather conventional one. According to the Oxford dictionary, in general, „access" is determined as „the means or opportunity to approach or enter a place"[63]. Moreover, the Oxford dictionary coins definition of „access" specific to the use of a computer: „the process of obtaining or retrieving information stored in a computer's memory"[64]. Therefore, there is hardly any room for an argument with respect to the plain meaning of the collocation „access to data".

In the Big Data framework opportunities of data mining users are largely pre-determined by the availability of data. The easier from both technical and legal points is access to data, the lower costs of realising Big Data projects. The main criterion applied by IP law to deal with the question of access to a work is the notion of a lawful use. Recital 33 of the InfoSoc Directive explains that „a use should be considered lawful where it is authorised by the right holder or not

[62] Let us disregard for a moment EU *sui generis* protection. Moreover, on the worldwide scale EU databases are only a drop in the sea of information.
[63] Definition available at: https://en.oxforddictionaries.com/definition/access. Accessed on 23.01.2018.
[64] Ibid.

restricted by law‘. Arguably, the conjunction ‗or‘ in this sentence is not precise, and an alternative ‗and/or‘ could better reflect relationship between contractual authorisations and legal stipulations, because latter can be either default or mandatory by nature. More important for the current discussion is who qualifies for the ‗lawful use‘ status. Useful explanation related to the use of databases is suggested by Professor Estelle Derclaye:

> ‗The interpretation of a lawful user as a lawful acquirer leaves us with very few unlawful users. The majority of users will be lawful. [...] Unlawful users will be those stealing a database incorporated in a tangible medium, those subsequently acquiring a stolen or an infringing database (a pirate copy). It could also be the persons using someone‘s password to access a database or using, knowingly or not, a website pirating a protected database‘[65].

Indeed, if one equates a lawful use with the lawful acquisition of a single work, there would be no issues with defining a status of a user concerned: there should be a licence, contractual or statutory one, to justify the use. If there is none, the use is unlawful. However, the data mining scenario involves, restating the obvious, ‗acquisition‘ of large datasets and there can be numerous right holders enjoying various IP rights in these data. Hence, more detailed categorisation of these data would be necessary for the objective risk assessment.

The EU study on text and data mining provides highly valuable classification on which I rely below[66]. The authors define four possible levels of granting access to data: (1) by all to all, (2) by many to many, (3) by one to many, (4) by one to one. This classification, as it is quite straightforward, appears to be very helpful for the risk assessment preceding actual data mining.

Access granted (1) by all to all refers, in general, to the entire data published on the web in the freely accessible manner. Web data, when defined as freely accessible, constitute a subject matter beyond the scope of the present research. However, probable rare situations may exist, when data are freely accessible, but data mining activities are restricted contractually or technically. In any case, defining existing restrictions of data mining is already the next step to examine. Furthermore, such restrictions are dictated by the will of private parties and usually are imposed contractually. In absence of clear legislative stipulations there is hardly any need to examine these freely exercisable acts of will. Thus, on this stage, it is important to establish that the access as such is lawfully provided.

Access granted (2) by many to many refers to contractual settings existing in social networks. For instance, a user may set accessibility of data published on his or her Facebook profile in a way which suits better his/her requirements of privacy or other individual interests. These settings will contractually oblige Facebook to set up access settings correspondingly with either user‘s preference to restrict access to data or with his/her will to share data. With respect

[65] P. 125. Derclaye, Estelle. *The Legal Protection of Databases: A Comparative Analysis*. Edward Elgar, 2008.
[66] See, pp. 18-27, *Study of the legal framework of text and data mining (TDM).* 2014.

to this category of data more legal implications arise. One could think about the situation when profile settings of two users from the same social network are not compatible, and, at the same time, one of them may share a work or material with less restrictive settings than those initially set up by the source of the work or material. In such situation data miners could gain access to the work or material, a lawful use of which is contractually restricted by the source. The same also applies to situations when users of social networks share materials available from the sources outside the network (e.g. YouTube) under distinctive contractual terms.

Hence, in the ‗many to many' scenario data mining entities cannot rely on the mere fact that data are technically accessible, but rather need to verify compliance with contractual requirements of the source of a work or material. In the modern digital world social networks generate huge amount of data, and this makes them valuable field for a data mining research. Roughly speaking, there are as many sources of data – with potentially restricted access – as users in the given social network. Costs of verifying an absence of contractual constrains on the access to data would be prohibitive for almost any Big Data projects. Therefore, the legal uncertainty exists with respect to the question whether data mining users have a lawful access to data technically accessible in social networks.

While the (4) ‗one to one' scenario related to confidential arrangements will be scrutinised later[67], situation, when access granted (3) by one to many, deserves some commentaries. The authors of the EU study on text and data mining categorise under this title data which are gathered or generated by publishers or repositories. Consequently, a lawful access to such data in the online environment is usually pre-determined by the acceptance of terms of use. Prior to the acceptance there is no lawful access to works or materials concerned.

Depending on the size of a database, costs of verifying or negotiating lawful access to collected materials would be lower or higher. Quite apparently, the bigger is the size of a database, lower costs of the verification or the negotiation process. Therefore, in the ‗one to many' scenario position of data mining entities is not legally uncertain as such. Their willingness to either check contractual terms or to negotiate a licence for accessing materials would be in the direct dependence on the interest involved.

The analysis of levels of access to data in the digital environment justifies following observations. In all four scenarios possibility of the lawful access to data is pre-determined by the willingness of a source of information to grant the access. I believe, in most cases persons granting or restricting the access to data would be guided by the subjective value of either

[67] See, *infra* Paragraph 3.

keeping secret information concerned[68], or sharing it. Two categories of the access restrictions exist in digital settings: contractual and technical. IP laws play no significant role with respect to the question of the access. Indeed, if the information was published on the Internet, it is considered as communicated to the public. Hence, the public can access the information and use it further if additional restrictions are not in place.

One could compare access to data with the fence on the way to a piece of land. At the times when no ownership titles existed, a fence was a useful mean for protecting one‘s economic advantage. Society, at the time, did not need to grant explicit property rights in land, because, as there were enough non-occupied lands available, fences were perfectly suitable to restrict access to occupied areas. From my point of view, private persons and businesses do not need ownership titles in their data, but they need safeguards (fences) – legislative, contractual or technical, further research can demonstrate – allowing them to design access to their data as they deem appropriate.

To summarise, from the standpoint of Big Data projects, examined classification, suggested by authors of the EU study on text and data mining, proves to be a useful check-list for verification of existing legal risks with respect to data mining. To put it simply, two steps are needed: categorising data depending on access settings and evaluating risks associated with a defined access category. I believe such roadmap is a useful tool for the risk assessment even in absence of explicit legislative solutions. In addition to the levels of access to data, data mining entities can categorise data by different levels of IP protection. As was mentioned in the previous subsection, depending on categories of data rights of data producers will be different.

3. LAW ON CONFIDENTIALITY: APPLYING TO BIG DATA AND DATA MINING PRACTICES

Securing confidential information is a core strategy of almost every commercial undertaking. Indeed, if a company has a list of customers, even a modest one, it would be sensible to secure this information from its competitors. Furthermore, developments of technologies in the digital environment have provided market players with technical tools to protect commercially valuable information. Consequently, confidentiality also represents a useful mechanism for securing Big Data assets through implementation of legal and technical secrecy measures.

Before coming to the substance of this section, some terminological clarifications are due. Throughout this paper I use the terms ‗trade secrecy‘ and ‗confidentiality‘ interchangeably.

[68] At least, it seems to be true for individuals or corporations not deriving substantial benefits from their data assets, e.g. users of Facebook for private purposes or small businesses using personal websites only to manifest their economic presence. To the contrary, information agencies might be able to assess objective value of data.

This is so, because the synonymous meaning of these terms is largely acknowledged in the legal literature[69]. Being more precise, trade secrets should be considered as a division of more general category, confidentiality, but for the purposes of the present research I assume the meaning of these terms to be synonymous.

This paragraph pursues two main objectives. First objective is to assess which legal risks and associated costs will arise for data mining entities if they are interested in using confidential datasets in Big Data projects. Second objective is to assess whether data mining practices themselves constitute an asset which can be protected via confidentiality clauses.

To address the former objective, the general legal framework of confidentiality under EU and US laws is clarified in the first section, and risks and costs associated with the confidentiality regime over Big Data assets are defined and scrutinised in the second section. To address the second objective, third section provides insights on data mining practices, and fourth section addresses how these practices can be protected via confidentiality.

3.1. General Legal Framework of Confidentiality in the EU and the US

The common legal framework of confidentiality has been lacking within the EU until very recent legislative changes. Directive No. 2016/943 of the European Parliament and of the Council on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure (hereinafter – Confidentiality Directive) has been adopted on 8 June 2016, obliging member states to incorporate its provisions in national laws within 2 years.

Adoption of the Directive was traditionally preceded by the study of relevant legal and economic objectives of the European internal market. Study on trade secrets in the EU internal market summarises certain common characteristics of the confidentiality definition in various national jurisdictions: „(i) it is technical or commercial information related to the business; (ii) it is secret in the sense that it is not generally known or easily accessible; (iii) it has economic value consisting of conferring a competitive advantage to its owner; and (iv) it is subject to reasonable steps to keep it secret“[70].

Apparently, due to the lack of conformity, precise legal definitions across national jurisdictions had had distinctive features. For instance, authors of the EU study note: „while most Member States make a reference to the need of information to have a commercial or economic value, some other Member States have instead a reference to the interest of the trade secret

[69] See, p. 9 Fenerly, D.D. *Historical Perspectives on Criminal Laws Relating to the Theft of Trade Secrets* (1970) 25 Bus. Law. at 1535. and citations provided therein. *Forms of trade secret protection*. Master thesis. 2005.
[70] P. 5, *Study on Trade Secrets and Confidential Business Information in the Internal Market*. April 2013. Contract number: MARKT/2011/128/D.

holder‘[71]. Although these terminological differences initially had no independent influence on data mining activities, rather being typical for any cross-border transactions, the Confidentiality Directive has brought around a new level of harmonisation. Article 2(1) of the Directive stipulates:

> (1) „trade secret‘ means information which meets all of the following requirements:
> a) it is secret in the sense that it is not, as a body or in the precise configuration and assembly of its components, generally known among or readily accessible to persons within the circles that normally deal with the kind of information in question;
> b) it has commercial value because it is secret;
> c) it has been subject to reasonable steps under the circumstances, by the person lawfully in control of the information, to keep it secret.

As detailed analysis of the provision is hardly relevant for this research, let us provide definition existing under the US legal order and make brief comparative remarks instead. Section 1(4) of the Uniform Trade Secrets Act states:

> „Trade secret‘ means information, including a formula, pattern, compilation, program, device, method, technique, or process, that: (i) derives independent economic value, actual or potential, from not being generally known to, and not being readily ascertainable by proper means by, other persons who can obtain economic value from its disclosure or use, and (ii) is the subject of efforts that are reasonable under the circumstances to maintain its secrecy.

Two differences can be noted from the surface of the cited provisions. First, the EU definition clarifies meaning of the term „secret‘, while the US definition only provides some examples of confidential information. Second, the EU definition contains the puzzling addition stipulating that a trade secret „has commercial value because it is secret‘. I believe that the EU legislator refers here to the subjective commercial value of a trade secret from its owner‘s perspective. This is so, because a disclosed trade secret still has a commercial value of the objective nature: the value from the perspective of those to whom it was disclosed or from the society‘s perspective, in general.

All in all, I believe that existing terminological discrepancies do not raise serious concerns in the Big Data framework. Any risks arising from differences between national legislations can be mitigated through careful drafting of confidentiality clauses, which is not an issue for skilled legal counsels. However, this overview, revealing peculiarities of the confidentiality concept, would serve as the helpful background for the subsequent analysis.

[71] Ibid.

3.2. Confidentiality and Availability of Big Data

Data mining entities bear economic costs due to the confidentiality imposed on the certain types of data which, if available, could be beneficial for Big Data projects. The authors of the article „Access Barriers to Big Data" opine: „Barriers to data collection […] arise from limited information on who owns the relevant data, or on the costs of locating and contracting with such data holders"[72]. Hence, there are two factors resulting in the additional costs for Big Data projects: (1) defining which confidential data might have a value for the research and (2) approaching/contracting market players controlling these data.

There are no legal risks associated with mentioned costs. If data mining entities are not able to define data or if costs of contracting are prohibitive, they simply would proceed without using confidential datasets. At the same time, the associated economic risk is that final results of data mining without crucial confidential datasets might be obsolete. Furthermore, the following economic consideration sheds more light on the issue of confidentiality arrangements in the Big Data framework: „Where there is demand, there is a market—while the costs of producing big data may be marginal for the companies, the licensing or propriety costs for accessing this information may be significant"[73]. The authors of the cited article also emphasise: „Increasingly, private companies sell their data for a profit; over time the generation of big data may become a major objective, and not just a by-product, of their activity"[74].

These commentaries imply that the confidentiality regime over Big Data assets is becoming a rule rather than exception. In other words, entities controlling Big Data assets tend to keep large datasets confidential to ensure subsequent negotiations over licensing fees. Consequently, data mining entities need to deal with the market reality instead of simply excluding datasets which are not available due to maintained confidentiality.

In general, a negotiation process over trade secrets is considered as a complex and non-transparent. This is so, because the precise value behind a trade secret at the time of negotiating is apparent only for its owner, a potential licensor, while a prospective licensee is not able to assess objective value of the secret information and, therefore, not able to decide whether demanded royalties are reasonable and – even more extremely – whether secret is at all useful. This controversy of the negotiation process over trade secrets was aptly described by commentators in following words: „Licensing trade secrets requires a careful consideration of a

[72] P. 359, Rubinfeld, Daniel L.; Gal, Michal S. *Access Barriers to Big Data*. 59 Ariz. L. Rev. 339, 382 (2017).
[73] P. 22. Cornelia Hammer, Diane C.; Kostroch, Gabriel Quiros. *Big Data: Potential, Challenges and Statistical Implications*. September 2017. International Monetary Fund.
[74] Ibid.

so-called black box dilemma – the trade-secret owner cannot *let the cat out of the bag*, and the potential licensee will not want to *buy a pig in a poke*'[75].

The core justification of the dilemma's existence is a risk of misappropriation of the secret during the negotiation process. I believe, however, that this dilemma does not exist in the negotiation process over digital databases for the purposes of their use in Big Data projects. As I have already explained earlier, the objective of virtually all Big Data projects is the knowledge discovery process. The word „discovery' bears significant distinct meaning here, as it is literally unknown for entities involved in a Big Data project, before finalising data analysis, whether a valuable information, or knowledge, would be discovered. At the same time, for entities involved in Big Data projects volume and variety of data are of a paramount importance, and a potential licensor can contractually guarantee both size of licensed datasets and categories of data included without endangering its proprietary rights in confidential information during negotiations.

To provide an example, let us assume that a pharmaceutical company possesses vast amounts of data concerning its customers' reactions on the certain types of drugs. If there is another party interested in these side effects, e.g. a research institution, negotiations over licensing a dataset could be substantiated by plain revealing an amount of customers and categories of drugs involved. It even might be feasible for the negotiation process to provide specific – de-identified for the sake of privacy – excerpts of information from a database.

If one compares the confidentiality regime with previously discussed regimes under copyright and database laws, the former fits perfectly into the Big Data framework. The reason is obvious: confidentiality is a creature of private arrangements, while copyright and database regimes are heavily regulated by statutory provisions. Legislation of countries needs to be amended to better reflect economic demands, and confidentiality arrangements are here from the very beginning to fill various gaps in the legislation.

The important conclusion for the present research is that the confidentiality legislation does not need to be adapted for the needs of Big Data projects. Parties generating and controlling information are free to decide which types of information to keep secret in attempts to achieve competitive advantage and which types of information to license out for the use in Big Data projects. In the information society such approach guarantees stability of the market, and it should not be abandoned even in attempts to support Big Data innovations.

[75] See, P. 6 and citations therein, Tsotsorin, Maxim, Practical Considerations in Trade Secret Licensing (October 1, 2012). Available at SSRN: https://ssrn.com/abstract=2334060.

3.3. Data Mining Practices: Independent Asset of Big Data Projects

It might sound surprising but even in the era when Big Data, machine learning, artificial intelligence and other complex concepts of computer science are more and more praised by scientific and business circles, human skills and judgement still play the central role in the progress of data analytics. To put it simply, the present state of art is still perfectly captured by the following statement of Babbage C., who lived in 19th century and authored the concept of a digital programmable computer:

> On two occasions, I have been asked [by members of Parliament], "Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?"...I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question[76].

Previous section was dealing with the question of the lawful access to data, providing advice how to speculate between distinctive access settings in the digital environment. Once data mining users lawfully accessed data, they would perform an actual data analysis. At this stage it would be also necessary to assess whether specific data mining practices are exempted from the duty to acquire authorisation under applicable legislation and to assess economic value of these practices. Detailed examination of risks and uncertainties related to data mining in the EU and the US were dealt with in the first paragraph. In this section I would provide some illustrations of data mining practices, revisiting previous conclusions where appropriate.

a) Data Mining Practices: Automated Algorithms and Human Skills and Judgement

Throughout this paper I have used term ‗data mining practices' many times. If data mining is defined as a whole knowledge discovery process performed on digital databases, quite evidently, ‗data mining practices' would refer to specific actions required to achieve objectives of the knowledge discovery. Current state of art demonstrates that these specific actions represent the mixture of a computer algorithms' employment with a human decision-making.

Professor M. Mattioli in his article ‗Disclosing Big Data' discusses practical insights of the data mining process. He was able to gather information through personal interviews with entrepreneurs. Marketing director at TrueLens, the company providing sociographic insights on the customers[77], commented: ‗Data scientists at the company might have a hunch, for example, that customers most interested in the airline's new route are those who live in major cities and who also enjoy skiing. Relying on this hunch, they will create a selection of customers who

[76] P. 67, Babbage, Charles. *Passages from the life of a philosopher*. Longman, Green, Longman, Roberts, & Green. 1864.
[77] See, https://www.crunchbase.com/organization/truelens. Accessed on 16.01.2018.

match these criteria‘[78]. This would be an example of data selection which, in some instances, is predetermined by a human expertise.

Another step would be cleansing of information, or determining which information is relevant and reliable: ‗A diagnosis of prostate cancer would lead us to decide that an individual was male, regardless of data that suggested otherwise, such as a petite body size‘[79] – commented a specialist working with data in a healthcare industry.

A step which takes place after the actual data mining was performed is classification. As M. Mattioli comments, this step is necessary during data analysis conducted by start-up companies specializing in the so-called ‗sentiment analysis‘. One of interviewees of Professor M. Mattioli explained: ‗There is no simple formula for creating these types of classifications; rather, what is needed is a deep knowledge of the subject matter and the ability to find patterns in the data. Classifying data often requires an appreciation for context that only a human can judge‘[80].

As one may see, human expertise plays central role in data mining practices. Opinions of specialists from various economic sectors demonstrate that there is no automated algorithm developed so far which could exclude necessity of human‘s assistance. One could argue that data analysts perform what EU database law and US copyright law on collections and compilations consider as selection and arrangement. Accordingly, the question arises whether operations performed by data analysts satisfy originality requirement either under EU database law or under US law. Unfortunately, straightforward answer to this question hardly exists. As was pointed out in the literature, even across relatively harmonised EU legal landscape, national courts find various originality thresholds to be appropriate for the copyright protection[81]. Anyhow, above illustrations of data mining practices do demonstrate that they can be viewed as an independent asset for data mining entities, and this observation is valuable for further discussion of confidentiality over data mining practices.

b) ETL: Extract, Transform and Load

Before scrutinising data mining practices in the confidentiality framework, another example of these practices would be helpful. As I already noted in the first paragraph, the Big Data impact on the global economics has been noticed by the EU legislator, and the legislative proposal, which objective is to incentivise data mining employment in the course of scientific

[78] P. 560, Mattioli, Michael, *Disclosing Big Data*, 99 Minn. L. Rev. 35, 584 (2014).
[79] Ibid, p. 561.
[80] Ibid, p. 569.
[81] See, e.g., p. 6 Margoni, Thomas, *The Harmonisation of EU Copyright Law: The Originality Standard* (June 29, 2016). Available at SSRN: https://ssrn.com/abstract=2802327: ‗the determination of the precise meaning [of originality] is left to national laws and tribunals‘.

research, is currently pending before legislative bodies of the EU. This proposal was preceded with a thorough study of the text and data mining legal framework. According to this study, a model of data mining named ETL is recognised as more widespread than other data mining practices. Hence, it would be useful for the current study to provide some observations with respect to this model. It manifests itself in the following steps:

1. Individual content is extracted from outside sources (extraction);
2. If necessary, content is transformed to fit operational needs (transformation);
3. Content is loaded into a data set, repository or collection (loading);
4. Data miners gain access to the data and the mining (analysis) tools are applied to the data set;
5. New knowledge is created (usually a report can be drafted)[82].

First three steps – extraction, transformation and loading – coincide with actions of selection and cleansing. However, the latter two refer more precisely to technical activities of data analysts.

All steps of the ETL model of a given Big Data project require the careful risk assessment. The authors of the EU study examine steps of the model at length, testing them against relevant provisions of the InfoSoc and Database Directives[83]. In general, there are two important questions to answer: (1) whether reproduction under the applicable law takes place? (2) if it takes place, whether it still can be exempted under exceptions and limitations to copyright? The second paragraph of this paper, applying EU copyright and database law to Big Data transactions, partly answers these questions[84]. The issue remains that a high level of legal uncertainty is not excluded, and costs of eliminating this uncertainty, without introducing either pro-active contractual framework or legislative solutions, are exceedingly high.

Previous analysis has emphasised that classifying levels of the access to data is useful for preliminary risk assessment, before even starting data mining. The ETL model, in turn, is helpful to assess risks which could arise in connection with data mining itself. Although the level of legal uncertainty is higher than in the ‗access to data' situation, decision makers involved in Big Data projects frequently would be able to evaluate whether they are willing to take a risk or not.

This section has further emphasised that there are legal uncertainties with respect to reproduction of pre-existing datasets for the purposes of data mining. Furthermore, in absence of statutory exemptions data mining can be restricted as such through contractual arrangements. It is not entirely clear whether steps of the ETL model or other specific steps of data mining are exempted from the duty to acquire authorisation in absence of explicit legislative provisions concerning data mining. Considering previous analysis of the applicable laws, the risk

[82] P. 28. *Study of the legal framework of text and data mining (TDM).* 2014.
[83] See, for the detailed analysis p. 28 et seq. *Study of the legal framework of text and data mining (TDM).* 2014.
[84] See, s*upra* Paragraph 2.

assessment should be exercised by interested parties on the case-by-case basis, and I believe that available methods of the risk assessment discussed above can serve that purpose. At the same time, this section highlights that data mining practices themselves constitute an independent asset – unique combination of computer analysis with human skills and judgement - which, in some instances, may deserve protection, and this issue would be in focus of the following section.

3.4. Protecting Data Mining Practices via Confidentiality

Data mining practices can be viewed as an independent asset of a company deriving benefits from Big Data. Recalling previously provided definitions of data mining, it is referred to as (1) automated processing of digital materials, (2) statistical tests on quantitative data without proper theoretical preparation, (3) the whole process of the knowledge discovery[85]. Furthermore, previous section illustrates that data mining is a combination of automated algorithms with humans' creative choices.

Mentioned characteristics of data mining practices exclude or, at least, highly implicate possibility of their reverse engineering. Professor M. Mattioli notes in this respect: „Secrecy over [data mining practices] may be even easier to maintain than secrecy over software methods. The recent commentary describing big data's disclosure problem indicates that, unlike software, big data practices cannot be reverse-engineered"[86]. Therefore, at the current stage of technological developments, confidentiality is a highly reliable mechanism for securing economic advantages derived from data mining practices.

By the same token, it is less likely that former employees could misappropriate confidential data mining practices. Such misappropriation would require possessing both exceptional skills in data analysis and detailed technical knowledge of the specific automated algorithms involved.

As one may see, trade secrecy is the exceptionally valuable strategy for protecting data mining practices. From the perspective of data mining entities, such protection almost excludes probability of revealing secret practices to competitors. In this scenario, competitors are players of the same market, which also employ data mining or plan its implementation in the operation of their businesses. Hence, applying risk assessment in the light of aforementioned, the strategy would be to enter into non-compete/non-disclosure agreements with employees possessing exceptional knowledge about specifics of data mining practices in question. The same applies for collaborative data mining projects, when cooperating parties need to employ essentially the same data mining algorithms and develop similar analytical techniques.

[85] See, *supra* pp. 11-12 and citations provided therein.
[86] P. 553, Mattioli, Michael. *Disclosing Big Data*, 99 Minn. L. Rev. 535, 584 (2014).

It is interesting to observe how confidentiality regimes imposed on both Big Data assets and data mining practices simultaneously could hinder and foster Big Data transactions. As I have shown before, the confidentiality regime fits perfectly in the framework of Big Data transactions: it secures valuable databases from unauthorised access, hindering Big Data utilisation, but at the same time, it does not complicate a negotiation process over digital datasets, fostering Big Data transactions. All in all, there are two considerations to make. First, the confidentiality regime imposed on data tends to be a common practice in the digital environment. Second, if a data mining entity performs successful data analysis, actual practices of its employees will constitute an additional value (know-how).

Let us assume, for the sake of this illustration, that accessed databases contain some unified types of information. Hence, specialists performing data mining possess a unique knowledge how to operate with this information. For instance, in the beginning of this paper I have provided example of the Big Data project focused on eliminating gaps between actual and assessed times of flights arrival[87]. Specialists who worked with flights schedule and forecast data are most likely to succeed in similar projects.

Therefore, fostering of Big Data transactions would take place, because Big Data projects are most likely to be successful when specialists apply their previous expertise to similar sets of data. Nevertheless, hindering of Big Data transactions results from the fact that Big Data projects, in which less experienced data analysts are involved, bear higher costs when competing in the market place.

To sum up, data mining practices represent an independent asset which, though in rare situations, might require protection against misappropriation. Most importantly, developments of Big Data transactions turn data mining practices into the negotiation leverage for data mining entities, although such situation is possible mainly in cases when a database owner is interested in results of data mining.

All in all, this paragraph was devoted to the law on confidentiality. I believe it was more appropriate to discuss issues of confidentiality in the separate part of the thesis, mainly because law on confidentiality represents a unique set of rules, not strictly comparable with copyright and database law scrutinised in other parts of this paper. This paragraph highlighted that confidentiality arrangements provide useful tools for Big Data transactions. There is hardly any need to adapt trade secrecy legislation to the changes on the digital market. Confidentiality suits both data producers and data mining entities, as it can be applied both to digital datasets and to data mining practices. Hence, at this stage of the discussion it can be concluded that confidentiality is the most suitable legal instrument for Big Data innovations.

[87] See, s*upra* Section 1.1.

4. IP RIGHTS IN BIG DATA: GENERAL LIMITATIONS UNDER EU AND US LAW

In previous paragraphs I have mainly discussed which pieces of EU and US legislation safeguard rights of data producers and how to balance interests of data mining entities against these safeguards. The risk assessment was mainly at heart of the discussion concerning database and copyright laws. With respect to the confidentiality arrangements, there was more focus on the business strategy as such, because law on confidentiality – even taking into account its relatively non-harmonised state – appears to be more beneficial for both data producers and data mining entities. Moreover, I attempted to clarify concepts of ownership and access to data in the digital settings. Most importantly, previous analysis can be used as a part of the risk assessment roadmap.

This paragraph, in turn, approaches exceptions and limitations[88] enshrined in EU and US IP laws. First section examines whether limitations under the EU Database and InfoSoc Directives treat data mining as unrestricted act. Second section analyses whether Big Data projects may benefit from the fair use doctrine developed in the US to limit monopoly of copyright holders.

4.1. Incentivising Data Mining in the EU: Sui Generis and Copyright Limitations

This section would analyse whether the EU Database and InfoSoc Directives provide limitation to *sui generis* rights and copyrights respectively to the benefit of data mining entities. Accordingly, first subsection addresses exceptions under the Database Directive. Second subsection is devoted to copyright limitations under the InfoSoc Directive.

4.1.1. Data Mining Exception under the EU Database Directive

Could one argue that data mining practices cannot be limited by the right holder, at least, in some specific cases? Not an answer but a clue is provided in Articles 6 and 9 of the Database Directive for copyright and *sui generis* rights in a database, respectively.

> Art. 6(2)(b): Member States shall have the option of providing for limitations on the rights set out in Article 5, where there is use for the sole purpose of illustration for teaching or scientific research, as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved.
>
> Art. 9(b): Member States may stipulate that lawful users of a database which is made available to the public in whatever manner may, without the authorization of its maker, extract or re-utilize a substantial part of its contents in the case of extraction for the purposes of illustration for teaching or

[88] For the purposes of this paper terms ‗exception' and ‗limitation' are used interchangeably, regardless possible terminological differences between them.

scientific research, as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved.

I consider necessary to introduce the whole wording here in order to demonstrate how restrictive the approach of the Directive is. Moreover, even in relation to the exception concerning scientific research the wording is uncertain and requires clarifications of applicable legal standards. Questions arise, whether teaching or scientific research themselves constitute justified non-commercial purposes? What if, as a result of data mining, a university was able to patent a new drug and, moreover, further commercialize the invention? Is it already not justified? If so, from which stage of the research a university performing data mining runs the risk of infringement? This goes without saying that the successful scientific research would, most likely, have commercially beneficial outcomes.

Another drawback which adds no legal certainty is the notion that limitations under Articles 6 and 9 of the Database Directive are left for the discretion of the EU Member states. Hence, at the time of writing there can be 28 distinctive interpretations whether data mining for scientific purposes constitutes a restricted act or not. Fortunately, there is a scientific article discussing how numerous jurisdictions around the globe treat data mining under copyright laws.

Handke Christian and other authors of the article with the ironic title „Is Europe Falling Behind in Data Mining?" attempted to answer the question whether data mining for scientific purposes is allowed under different national legal orders. Authors conclude that in the European context data mining without an author's expressed consent would most likely not be allowed[89]. However, an example of the United Kingdom (still the member state at the time of writing) provides the interesting exception.

In 2014 the UK legislator has amended Copyright, Designs and Patents Act (hereafter – the UK Copyrights Act) adding Article 29A with the title „Copies for text and data analysis for non-commercial research". This amendment provided persons who have lawful access to the work with the right „to carry out computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose"[90]. Consequently, Handke Christian and others, conducting their research, listed the UK among countries allowing data mining without authorisation of a right holder.

There are several points necessary to highlight, on the basis of what was already said about data mining and limitations under the Database Directive, especially those related to copyright in a database. First, it should be noted that scholarly discussion so far has focused mainly on limitations protecting scientific research. I could find no scholarly works discussing

[89] See, pp. 13, 16-17. Handke, Christian and Guibault, Lucie and Vallbé, Joan-Josep, *Is Europe Falling Behind in Data Mining? Copyright's Impact on Data Mining in Academic Research* (June 7, 2015).
[90] The UK data mining exception would be examined in further detail in Subsection 4.1.1of this paper.

how to approach data mining for the use in commercial purposes. Nevertheless, I believe that social benefits of Big Data innovations manifest themselves both in commercial and scientific settings. To lay the emphasis here, even if one might not think about rendering commercial data mining a lawful use, one should definitely contemplate how to incentivise innovative and socially beneficial commercial Big Data projects. I would elaborate on this statement more in subsequent paragraphs of the paper.

Second, assessing the wording of the UK Copyrights Act provokes the question: do we need explicit exemption concerning data mining at all? As long as the data mining entity has a lawful access to the work at issue, couldn‘t one argue that data mining is beyond the scope of copyright protection? Indeed, the UK copyright act allows to a lawful user to perform ‗computational analysis of anything recorded in the work‘ if the purpose of the analysis is a scientific research. I believe that questions which I pose here concern a disruption of the expression-idea dichotomy. Computational analysis, or data mining, performed in order to extract and reuse facts and ideas not the expression. As long as the access to the work is lawful, why one should acquire separate authorisation – by law or by a right holder – for exploiting facts and ideas?

The answer to all these questions may lie in the plain notion of legal certainty. While it can be argued that data mining does not infringe copyright in a database in question, entering into argument is not in researchers‘ and businesses‘ interests. They would rather appreciate a clear legislative solution designating safe harbours for Big Data projects.

It is essential to provide some points summing up discussion about Big Data and EU database law. The Database Directive is not beneficial for Big Data projects, limitations under the Directive do not clearly qualify data mining as a lawful use, even if it is conducted for the scientific purposes. On the national level, legal protection to data mining practices was granted only in the UK in 2014. Thus, to restate the ironic question posed by Handke C. and others: is the EU still falling behind in data mining? Fortunately, the EU legislator seems to feel the need for a change.

The Proposal for a Directive on copyright in the Digital Single Market is currently pending before legislative bodies of the EU stipulating in Article 3 that the EU member states ‗shall provide for an exception to the rights provided for in Article 2 of [the InfoSoc Directive], Articles 5(a) and 7(1) of Directive 96/9/EC […] 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‘[91].

[91] Text of the Proposal is available at: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52016PC0593.

The date of the Proposal is 14th September 2016. Taking into consideration political hurdles of the EU legislative process it is uncertain whether and when the proposed Directive would be enacted.

4.1.2. Copyright Limitations under the InfoSoc Directive

There are two legislative provisions in the InfoSoc Directive which are relevant for the Big Data discussion: Articles 5(1) and 5(3)(a). Thus, the subsequent analysis is devoted to these two provisions.

a) Limitations of Reproduction Right: Article 5(1) of the InfoSoc Directive

Could one argue that existing limitations under the InfoSoc Directive already save the day for data mining? There are two limitations of the reproduction right under the Directive which should be examined to answer this question. First limitation is provided in Article 5(1)(b) of the Directive:

> Temporary acts of reproduction, which are transient or incidental [and] an integral and essential part of a technological process and whose sole purpose is to enable a lawful use of a work or other subject-matter to be made, and which have no independent economic significance, shall be exempted from the reproduction right provided for in Article 2.

The cited wording implies that as long as copying performed in the course of data mining bears temporary character, it might comply with remaining requirements of Article 5(1) of the Directive. Indeed, copyright law does not explicitly suggest that data mining is unlawful use. Furthermore, a copy which is made by data mining entities bears no independent economic significance, as long as it is used exclusively for internal analytical activities. However, there are at least two issues which render application of the limitation under Article 5(1) legally uncertain for data mining entities.

First, as one may see, the EU legislator construes the limitation in the narrowest way possible referring only to those temporary acts of reproduction which are either transient or incidental. Which of these types of temporary copying could take place as a part of data mining? The ECJ explains in *Infopaq I* that copying can be considered transient only when it is an essential part of a technological process and a copied material is automatically deleted without human interruption with the completion of a process[92]. To the contrary, data mining, as I demonstrated earlier, is not completely automated process, and human interventions occur at all steps of the analysis. Hence, it is problematic to argue that a copy being made is transient. Alternatively, one could argue that copying during data mining is incidental. For instance, the authors of the EU study on text and data mining opine: _It is not excluded that the temporary

[92] § 64. The ECJ, *Infopaq International A/S v Danske Dagblades Forening*, 16 July 2009, Case C-5/08.

copy made during the extraction is incidental to the main act of exploitation of the work, i.e. the analysis of the work'[93].

Second, even if copying performed by data mining users is incidental and all other requirements under Article 5(1) are complied with, Recital 44 of the InfoSoc Directive gives more food for thought, stating that in the context of new electronic environment the scope of certain exceptions or limitations provided for in the Directive may have to be even more limited when it comes to certain new uses of copyright works and other subject-matter.

In other words, Recital 44 provides that in the changing digital environment limitations should be construed in a way guarding monopoly of copyright holders. In my opinion, the better approach could be if this recital had restated the balancing purpose behind copyright law: a legitimate interest of copyright holders on the one side, fair remuneration, shall be balanced with a legitimate interest of the society on the other, the economic progress with a minimum deadweight loss. A deadweight loss for the society is the situation when authors receive remuneration which is higher than necessary to incentivise production of new works. Moreover, I do not consider fair to presume that an author of a specific, single literary or artistic work, is entitled to restrict „free riding' on facts which are collected in the copyrighted subject-matter as long as the lawful access to a work is not in question. Arguably, an author could prohibit data mining contractually, but there should be no such general prohibition on the level of the copyright legislation.

To summarise, limitation under Article 5(1) of the InfoSoc Directive is construed as extremely narrow. This is justified by the fact this limitation aims to render lawful technical copying which is presupposed by functioning of many digital services. The most relevant example would be: streaming video from the website which has a licence to exploit a work. There still would be an additional act of copying behind this explicitly authorised by an author: the copy which is kept on servers of the website owner. The additional copy being made – which could be kept infringing in absence of Article 5(1) – is the automatic copy in the cash memory of a computer. Shortly speaking, at the time of adopting the InfoSoc Directive the EU legislator had in mind this and other technical instances of copying of a work in the digital environment. Law-makers have had no intention to predict such ways of works' exploitation, as data mining presupposes.

b) Limitations of Reproduction Right: Article 5(3)(a) of the InfoSoc Directive

Another limitation relevant for the Big Data discussion is enshrined in Article 5(3)(a) of the InfoSoc Directive and implies that Member States may limit reproduction right in the case of

[93] P. 46, *Study of the legal framework of text and data mining (TDM).* 2014.

work‘s use for the sole purpose of illustration for teaching or scientific research, as long as the source, including the author‘s name, is indicated, unless this turns out to be impossible and to the extent justified by the non-commercial purpose to be achieved.

In general, the analysis in the previous subsection concerning similar provision in the Database Directive is wholly applicable here. However, further details should be provided. The meticulous examination of the limitation under Article 5(3)(a) is presented in the EU study on text and data mining, which I use throughout my research as one of the sources of reference. A reader is further referred to this study[94], if interested in more details, and hereafter I would highlight only the most problematic points.

First, non-mandatory character of the provision creates discrepancies for data mining cross-border collaborative projects. Implementation of the limitation is left for the discretion of member states. Hence, there is no harmonisation of the limitation across various jurisdictions within the EU. For instance, the Netherlands and Spain have not implemented limitation at all, while all other states provide limitation of a different scope depending on the subject-matter at issue – extracts from works or only listed categories of works. This leads to the situation of legal uncertainty: data mining entities may be protected by the limitation within one jurisdiction, while data miners in other jurisdiction, copying the same data, may not. Legal basis for such controversy is the principle that the extent of copyright protection shall be governed exclusively by the laws of the country where protection is claimed, as per Article 5(2) of the Berne Convention.

Second, there is no coherent interpretation of the proviso ‗to the extent justified by the non-commercial purpose to be achieved‘, while for decision-makers involved in Big Data projects definite borders of this purpose limitation are of a paramount importance. The UK High Court of Justice opines: ‗Presumably, any research which, at the time it is conducted, contemplated or intended, should be ultimately used for a purpose which has some commercial value will not be within the permitted act‘[95]. This opinion does not provide any final decision but rather suggests directions for further arguments, in favour or against data mining activities at issue.

To simplify the matter, one could argue that fundamental scientific research is conducted for non-commercial purposes, while applied scientific research, presumably, is intended for economic or commercial advantage. However, there are various economic activities which cannot be plainly classified as serving for either commercial or non-commercial purposes. For

[94] For the exhaustive analysis of Article 5(3)(a) of the InfoSoc Directive see, pp. 50-67, *Study of the legal framework of text and data mining (TDM).* 2014.
[95] §23. *Ordnance Survey v Green Amps Limited*, Case No: HC07C00249, High Court of Justice, Chancery Division Intellectual Property, 5 November 2007, [2007] EWHC 2755 (Ch).

instance, the EU study suggests that if a private company funds a research for philanthropic purposes in the field non-related to its business activities – telecommunication company funds a healthcare project – such data mining research most certainly fulfils the ‗non-commercial purpose' limitation. Nevertheless, the private company derives reputational benefits which, anyhow, provide some economic advantage over competitors.

All in all, data mining performed as a part of scientific research for non-commercial purposes can benefit from the limitation under Article 5(3)(a) of the InfoSoc Directive, keeping in mind discrepancies of non-harmonisation of the limitation within the EU and unclarities of the legal test under the Directive.

Following final considerations sum up the analysis of the EU database and copyright laws in relation to Big Data innovations. More specifically, the discussion concerning author's right for fair remuneration in the case of work's reproduction for the scientific research can provide interesting highlights when it is viewed from the standpoint of copyright limitations.

According to Article 5(5) the exceptions and limitations under the InfoSoc Directive shall only be applied in certain special cases which do not conflict with a normal exploitation of the work or other subject-matter and do not unreasonably prejudice the legitimate interests of the copyright holder. Interpreting this article, M. Walters and S. Lewinski state: ‗[although the] Directive does not require that the right holders receive fair compensation, [for exploitation of their works under Article 5] the application of the three-step test under Article 5(5) may result in an obligation of the Member States to provide for some form of fair compensation or remuneration'[96].

Indeed, if this notion is applied to situations of the scientific research exception, in general, some authors may be eligible for fair compensation for the use of their works. However, I would argue that this is not the case if data mining was applied to a work at issue. One should distinguish the protected subject-matter under the Database Directive from the one under the InfoSoc Directive. The former protects copyright in selection and arrangement of a database and investment in creating a database, but the latter guaranties protection of the expression enshrined in an artistic or literary work.

Understanding of the distinction is crucial. When data mining entities analyse large sets of data, they indeed ‗free ride' to the certain extent on both copyright and *sui generis* rights protected under the EU database law. To the contrary, they do not ‗free ride' on the original expression which is embodied in numerous works. Hence, in my opinion, one cannot reasonably argue that data mining of a single work in question conflicts with its normal exploitation.

[96] P. 1045, Walter, W.; Lewinsky, S. V.; *European Copyright Law: A Commentary*, Oxford University Press, 2010.

All in all, the existing copyright and database laws in the EU arguably could provide copyright limitation with respect to data mining performed for the scientific research. However, a requirement of a non-commercial purpose complicates the application of this limitation. Moreover, wording of the Directives implies that exceptions and limitations should be construed as narrowly as possible, which raises additional legal uncertainties. Thus, explicit provisions balancing copyright and the *sui generis* right with data mining activities are needed.

4.2. Big Data and US News Media: Unlawful Contents Scrapping or Fair Use?

In the US legal order closed list of copyright limitations is not provided, but instead the fair use doctrine has been developed as a legal test assessing whether some specific uses of copyrighted subject matter can be exempted from liability. In this section I would examine the fair use doctrine in the Big Data framework. To address issues of data mining in the context of the fair use doctrine, the US case law would be helpful.

The case concerning commercial exploitation of Big Data scrapping techniques, *Fox News Network, LLC v. TVEyes, Inc.*[97] (hereinafter – *TVEyes*), was adjudicated by the US Court for the Southern District of New York. The court considered the action of a copyright infringement brought by Fox News against TVEyes and the action of a fair use defence raised by the defendant. Subsequently, the ruling of the first instance was reversed by the court of appeal.

a) Background of the Case

TVEyes creates a searchable database of radio and television contents through closed-caption technology. It grants access to captured contents via pre-paid subscription. At the time of the judgement there were over 22 000 subscribers. It is important to note that TVEyes does not provide subscription packages for the general public. As stated on the official page of TVEyes, their subscription is tailored for interests of following actors:

- Mid-sized organisations that generate publicity in multiple media centres;
- National brands, government agencies or political campaigns running multiple PR initiatives;
- Multinational, military and government organisations tracking ongoing PR campaigns[98].

TVEyes subscribers have following benefits: (1) searching through database of news media back to 32 days from the date of search and, accordingly, displaying media which are objects of interest, (2) archiving media files in the personal digital library, (3) sharing links to media files with others via e-mails, and (4) downloading media files directly on subscribers' computers. The court also considered to be important the opportunity to search the media database using search terms of time and date.

[97] *Fox News Network, LLC v. TVEyes*, Inc. United States District Court, S.D. New York. August 25, 2015 124 F.Supp.3d 325 2015.
[98] https://www.tveyes.com/which-tveyes-is-right-for-you/. Accessed on 24.12.2017.

Fox News, in general, alleged that „by making Fox News' content available to TVEyes subscribers, TVEyes is diverting potential licensees, website visitors, and therefore revenue, from Fox"[99].

b) Fair Use and Data Mining

The unique legal standard defining boundaries of the copyright holder's monopoly over exploitation of its work is developed under the US copyright law. The fair use doctrine, originally elaborated by the case law, is presently enshrined in 17 U.S.C. § 107. This legislative provision provides that to establish whether a particular use of a work constitutes a fair use, four following factors shall be considered:

(1) the purpose and character of the use, including whether such use is of a commercial nature or is for non-profit educational purposes;

(2) the nature of the copyrighted work;

(3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and

(4) the effect of the use upon the potential market for or value of the copyrighted work.

Before continuing examination of *TVEyes*, let us briefly consider how the fair use doctrine would apply to Big Data practices at large. First factor highlights that commercial exploitation of a copyrighted material will not automatically exclude the fair use defence. Hence, exploiting Big Data for commercial purposes still may benefit from the fair use.

Second factor, establishing that the nature of a copyrighted work shall be considered, in practice, means that works of factual nature enjoy stronger protection than fictional works[100]. This criterion is highly beneficial for Big Data projects, because, as I have previously demonstrated, data mining is performed to extract facts from large digital datasets, which justifies following considerations. First, data mining entities would be primarily interested in works representing compilations of facts and figures (e.g. databases, news media). Second, even if data mining entities are interested in fictional works, their focus would be primarily on facts and figures manifested in such works rather than on the original authors' expression. For the sake of precision, I should mention that techniques of mining the original expression with assistance of machine learning do exist[101]. However, probable legal implications of resultant

[99] *Fox News Network, LLC v. TVEyes*, Inc. United States District Court, S.D. New York. August 25, 2015 124 F.Supp.3d 325 2015.

[100] See, ibid. See also, *Harper & Row Publishers, Inc. v. Nation Enters.* 471 U.S. 539, 563, 105 S.Ct. 2218, 85 L.Ed.2d 588 (1985).

[101] See, e.g.: *Using Machine Learning to Generate Lyrics in the Style of your Favourite Artist*. Available at: http://www.thedurkweb.com/using-machine-learning-to-generate-lyrics-in-the-style-of-your-favourite-artist/.

derivative works are beyond the scope of my research, which is focused primarily on the knowledge discovery process and not on available techniques of the expression mockery.

Third factor of the fair use legal standard concerns the size of the portion reused by the party raising the fair use defence. This strongly corresponds with substantiality requirements for establishing infringement under the EU Database Directive[102]. All the arguments, which I raised in this respect previously, are relevant here as well. Data mining entities may copy the substantial part of a database in order to perform data analysis, but they will usually reuse further a quantitatively insignificant part of an original work. What concerns qualitatively substantial reuse of a work at issue, it seems that under the fair use doctrine US courts would not defer such an argument, might it be posed. As one may remember, qualitative substantiality under the EU database law is to be found when investment in the creation of a database's part was significant. However, as I discussed earlier[103], protection of investment, or „sweat of a brow", under the US IP law does not exist. Thus, the only implication raised by the third factor with respect to data mining is how to justify copying of an entire work before performing data analysis.

Fourth factor under the US fair use doctrine is related to economic consequences of a reuse within the market at issue. This factor more comprehensively than others reflects the main purpose of the fair use doctrine: to promote the progress of science and useful arts[104]. The US Supreme Court has also commented that fourth factor is „undoubtedly the single most important element of fair use"[105]. Joseph A. Tomain, analysing fourth factor in his meticulous examination of *TVEyes*, observes that: „Copyright holders must generally establish substitution market harm to defeat the fair use rights of another, [...] courts must consider [...] the potential harm to the copyright holder if others engaged in similar activities as the defendant, [and ...] courts must balance the benefit the public will derive if the use is permitted and the personal gain the copyright owner will receive if the use is denied"[106].

Evidently, the legal standard under fourth factor would be highly beneficial if applied to Bid Data practices. Copyright holders in majority of cases would not even be able to prove that exploitation of a work, created by means of data mining, causes substitution market harm. Even in the instances, like in *TVEyes,* when plaintiff might argue about a substitution market harm as such, remaining factors cited above further limit copyright protection. However, let us examine these factors below, taking closer look at legal implications in *TVEyes*.

---

[102] See, *supra* pp. 17-20.
[103] See, *supra* pp. 22-23.
[104] See p. 59 and citations therein, Joseph A. Tomain. *Big Data and the Fourth Estate: Protecting the Development of News Media Monitoring Databases*, 12 J. Bus. & Tech. L. 53, 72 (2016).
[105] *Harper & Row,* 471 U.S. at 566.
[106] Pp. 62, 64, Joseph A. Tomain. *Big Data and the Fourth Estate.* See also, *Bill Graham Archives v. Dorling Kindersely Ltd.*, 448 F.3d 605, 613 (2d Cir. 2006).

c) TVEyes and Data Mining: Ruling of the US Court for the Southern District of New York

TVEyes states on their official webpage that its ‗clips are for Internal Review, Analysis and Research only. Any editing, reproduction, publication, rebroadcast, public showing, public display or placement on any website is prohibited and may violate copyright laws'[107]. This is questionable whether this disclaimer appeared on the webpage in the recourse of the lawsuit or TVEyes initially maintained that its business practices were covered by fair use. Arguably, freedom of speech defence also can be relevant legal justification behind defendant's business activities[108]. Freedom of speech is enshrined in the First Amendment to the US Constitution. However, this concept is more related to philosophy of law, while the academic focus of this subsection is interplay between law of economics and the fair use doctrine. It was already established that data mining entities can benefit from fair use. Let us now examine specific defendant's actions, which were at issue in *TVEyes*.

In general, New York S.D. Court found that ‗TVEyes' core function — recording content, putting it into a searchable database and, upon a keyword query, allowing users to view short clips of the content up to 32 days from the date of airing' survived legal test under fair use doctrine. However, the court then focused on additional features of TVEyes' news media database. The court analysed whether functions of (1) archiving videos, (2) downloading them, (3) sharing them by e-mail and (4) performing search queries, using ‗date and time' search terms, were fair use.

The district court finally found that archiving function is fair use, and e-mail function can be fair use if reasonable measures against subsequent violations would be implemented. However, the court found that downloading and date-time search functions do not enjoy fair use defence. I would farther focus on the functions which were denied fair use protection.

The court denied that 3rd and 4th functions constitute fair use, because the court did not view them to be integral part of the core TVEyes' database functioning. In relation to e-mail sharing TVEyes argued that this function is crucial for offline criticism. The court rejected the argument stating that ‗very few remaining locations in the United States lack internet connectivity'[109]. To protect the date-time search feature, TVEyes argued that it helps subscribers to perform a search as accurate as possible. This feature, for instance, cures problems of misspelled names, as they fixed in TVEyes' search indexes. The court denied this argument, as well, stating that the time-date search makes TVEyes' database substituting product and effectively deprives Fox News from its licensing fees.

---

[107] https://www.tveyes.com/. Copyright Statement. Accessed on 24.12.2017.
[108] See, pp. 67-71 Joseph A. Tomain. *Big Data and the Fourth Estate.*
[109] *Fox News Network, LLC v. TVEyes*, Inc.

J. Tomain, analysing at length court‘s reasoning, takes pro-defendant‘s position. The author suggests additional arguments in support of the fair use defence for features which were denied protection by the New York S.D. Court. Arguing for downloading feature, J. Tomain simply states that ‗being connected to the internet should not be a requirement of fair use‘[110]. Protecting the date-time search feature, J. Tomain opines that ‗date and time searches help discover the absence of coverage at a specific time and allow for a comparative analysis of news coverage across stations at a specific time‘[111].

In the nutshell, although the court finds largely in favour of the defendant, its application of the fair use doctrine is quite cautious. The court attentively scrutinised separate technical features of the news media service to strike the proper cost and benefit balance in its final ruling. J. Tomain and other scholars who submitted amicus briefs to support TVEyes‘ business model, in turn, argue for more liberal approach, which could be substantially favourable for other innovative users of Big Data.

It is important to note that the *TVEyes* example is concerned about balancing rights of the data mining entity with rights enshrined in particular identifiable works. TVEyes used and still uses scrapping techniques for copying materials which are protected by copyright as such. For the purposes of criticism, TVEyes is interested not only in ideas and facts but also in the way how this information is delivered to the general public – including gesticulation of reporters, choices of background scenes, etc. Keeping in mind the purpose of criticism and, perhaps, wider rational of the freedom of speech, the fair use doctrine allows data mining entities ‗free riding‘ on the author‘s original expression.

d) TVEyes and Data Mining: Ruling of the US Court of Appeal for Second Circuit

The decision of the first instance with its partial rejection to find for the fair use was heavily criticised as I demonstrated above. Nevertheless, the decision of a court of appeal came as even bigger disappointment for TVEyes and its proponents. In February 2018 the US Court of Appeal for Second Circuit has reversed the decision of the district court to the extent that it found actions of TVEyes to constitute fair use. The following citation reflects the main ground of the appellate court‘s vision:

> ‗TVEyes‘s re-distribution of Fox‘s content serves a transformative purpose insofar as it enables TVEyes‘s clients to isolate from the vast corpus of Fox‘s content the material that is responsive to their interests, and to access that material in a convenient manner. But because that re-distribution makes available to TVEyes‘s clients virtually all of Fox‘s copyrighted content that the clients wish to

[110] P. 66 Joseph A. Tomain. *Big Data and the Fourth Estate.*
[111] Ibid.

see and hear, and because it deprives Fox of revenue that properly belongs to the copyright holder, TVEyes has failed to show that the product it offers to its clients can be justified as a fair use'[112].

This citation was the result of applying the fair use test to the _watch' function of the TVEyes database. The purpose of this paper is not to criticise court decisions in *TVEyes* but rather to find whether these decisions can provide some guidance for Big Data projects. Although, the court of appeal refused to find for fair use in *TVEyes*, its ruling is still a food for thought.

First, one should bear in mind that *TVEyes* concerns content scrapping techniques. Such techniques realise wholesale copying without any noticeable transformation or adaptation of the copied contents. On pages of this paper I am examining primarily different data mining scenarios: when all data may be copied but not all data are re-utilised. Data mining technicques, undoubtedly deserving protection, perform the knowledge discovery on digital datasets, which is way ahead of the plain copying of the copyrighted materials.

Second, the decision of the court of appeal examines important factor: whether reuse of a copyrighted material serves a transformative purpose. This factor corresponds to the first step under the fair use legal test and highly relevant for the Big Data discussion. The court of appeal notes: _To be transformative, a use must do something more than repackage or republish the original copyrighted work; it must add something new, with a further purpose or different character, altering the first with new expression, meaning or message'[113].

While this is arguable whether transformative purpose could have been established in *TVEyes* and probably the US Supreme Court would have its final say in this respect, this is clearly the case for majority of data mining practices. If data mining entities analyse vast amounts of copyrighted materials, re-utilising only their factual contents, transformative purpose of the use would be quite straightforward.

To summarise the discussion, let us emphasise that four steps under the fair use legal test are highly useful for the risk assessment related to data mining. If mined copyrighted materials fall under the US legal protection, it would be wise to apply 17 U.S.C.A § 107 as a part of pro-active risk assessment. Nevertheless, one should bear in mind that the fair use legal test is highly circumstantial. Although this legal instrument provides open-ended list of exceptions under 17 U.S.C.A § 107, it also stipulates quite speculative steps for legal analysis. Hence, under the US fair use doctrine there would never be a final answer whether specific data mining practices infringe copyright or not.

[112] *Fox News Network, LLC v. TVEyes, Inc.*, No. 15-3885 (2d Cir. 2018).
[113] Ibid. (Quotation marks omitted)

5. FOSTERING BIG DATA INNOVATIONS: LEGISLATIVE EXCEPTIONS AND ALTERNATIVE SOLUTIONS

I have previously analysed general IP laws – copyright, confidentiality, database law – relevant for the Big Data discussion. I have further examined general exceptions and limitations of IP rights from which Big Data innovations, arguably, could benefit. This paragraph, taking another step forward, would be concerned about legislative and alternative solutions which could incentivise investments in Big Data projects and, accordingly, foster Big Data innovations. The paragraph is divided on two sections, first of which deals with recent legislative solutions, assessing how they affect Big Data transactions. Second section discusses possible alternative mechanisms of regulating Big Data markets, focusing on independent social initiatives and probable licensing solutions akin to those adapted by the open source software community and by advocates of creative commons.

5.1. Legislative Solutions and Scientific Arguments to Foster Big Data Transactions

Earlier on the pages of this paper I have already mentioned legislative solutions dealing with the issues of Big Data economic developments. These solutions mainly concern introduction of specific data mining limitations to copyright and other IP rights. At this stage it is necessary to revisit existing legislative initiatives to define which of them better incentivise investments in Big Data projects.

The UK legislator has already introduced data mining limitation to copyright and database rights for purposes of non-commercial research. In Japan stricter legislative solution was implemented, effectively excluding a need to acquire authorisation for data mining irrespective of its purpose. First subsection is devoted to the comparative analysis of UK and Japanese legislative solutions. Second subsection, in turn, would bring to the table initiatives argued for by scholars.

5.1.1. Data Mining Legislative Solutions: Soft in the UK, Sweeping in Japan

On 1st June 2014 the Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014 came into force, implementing Article 29A in the UK Copyrights Act which reads as follows:

> **Copies for text and data analysis for non-commercial research**
>
> (1) The making of a copy of a work by a person who has lawful access to the work does not infringe copyright in the work provided that—
>
> (a) the copy is made in order that a person who has lawful access to the work may carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose […]

(5) To the extent that a term of a contract purports to prevent or restrict the making of a copy which, by virtue of this section, would not infringe copyright, that term is unenforceable.

As it can be seen from the surface of the provision, data mining – or in words of the UK legislator, text and data analysis – is allowed for non-commercial research and cannot be contractually restricted as long as a data mining entity has a lawful access to a work at issue. Such provision, although devoted exclusively to situations of a research for non-commercial purposes, still eliminates some legal uncertainties related to data mining of specific copyrighted works.

Another question is how Article 29A affects *sui generis* right in databases under the EU Database Directive[114]? *Sui generis* right was transposed into the UK legislation through the adoption of the Copyright and Rights in Databases Regulations 1997. At the same time, the UK legislator preferred to employ slightly different terminology, opting for the term „database right". It follows from the explanatory note to the Regulations 2014 that Article 29A serves the purpose „to cover all types of copyright work"[115]. One would not find in the amended text of the UK Copyrights Act any straightforward information whether Article 29A applies to a database right. I believe, however, that this article does limit a database right.

Article 29A allows copying of works, and thus, one should establish whether a database is treated as a work under the UK Copyrights Act. According to Article 3A of the UK Copyrights Act, a database „means a collection of independent works, data or other materials". Upon reading this provision the conclusion follows that the term „database" is wider than the term „work". Nevertheless, Article 3 of the UK Copyrights Act mentions a database as one of forms of a literary work. Hence, Article 29A of the UK Copyrights Act exempts wholesale copying of a database in situations of text and data analysis for non-commercial research. Consequently, this is also not possible under UK law to assert a database right against data mining entities. In addition, this conclusion is supported by the fact that acts infringing database right – extraction and re-utilisation[116] – are narrower than wholesale copying of a database. Thus, if UK law treats wider act – wholesale copying – as non-infringing, this should overcome narrower database owner's monopoly under the database right.

To sum up discussion so far, I believe, the above examination of the UK law implies that a database right does not fit perfectly in the traditional IP legal framework. A database right was already subject of critics on previous pages of this paper, and, moreover, it is heavily criticised,

[114] Brexit is still a future, and, moreover, comparison of national and European laws is valuable for the academic analysis of existing issues of interpretation.
[115] Explanatory note to the Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014.
[116] See, Regulation 16 of the Copyright and Rights in Databases Regulations 1997.

by scholars and some stakeholders[117]. Notwithstanding these issues of doctrinal compatibility, the analysis of relevant legal provisions demonstrates: in the UK both copyright in any works and *sui generis* rights in databases cannot be asserted against data mining entities conducting non-commercial research.

Comparison of the UK piece of legislation with copyright and *sui generis* limitations in the EU law highlights another important observation. Article 29A exempts data mining conducted as a part of ‗non-commercial research', while both Database and InfoSoc Directives speak only about ‗scientific research [...] to the extent justified by the non-commercial purpose'[118]. Quite obviously, the term ‗non-commercial research' is wider than ‗scientific research for non-commercial purpose'. According to Database and InfoSoc Directives member states have limited powers to provide IP rights' limitations. Thus, arguably, the UK law provides wider limitation to copyright and database right than is allowed under the EU law.

Now, let us turn to the examination of Japanese data mining exemption. It was introduced to Japanese copyright law already in 2009. This very fact justifies a need to analyse Japanese legislative provisions, although referred to so late in the course of the present discussion. Article 47septies of the Japanese copyright law is entitled ‗Reproduction, etc. for information analysis' and reads as follows:

> For the purpose of information analysis (‗information analysis' means to extract information, concerned with languages, sounds, images or other elements constituting such information, from many works or other much information, and to make a comparison, a classification or other statistical analysis of such information; the same shall apply hereinafter in this Article) by using a computer, it shall be permissible to make recording on a memory, or to make adaptation (including a recording of a derivative work created by such adaptation), of a work, to the extent deemed necessary[119].

The Japanese provision on data mining employs the term ‗information analysis' which can be considered synonymous to the term ‗data analysis' discussed previously[120]. This is disputed in the literature how the meaning of terms ‗information' and ‗data' is related. In the nutshell, information can be determined as data in the context[121]. More importantly for the present study, one should not overlook more extensive character of Japanese provision in comparison with Article 29A of the UK Copyrights Act: no similar purpose limitation was enacted under Japanese copyright law. Hence, within Japanese jurisdiction data mining entities

---

[117] See, e.g. Banterle, Francesco, https://iplens.org/2017/12/22/first-results-of-the-public-consultation-on-revision-of-the-eu-database-directive/: ‗It is known that the introduction of the *sui generis* right has been largely criticized (going in the opposite direction than that endorsed by *Feist* in the US)'.
[118] See, Articles 6 and 9 of the Database Directive and Article 5 of the InfoSoc Directive.
[119] Translation is available at: http://www.cric.or.jp/english/clj/cl2.html#cl2_1+A47septies.
[120] See, *supra* Section 1.2.
[121] See, e.g. https://www.diffen.com/difference/Data_vs_Information.

are explicitly allowed to conduct data mining for any research irrespective of its commercial or non-commercial purpose.

Article 47septies has, on its surface, two details which raise some issues of interpretation. The proviso ‗by using a computer' restricts the copyright exception to data mining scenarios when a computer is involved as a technical mean of the analysis. This can be seen as a disadvantage, taking into account rapid growth of technologies which can develop not only computer-based platforms as technical devices for the analysis. Another proviso ‗to the extent deemed necessary' can be considered problematic by virtue of its vague wording. If one speaks about reproduction of data for its subsequent data mining, there would be no objective criterion to establish ‗the extent deemed necessary'. To pose a rhetorical question, how could a researcher possibly restrict copying if a research model itself is predictive? In other words, necessary amount of copying is not known by a researcher prior to actual data mining. Authors of the EU study on text and data mining also opine that this proviso can create some problems when applied in practice[122]. In any case, Japanese blanket copyright exemption explicitly allowing both commercial and non-commercial data mining is unprecedented sweeping legislative solution, which was introduced to the world as early as in 2009. If one takes into account rapid technological developments, nine years of legal certainty bring a huge economic advantage to Japan-based Big Data projects.

Although this paper cannot go into analysis of all legal solutions implemented around the globe, following citation will serve as a valuable summary for the discussion. Sergey Filippov and Paul Hofheinz, authors of the Lisbon counsel study on text and data mining, after examining approaches to the Big Data phenomenon in various jurisdictions, have made the observation:

> The U.S. retains its long-held leadership position in the study of text and data mining, [...] Europe is slipping further behind and [...] Asia is rising strongly. In Asia, relatively recent entrants [...] – China and India – challenge the leadership of the traditional Asian centres of excellence in the field – Taiwan, Korea and Japan. And there is increasing evidence that European research institutions are being forced to reach outside of Europe to build better teams for text-and-data-mining-related consortia – not because the foreigners' researchers are more able, but because their laws are smarter and more straightforward than the ragged patchwork of rules which apply in Europe[123].

So, in view of scholars, the US retains leading position on the Big Data market, although, as I discussed earlier, the fair use doctrine does not provide explicit exception allowing data mining. At the same time, mixture of the fair use doctrine with an absence of a database right, similar to the EU *sui generis* right, can be seen as a strong basis of the US competitive position.

[122] See, p. 12. De Wolf & Partners. *Study of the legal framework of text and data mining (TDM).* 2014.
[123] P. 14. Filippov S., Hofheinz P. *Text and Data Mining for Research and Innovation*. Issue 20/2016. The Lisbon Counsel.

Asian jurisdictions compete with the US in the market of Big Data investments. Japan can compete due to the clear blanket exception, granting data mining entities freedom to engage in both commercial and non-commercial Big Data projects. Other Asian countries, including Israel, can compete, because their legal orders provide rules akin to the US fair use or UK fair dealing doctrines[124].

The European Union ‗is slipping further behind' due to the absence of any legal provisions which could guarantee some legal certainty for Big Data projects. Article 3 of the EU Proposal provides modest limitation which unchains only ‗research organisations [...] to carry out text and data mining [...] for the purposes of scientific research'. Furthermore, wording of the Proposal is not fully consistent with the language of Infosoc and Database Directives, which exempt ‗scientific research [...] to the extent justified by the non-commercial purpose'. Literal interpretation of Article 3 of the EU Proposal leads to the conclusion that data mining would be allowed only for research organisations, although, arguably, they would be allowed to use data mining for a commercial purpose.

To the date, the UK solution in Article 29A of the Copyrights Act already eliminates some legal uncertainties associated with Big Data transactions, allowing data mining for non-commercial research. Nevertheless, one can still argue that such limitation is not in line with the EU Directives regulating copyright and database law.

Anyhow, I believe that both UK solution and EU proposal are too narrow and cannot attract Big Data investments in current international settings. Let us now turn to initiatives argued for within the scientific literature devoted to the subject.

5.1.2. Fostering Big Data Transactions: Vision of Scholars

In this subsection I would address interesting recommendations provided by scholars as probable ways to foster Big Data transactions and research. I would primarily discuss two articles. First is devoted to the novel concept of ‗dataright'. In view of concept's advocate, Professor M. Mattioli, this new right could incentivise disclosure of data mining practices in return for greater monopoly over data reuse granted to investors by society. Second article does not address the term ‗Big Data' as such. However, the article is concerned about data reuse via licensing strategies and without specific legislative changes. Such approach makes this article interesting for the current discussion.

a) Big Data trade-offs: Greater Disclosure or Traditional Confidentiality

Before commencing examination of the ‗dataright' introduced by Professor M. Mattioli, this is worth mentioning that the objective of my analysis is different from that pursued by the

[124] See, p. 4, Ibid.

scholar. While Professor M. Mattioli examines ‗dataright' in his article from the perspective of incentivising greater disclosure of data mining practices, I am concerned about incentivising investments in Big Data projects and establishing more legally certain environment for Big Data innovations. I believe that the existing confidentiality framework analysed earlier better addresses demands of the Big Data market. Nevertheless, the ‗dataright' concept is interesting subject of scrutiny, as it is novel and extreme in comparison with the ideas and concepts discussed so far.

First of all, this is important to mention that the scholar develops a legal concept of ‗dataright' not arguing for its instant incorporation into intellectual property laws but with the objective to fuel the debate[125]. Hence, the criticism of the concept, which I provide hereafter, perfectly suits the purpose of the scientific discussion initially foreseen by Professor M. Mattioli. As demonstrated by the scholar, three important elements are to be taken into consideration when construing a new *sui generis*[126] right: ‗(1) subject matter covered by the right; (2) exclusive rights conferred to publishers of this subject matter; and (3) a set of acquisition rules upon which exclusivity is conditioned'[127].

Subject matter of the right is viewed as any data that have been collected with the use of any methods not apparent to a person of ordinary skills. Exclusivity of the ‗dataright' could establish limited control of data producers over downstream uses and applications of their data. Finally, exclusivity should be granted only upon disclosure of ‗all data collection and organization practices relevant to each piece of data they seek to protect'[128].

This brief description of the ‗dataright' concept, as it was introduced by Professor M. Mattioli, already gives a material for the criticism. I believe that, at least, two problematic points can be highlighted. Firstly, the scope of a subject matter falling under hypothetical protection is not clear. The strict adherence to characteristics explained before suggests that any data could be protected by ‗dataright'. No distinction is made between categories of data, although such distinction could be crucial to define probable threshold of protection – e.g. personal data and data containing copyrighted works are definitely of a distinctive legal nature which implicates unified overarching approach as to their reuse or prohibition of reuse. Moreover, if one compares the ‗dataright' with the already existing *sui generis* right under the EU Database Directive, conclusion can be drawn that the former provides even wider monopoly for data producers. This, in my view, contradicts to general trends under US law to treat serious legislative interventions cautiously.

[125] See, pp. 537-538, Mattioli, Michael. *Disclosing Big Data*, 99 Minn. L. Rev. 535, 584 (2014).
[126] Fitting neither into the framework of copyright nor of patent law, hypothetical of ‗dataright' can be better constructed under a separate *sui generis* regime.
[127] p. 578, Mattioli, Michael. *Disclosing Big Data*, 99 Minn. L. Rev. 535, 584 (2014).
[128] Ibid, p. 579.

Secondly, even if a subject matter of ‗dataright' is further clarified, issues would still remain with regard to the grounds of granting such monopoly to data producers. The crucial question to pose: what society gets in return for providing additional exclusive rights in data? Indeed, I have shown earlier that data mining practices do constitute independent valuable assets for entities involved in Big Data projects. These practices are the mixture of various analytical algorithms with human skills and judgement. Hence, their value can be seen as subjective one: generally known analytical methods are intertwined with skills and labour of analysts/researchers assisting Big Data projects. This value can be of a great interest for some other market players.

To provide an illustration, if one telecom company has successfully conducted a Big Data project in one region, it might be willing to license out developed data mining practices – combination of human skills and analytical algorithms – to another telecom company located in the region, where the former does not compete for the customers. To put it simply, data mining practices constitute great value for either a data mining entity itself or for a small group of companies involved in similar business activities. Nevertheless, this is highly questionable whether data mining practices can be seen as an asset of the same great value for the society in general – such value which could justify monopoly for data producers on the legislative level.

Recalling that the objective of my analysis is searching for better ways to incentivise investments in Big Data projects, one might argue that ‗dataright' indeed could encourage greater investments in return for widely construed protection. However, this is not apparent whether social benefits which might be derived due to these investments could outweigh social costs of granting ‗dataright' to data producers. Furthermore, I believe that if there is no economic certainty whether new initiatives would benefit society at large, it would be wise to leave the question for private arrangements rather than implement experimental solutions.

All in all, there are no credible indications that the confidentiality regime cannot cope with demands of the Big Data market. Moreover, as was highlighted previously, confidentiality is perfectly suitable for the negotiations concerning both large sets of data and data mining practices[129]. One of scholars has also observed that ‗trade secret law actually encourages broader disclosure and use of information, not secrecy. [T]he legal protection trade secret law provides serves as a substitute for investments in physical secrecy that companies might otherwise make'[130]. Thus, to the date, the ‗dataright' mechanism can be seen as an artificial substitute to existing confidentiality settings. From my point of view, there is no justification for the imposition of that mechanism without serious empirical economic evidence.

[129] See, *supra* Sections 3.2., 3.4.
[130] Pp. 333-34, Lemley, Mark A. *The Surprising Virtues of Treating Trade Secrets As IP Rights,* 61 STAN. L. REV. 311, 332 (2008).

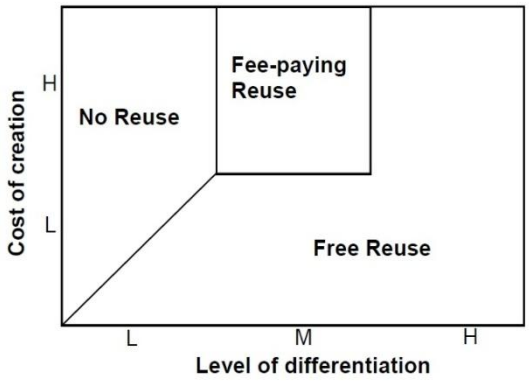
b) Big Data Reuse[131] and Licensing Strategies

Inventor of the web, Tim Berners-Lee, stated, „the exciting thing is serendipitous reuse of data: one person puts data up there for one thing, and another person uses it another way"[132]. This statement demonstrates that the suitability of data for coincidental reuse was commonly known long before Big Data came around as the phenomenon. If such characteristics of data are well-known already for some time, this justifies an assumption that there could be strategic solutions for the Big Data utilisation without resort to any legislative changes. The Big Data phenomenon simply realises the objective of the coincidental reuse, or predictive analysis, on the effectively new scale due to volume, velocity, variety and veracity of data[133].

Authors of the article „Legal Challenges and Strategies for Comparison Shopping and Data Reuse" have suggested valuable ways of data exploitation on the basis of existing US case law and with some remarks on the EU database law. This article mainly concerns data reuse for comparison shopping, but I would apply strategies described by authors[134] to the relationship between Big Data producers and data mining entities. Big Data licensing strategies can be discussed on two levels: (1) larger framework of Big Data reuse and (2) specific steps to be taken by the participants of Big Data transactions.

When one considers larger Big Data framework from standpoint of economics, digital datasets can be licensed out either on the fee basis or free of charge depending on costs of data creation and level of differentiation between initial data and those produced by data mining entities. There can be also cases of wholesale copying of data with subsequent reconstitution of the similar database. However, the authors have argued that such wholesale copying should be prohibited, because in its result no or very little economic value is added. Let us, therefore, focus on free reuse of data versus fee based licensing of data when no wholesale copying of database followed by reconstitution of contents takes place.

[131] For the purposes of this subsection the term „reuse" is synonymous to the term „data mining", as the latter is one of possible ways to reuse data.
[132] Frauenfelder, M. *Sir Tim Berners-Lee*. Technology Review, Vol. 107, No. 8:40-45, 2004.
[133] See, *supra* Section 1.1.
[134] See, p. 234 et seq. Madnick Stuart E.; Zhy Hongwei. *Legal Challenges and Strategies for Comparison Shopping and Data Reuse*. Journal of Electronic Commerce Research, VOL 11, NO 3, 2010.



Differentiation of initial database and database created by data miners[135]

From the one hand, majority of Big Data projects would fall under the category eligible for free reuse of data. This is so, because no matter how high were costs of creating initial, or input, data, output data produced by data mining entities would differ significantly from mined datasets. Considering law of economics, society, in general, would benefit more from free data mining, and data producers still would bear no or very little economic costs due to the reuse of their data[136].

From the other hand, there are a handful of cases when data can be licensed out upon payment of fees. I have previously discussed the US case, *Fox News Network, LLC v. TVEyes*[137]. It was generally found by the court of the first instance that TVEyes content scrapping activities fall under the fair use exception. Nevertheless, from the purely economic consideration TVEyes reused exactly the same contents which were introduced by Fox News Network, only building around it more intelligible services for purposes of the news' analysis by businesses and decision makers. In other words, if one disregards the fair use exception, it can be economically justified to request TVEyes to pay licensing fees. At the same time, it could be also justified obliging Fox News to provide a compulsory licence, as the product introduced by TVEyes creates significant economic benefits for the society and does not hinder investments in creating news in the first place. When it comes to the European legal regime, the licensing situation is directly derived from database protection. Hence, the use of data after paying licensing fees always warranties legal certainty under the EU database law, although cannot be always justified from standpoint of economics.

In addition to the analysis of Big Data licensing from the larger perspective, there are specific steps which can be taken by the participants of Big Data transactions. First, data producers can always consider selling „private" data to data mining entities to ensure that a Big

[135] The source: Fig. 1 on p. 235. Madnick Stuart E.; Zhy Hongwei. *Legal Challenges and Strategies for Comparison Shopping and Data Reuse*. (The name of the figure is changed)
[136] It is worth mentioning that this consideration disregards the fact whether a data mining entity competes directly with a data producer. If this is the case, economic analysis would be more intricate.
[137] See, *supra* Section 4.2.

Data project has more chances to succeed and also to derive more benefits from their producers' position. This is what happened in the PASSUR Aerospace collaborative project[138] discussed above: publically available data on flights schedule and weather forecasts were mixed with confidential data possessed by a company managing the airport. Second, data producers can always consider reusing, or mining, Big Data on their own. Quite obviously, these specific steps simply represent a piece of advice on the business strategy aiming to foster Big Data transactions. There is nothing completely innovative in these steps. However, they help to make use of the existing market situation without waiting for legislative changes to come.

This section has examined existing legislative solutions dealing with Big Data implications and some scholarly opinions concerning the matter. Although legislative solutions regarding data mining are not widespread across various jurisdictions, interesting approaches do exist. Both EU and UK legislators implement quite limited solution: data mining is allowed without copyright or database holders' consent if it is performed either for scientific or non-commercial research respectively. Japanese legislator has introduced more straightforward approach: data mining for commercial purposes also does not require copyright holders' consent. With respect to database law, there is no protection mechanism in Japan akin to the EU Database Directive. I consider Japanese approach to be the most comprehensive taking into account that copyright concerns protection of author's original expression, and data mining is focused on factual dimension of an original work at issue. Hence, rejecting copyright holders any control over data mining of their works simply emphasises that this activity is beyond the scope of copyright protection.

Turning to the second part of this section, I have highlighted some scholarly opinions in order to demonstrate trends of the scientific discussion. There is little to add in this respect: some authors argue for revolutionary changes and others take more conservative position. What comes to my opinion, I believe that some legislative and alternative solutions do provide safe harbour for Big Data projects and would summarise my views on the matter in the conclusion to this paper.

5.2. Alternative Solutions: Borrowing from Unique Social Initiatives

Developments of Intellectual property law in the recent years and even centuries tend to strengthen IP protection and widen monopoly of authors and other IP right holders. On the one hand, one might argue that objective economic criteria applied by legislators all over the world have justified: to foster innovation and creativity, inventors and authors need strong legislative protection. On the other hand, various independent social initiatives appeared, such as Creative

[138] See, *supra* Section 1.1.

Commons and Free Open Source Software, in order to balance too restrictive legislative approaches. Hence, in this section, on the basis of foregoing discussion, I would approach probable alternative mechanisms able to foster Big Data transactions, focusing on FOSS and Creative Commons from the Big Data perspective.

Founders of the Free Software Foundation, clarifying ideas behind FOSS movement, stated: ‗Free software' is a matter of liberty, not price. ‗Free' as in ‗free speech', not as in ‗free beer'[139]. On the webpage of Creative Commons (hereinafter – CC) it is stated that ‗When we share, everyone wins'[140]. The FOSS licences are tailored for needs of software developers, while the CC licences can be applied virtually to all types of copyrighted materials. How to apply CC and FOSS principles in the Big Data framework? I believe, at least, two nuances should be tackled. Terms and conditions of CC and FOSS licences could be applied directly to the relationship between data producers and data mining entities (a). The main idea, anti-proprietary spirit behind CC and FOSS movements, can be followed by advocates of Big Data innovations (b).

a) Applying CC and FOSS Licensing Terms in the Big Data Framework

Let us briefly examine the text of the CC license, Attribution 4.0 International (CC BY 4.0)[141]. Examination of the main principles highlighted on the official webpage would suffice for the purposes of the present research. These principles read as follows:

> ‗**Share** — copy and redistribute the material in any medium or format
>
> **Adapt** — remix, transform, and build upon the material for any purpose, even commercially.
>
> The licensor cannot revoke these freedoms as long as you follow the license terms.
>
> **Attribution** — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use [...]'[142].

If applied to the relationship between data producers (controllers) and data mining entities, first two lines effectively authorise latter to mine licensed datasets. Furthermore, the prohibition to revoke the license can be seen as additional warranty of a legal certainty for Big Data projects. However, the problem still remains that such voluntary waiver of a right to revoke a license can be invalid in some jurisdictions.

Another issue arising already in connection with this summary, regardless of the actual license's text, is an ‗attribution' obligation. I believe that it could be troublesome for data mining entities to attribute results of data mining to specific authors of copyrighted materials. Should data miners list all mined works or only those which manifest themselves in the results of data

[139] See, https://www.gnu.org/philosophy/free-sw.en.html.
[140] See, https://creativecommons.org/.
[141] For the full text of the license, see, https://creativecommons.org/licenses/by/4.0/legalcode.
[142] See, https://creativecommons.org/licenses/by/4.0/.

mining? Moreover, should we give any deference to utilised works if only their factual part is gathered and reproduced?

If one answers the first question positively, it will lead to the obligation to attach endless list of names of right holders to results of data mining. Obviously, to comply with such obligation is troublesome. This is why UK legislators, foreseeing this „attribution" technical obstacle, have included in Article 29A of the UK Copyrights Act the phrase: „the copy is accompanied by a sufficient acknowledgement (unless this would be impossible for reasons of practicality or otherwise)". When it comes to the second question, I think it can be better addressed as a part of the issue (b).

b) Using CC and FOSS Ideas to protect Big Data innovations

In general, CC and FOSS advocates criticise a proprietary model of employing IP rights. For instance, CC promotes free sharing of works instead of royalty-based licensing, and FOSS fights against patenting software. In the nutshell, these movements attempt to cope with trends of widening IP right holders' monopoly. One should also keep in mind that CC and FOSS movements' main realm is copyright, and I believe that this is not the case for Big Data. I would rather argue that the main realm of Big Data is IT law. Now, applying to Big Data ideas of CC and FOSS, which are fighting against excessive monopoly of IP right holders, let us revisit the question: should we give any deference to utilised copyrighted works if only their factual part is gathered and reproduced? The answer is plain: no, and here is why.

The posed question is specific, while problem „copyright v. Big Data" is more general. I have already argued before that copyright cannot restrict data mining, because it is focused on facts – it is clear from the very term „data mining" – while copyright law protects one's original expression against misappropriation. More importantly, this is not only one researcher's opinion. Association of European Research Libraries, criticising legislative draft of the EU data mining exception, states:

> „**Allow Commercial and Non-Commercial Uses, Without Compensation** – *Since TDM is only used to extract facts and data, which are not copyrightable, there is no reason to limit TDM to non-commercial purposes*. The copy of the individual work is not being re-used or communicated to the public. Researchers are instead focused on setting the facts and figures in a larger context, and drawing conclusions from this work. In addition, it should be noted that research funders such as the European Commission under the H2020 programme aim for commercial impact in research. Research supports innovation which, by definition, is something that has commercial use"[143].

Barely any additional comments are needed. If lobbying of wider sweeping data mining exception is not successful after all, the feasible way for European researchers to succeed is to

[143] See, https://libereurope.eu/blog/2017/09/25/copyright-reform-help-us-ensure-an-effective-tdm-exception/ (emphasis added).

borrow CC and FOSS experience and fight against application of copyright laws to Big Data projects. This is also important to note that if lobbying of a wider exception on the level of the EU legislative bodies fails, it would not be possible to overcome this on the national level. The InfoSoc Directive provides closed list of copyright exceptions which can be enforced by member states[144], and this list hardly includes anything helpful for innovations based on data mining.

[144] See, Article 5(2) of the InfoSoc Directive.

CONCLUSION

On the pages of this master thesis I have attempted to resolve problems related to the phenomenon of Big Data. This was not the purpose of the paper to reveal whether Big Data is valuable as such. I conducted research with the assumption that Big Data *is* valuable. Taking into account that the subject of scrutiny is considerably wide, I have focused on specific types of IP – database/*sui generis* rights, confidentiality and copyright. Furthermore, I conducted the study mainly from the position of data mining entities. In other words, I analysed risks related to activities of data miners and sought solutions which could incentivise such activities. Before providing general remarks, let us focus on specific research questions posed in the beginning of the study and see whether they were properly answered. Research questions were formulated following the structure of the master thesis. Therefore, concluding remarks addressing these questions will also logically follow the structure. This is also helpful to keep in mind for comprehending concluding remarks that two main objectives were pursued: risk assessment and elaborating business strategy for data mining entities.

Which relationships define Big Data from practical and technical standpoints? How key terms related to Big Data can be defined?

Technical side of Big Data manifests itself in V-characteristics which were a subject of scrutiny in the first paragraph. These characteristics are volume, velocity, variety and veracity, related respectively to the enormous size of digital datasets, speed of their processing, various forms in which data are presented and trustworthiness of data. At least, characteristics of volume and variety emphasise that rights of numerous natural and legal persons intertwined in Big Data. Thus, these two characteristics can be seen as actual reason why subsequent scrutiny of IP laws is needed.

Big Data and data mining are key terms which were frequently employed throughout this study. Moving from buzzwords to legal concepts, I employed following definitions. Big Data refers to large sets of data fixed in digital form which cannot be processed by conventional computing tools. Data mining refers to the whole process of knowledge discovery on digital datasets through application of automated algorithms and human‘s skills and judgement.

Which risks of Big Data exploitation are associated with copyright and database laws and relevant exceptions and limitations?

Database law is the most relevant for the Big Data discussion, as virtually all data, to which mining algorithms can be applied, are represented in the form of digital databases. The study has demonstrated that database law restrictions are problematic for data mining entities. As

one may remember, database law is a unique creature of the EU legal order. Other similar laws adopted around the globe can be found but it was not in the scope of this paper to discuss them. The main concern related to the EU database law is that position of data mining entities under this regime is legally uncertain. To establish whether data mining constitutes extraction or re-utilisation, one should conduct an intricate legal analysis, weighing specific data mining activities against the test under relevant provisions of the EU Database Directive.

Situation is also not satisfactory for data mining entities under the EU legal order if one assesses general exceptions and limitations under the EU Database Directive. At this stage one should remember that two distinctive rights exist under this Directive: copyright in selection and arrangement and *sui generis* right to prevent extraction and/or re-utilisation of database's contents. Arguably, data mining does not infringe the former, or, at least, it would be troublesome to establish such infringement. With respect to the latter, if a database was copied under data mining activities, this would, most likely, constitute re-utilisation or extraction. Hence, the last resort of data mining entities would be to justify their activities under applicable exceptions and limitations. Unfortunately, relevant limitation under the Database Directive exempts only the use for the sole purpose of illustration for teaching or scientific research, as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved.

To put it simply, Big Data projects in the EU encounter high level of legal uncertainties due to the restrictive language of the Database Directive. At the same time, in the US data producers can benefit only from the limited protection granted under the hot news doctrine. Hence, as long as data mining entities do not misappropriate instant value of „hot" data, they will not encounter similar problems mining data which falls under the US legal order.

If one assumes that copyright law restricts data mining activities, the same level of legal uncertainty exists within the EU under the InfoSoc Directive, as well. However, in the US data miners can benefit from the fair use exception. The fair use legal test includes, among others, a step requiring assessment of the effect upon the potential market of the copyrighted work. Fulfilment of this step essentially requires establishing that a substitution market was created as a result of infringing activities. In majority of cases data mining would not lead to the creation of such substitution market. Therefore, data mining entities can benefit from the fair use doctrine.

Whether and how Big Data projects may benefit from using confidentiality clauses?

Law on confidentiality is much more beneficial for Big Data than copyright or database law. Confidentiality arrangements provide useful tools for Big Data transactions. There is hardly any need to adapt trade secrecy legislation to the changes of the digital market.

Confidentiality suits both data producers and data mining entities, as it can be applied both to digital datasets and to data mining practices. Due to the nature of Big Data projects – predictive analysis – negotiations between data producers and data mining entities do not endanger confidential information of data producers. This should allow to conduct negotiations in the safe and mutually trustful environment. Furthermore, confidentiality clauses can protect even data mining practices themselves if such protection is needed within specific market place. Thus, it can be concluded that confidentiality is the most suitable piece of law for Big Data innovations.

Whether data mining activities infringe copyrights in specific literary and artistic works?

Copyright cannot restrict data mining, because it is focused on facts – it is clear from the very term ‗data mining' – while copyright law protects one's original expression against misappropriation. The very fact that the ‗copyright v. Big Data' controversy exists twists basic idea behind copyright law: author's expression is protected against misappropriation, but facts are not. The US Supreme Court was already forced to clarify for the judiciary system of the US that copyright does not protect ‗sweat of the brow', i.e. investment in the creation of a work. I believe this is the time for the EU legislator to explicitly state that copyright also does not protect facts and figures. While, in general, necessity of such clarification might sound like nonsense, it seems that it is needed with respect to data mining.

Which modern initiatives, legislative and alternative, available to deal with the Big Data phenomenon?

I have discussed three legislative solutions on the pages of this paper: Japanese, UK and European. Japan has introduced into its copyright law the data mining exception already in 2009. Japanese blanket copyright exemption, explicitly allowing both commercial and non-commercial data mining, is unprecedented sweeping legislative solution. Bearing in mind aforementioned, I believe that this is the only right way to introduce data mining exception into copyright law.

UK legislator has introduced data mining exemption in 2014. The analysis of relevant legal provisions has demonstrated: in the UK both copyright in any works and *sui generis* rights in databases cannot be asserted against data mining entities conducting non-commercial research. At the same time, the UK term ‗non-commercial research' is wider than the term ‗scientific research for non-commercial purpose' adopted on the level of EU Directives. According to Database and InfoSoc Directives member states have limited powers to provide IP rights' limitations. Thus, arguably, the UK law provides wider limitation to copyright and database right than it is allowed under the EU law.

In the European Union any legal provisions which could guarantee some legal certainty for Big Data projects are not yet in force. Literal interpretation of Article 3 of the EU Proposal leads to the conclusion that data mining would be allowed only for research organisations, although, arguably, their scientific research may have commercial purpose.

Independent social initiatives serve as an alternative way to overcome hesitations of legislators. If lobbying of wider sweeping data mining exception is not successful in the EU, the only way for European researchers to borrow CC and FOSS experience and fight against application of copyright laws to Big Data.

Whether and how data mining entities could mitigate risks under IP laws currently in force? As I have stated several times, I approach Big Data and IP law from the perspective of data mining entities. Hence, I have attempted to provide some guidance, helping to mitigate risks of these market players. In this respect following pieces of advice can be given to data mining entities.

As this study has highlighted problems existing in some most developed legal orders when it comes to the Big Data exploitation, the first step for data mining entities would be to define jurisdictions of interest. Depending on the level of friendliness to data mining activities – on the basis of this study one can view EU regime as the least friendly and Japanese market as the friendliest – data miners can adopt their business model and organise research projects in the most efficient way.

When jurisdictions of interest are defined, specific licensing strategies can be adopted. If data miners are forced to deal with EU databases, options would be either acquire a license or to use available databases without a separate license, making risky assumption that no infringement takes place. If jurisdiction of interest does not provide for data producers a right akin to EU *sui generis* right, data mining entities would be able to make the next step: to approach owners of data kept confidential.

When it comes to confidential data, I believe that as long as its owners do not compete directly with data mining entities they may have strong incentives to license their data out. Moreover, data mining entities may have leverage in negotiations – developed data mining practices. For instance, situation may be that an owner of confidential data is interested in data mining but does not possess necessary knowledge to run Big Data project independently. Hence, data mining entities can use developed skills of its employees to procure better deal over Big Data assets.

All in all, in this study I have attempted to provide overview of three specific types of IP enshrined in Big Data – databases, copyrighted works and confidential information. This overview was provided mainly from perspective of data mining entities and, to the certain extent,

without meticulous analysis of specific legal mechanisms. The depth of the scrutiny was dictated by the objective that, from the very beginning, I have attempted to provide general guidance for businesses rather than to resolve some legal problems related to Big Data. Accordingly, a prospective research might focus more specifically on topics: Big Data and database law, Big Data and copyrights, Big Data and alternative social initiatives. As for this paper, I hope it will serve as a valuable source for future scientific inquiries.

BIBLIOGRAPHY

1. Babbage, Charles. *Passages from the life of a philosopher*. Longman, Green, Longman, Roberts, & Green. 1864.

2. Borghi, M.; Karapapa, S. *Copyright and Mass Digitization: a Cross-Jurisdictional Perspective*, Oxford University Press, 2013.

3. Born, Gary. *International commercial arbitration, Volume I-II.* Kluwer Law International, Alphen aan den Rijn, 2009.

4. boyd, danah and Crawford, Kate, Six Provocations for Big Data (September 21, 2011). A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 2011. Available at SSRN: https://ssrn.com/abstract=1926431.

5. Brooks et al (2017). *Artificial Intelligence vs. Machine Learning vs. Data Mining 101 – What's the Big Difference?* (Guavus Blog). Available at: http://guavus.com/artificial-intelligence-vs-machine-learning-vs-data-mining-101-whats-big-difference/.

6. Cornelia Hammer, Diane C.; Kostroch, Gabriel Quiros. *Big Data: Potential, Challenges and Statistical Implications*. September 2017. International Monetary Fund.

7. Corrales M.; Djemame K. *A Brokering Framework for Assessing Legal Risks in Big Data and the Cloud*. In: Corrales M., Fenwick M., Forgó N. (eds) *New Technology, Big Data and the Law. Perspectives in Law, Business and Innovation*. Springer, Singapore. 2017.

8. Chirita, Anca D. *The Rise of Big Data and the Loss of Privacy*. June 15, 2016. Available at SSRN: https://ssrn.com/abstract=2795992.

9. Derclaye, Estelle. *The Legal Protection of Databases: A Comparative Analysis*. Edward Elgar, 2008.

10. Dumbill, Edd. *What is big data?* Available at: http://radar.oreilly.com/2012/01/whatisbigdata.html.

11. Ekstrand, Victoria Smith; Roush, Chirstopher. *From Hot News to Hot Data: The Rise of Fintech, the Ownership of Big Data, and the Future of the Hot News Doctrine*. 35 Cardozo Arts & Ent. L.J. 303, 340 (2017)

12. European Union. De Wolf & Partners. *Study of the legal framework of text and data mining (TDM).* March 2014.

13. Eytan, Adar. *The Two Cultures and Big Data Research*. 10 ISJLP 765, 782 (2015).

14. Fagan, Frank. *From Policy Confusion to Doctrinal Clarity: Successor Liability from the Perspective of Big Data*. 9 Va. L. & Bus. Rev. 391, 456 (2015).

15. Fenerly, D.D. *Historical Perspectives on Criminal Laws Relating to the Theft of Trade Secrets* (1970) 25 Bus. Law.1535.

16. Filippov S., Hofheinz P. *Text and Data Mining for Research and Innovation*. Issue 20/2016. The Lisbon Counsel. Available at: http://www.lisboncouncil.net/.

17. Handke, Christian; Guibault, Lucie; Vallbé, Joan-Josep. *Is Europe Falling Behind in Data Mining? Copyright's Impact on Data Mining in Academic Research* (June 7, 2015). Available at: https://ssrn.com/abstract=2608513.

18. Interview with Kenth Engø-Monsen. *Big Demand for Big Data: New Telenor study on Dengue Fever in Pakistan*.
Available at: https://www.gsma.com/mobilefordevelopment/programme/digital-identity/big-demand-for-big-data-new-telenor-study-on-dengue-fever-in-pakistan.

19. Khoury, Amir H.; Bekkerman, Ron. *Automatic Discovery of Prior Art: Big Data to the Rescue of the Patent System*. 16 J. Marshall Rev. Intell. Prop. L. [i], 65 (2016).

20. Kröll, Stefan; Lew, Julian D. M.; Mistelis, Loukas A. *Comparative International Commercial Arbitration.* Kluwer Law International, 2003.

21. Leaffer, Marshall. *Database Protection in the United States is Alive and Well: Comments on Davison*. 57 Cas. W. Res. L. Rev. 855 (2007). Available at: http://scholarlycommons.law.case.edu/caselrev/vol57/iss4/10.

22. Lemley, Mark A. *The Surprising Virtues of Treating Trade Secrets As IP Rights.* 61 STAN. L. REV. 311, 332 (2008).

23. Madnick, Stuart E.; Zhy, Hongwei. *Legal Challenges and Strategies for Comparison Shopping and Data Reuse*. Journal of Electronic Commerce Research, VOL 11, NO 3, 2010.

24. Margoni, Thomas, *The Harmonisation of EU Copyright Law: The Originality Standard* (June 29, 2016). Available at SSRN: https://ssrn.com/abstract=2802327.

25. McAfee, Andrew; Brynjolfsson, Erik. *Big Data: The Management Revolution*. October 2012. Harvard Business Review. Available at: https://hbr.org/2012/10/big-data-the-management-revolution.

26. Michael, Mattioli. *Disclosing Big Data*. 99 Minn. L. Rev. 535, 584 (2014).

27. Michaux, M. *Droit des bases de données*. Kluwer, 2005.

28. Petroiu, Marius. *Forms of trade secret protection: a comparative analysis of the United States, Canada, the European Union and Romani*. Master thesis. 2005.

29. Rubinfeld, Daniel L.; Gal, Michal S. *Access Barriers to Big Data*. 59 Ariz. L. Rev. 339, 382 (2017).

30. Siegan, Bernard H. *Property Rights: From Magna Carta to the Fourteenth Amendment*. Social Philosophy and Policy Foundation, Transaction Publishers. 2001.

31. *Study on Trade Secrets and Confidential Business Information in the Internal Market*. April 2013. Prepared for the European Commission. Contract number: MARKT/2011/128/D.

32. Smits, Jan M., *What is Legal Doctrine? On the Aims and Methods of Legal-Dogmatic Research* (September 1, 2015). Maastricht European Private Law Institute Working Paper No. 2015/06. Available at SSRN: https://ssrn.com/abstract=2644088.

33. Tomain A., Joseph. *Big Data and the Fourth Estate: Protecting the Development of News Media Monitoring Databases*. 12 J. Bus. & Tech. L. 53, 72 (2016).

34. Tsotsorin, Maxim, Practical Considerations in Trade Secret Licensing (October 1, 2012). Available at SSRN: https://ssrn.com/abstract=2334060.

35. Tzanis, George. *Biological and Medical Big Data Mining* (January 2014). Available at: https://www.researchgate.net/publication/261958613.

36. Ursic, Helena; Custers, Bart. *Legal Barriers and Enablers to Big Data Reuse*. 2 Eur. Data Prot. L. Rev. 209, 221 (2016).

37. Virtanen, P. *Evolution, practice and theory of European database IP law.* Lappeenranta: Lappeenrannan Teknillinen Yliopisto, 2008.

38. Walter, W.; Von Lewinsky, S. V. *European Copyright Law: A Commentary*. Oxford University Press, 2010.

39. Xuqiong Wu, *E.C. Data Base Directive*, 17 Berkeley Tech. L.J. 571 (2002). Available at: http://scholarship.law.berkeley.edu/btlj/vol17/iss1/33.

40. *The Big Data Explosion: Maximizing information value while minimizing risk*. (2013) Information Management, Volume 42, (2), p s2.

41. *Using Machine Learning to Generate Lyrics in the Style of your Favourite Artist*. Available at: http://www.thedurkweb.com/using-machine-learning-to-generate-lyrics-in-the-style-of-your-favourite-artist/.

INDEX OF LEGAL SOURCES

Case Law

1. *Bill Graham Archives v. Dorling Kindersely Ltd.,* 448 F.3d 605, 613 (2d Cir. 2006).
2. *Dictionnaire Permanent des Conventiones Collectives.* Tribunal de Grande Instance Lyon, 28th December 1998.
3. *The British Horseracing Board and Others.* The European Court of Justice. C-203/02. 09.11.2004.
4. *Feist Publications, Inc. v. Rural Telephone Service Co.*, Inc., 111 S.Ct. 1282, 1295, 499 U.S. 340, 359–60 (U.S.Kan.,1991).
5. *Fox News Network, LLC v. TVEyes, Inc*. United States District Court, S.D. New York. August 25, 2015 124 F.Supp.3d 325 2015.
6. *Fox News Network, LLC v. TVEyes, Inc.*, No. 15-3885 (2d Cir. 2018)

7. Frauenfelder, M. *Sir Tim Berners-Lee*. Technology Review, Vol. 107, No. 8:40-45, 2004.
8. *Jeweler's Circular Pub. Co. v. Keystone Pub. Co.,* 281 F. 83, 88 (C.A.2 1922).
9. *Harper & Row Publishers, Inc. v. Nation Enters.* 471 U.S. 539, 563, 105 S.Ct. 2218, 85 L.Ed.2d 588 (1985).
10. *Infopaq International A/S v Danske Dagblades Forening*, The European Court of Justice. 16 July 2009, Case C-5/08.
11. *International News Service v. Associated Press*, 248 U.S. 215 (1918).
12. *National Basketball Association v. Motorola, Inc.*, 105 F.3d 841 (2d Cir. 1997).
13. *Ordnance Survey v Green Amps Limited*, Case No: HC07C00249, High Court of Justice, Chancery Division Intellectual Property, 5 November 2007, [2007] EWHC 2755 (Ch).
14. *Ruckelshaus v. Monsanto Co.,* 104 S. Ct. 2862, 2877, 467 U.S. 986, 1011 (U.S.,1984).

Statutory Law

1. The Berne Convention for the Protection of Literary and Artistic Works 1886, as amended in 1979.
2. The Code of Laws of the United States of America. Constitution of the United States.
3. The Code of Laws of the United States of America. Title 17: Copyrights.
4. Copyright, Designs and Patents Act 1988. The United Kingdom.
5. The Copyright and Rights in Databases Regulations 1997. The United Kingdom.
6. The Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014. The United Kingdom.
7. Copyright Law of Japan. Copyright Research and Infomation Center (CRIC) October, 2016 Translated by Yukifusa OYAMA et al.
8. Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society.
9. Directive No. 96/9/EC of the European Parliament and of the Council, of 11 March 1996 on the legal protection of databases.
10. Directive No. 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure.
11. Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market. Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52016PC0593.
12. The Uniform Trade Secrets Act with 1985 Amendments. The United States of America.

INTERNET RESOURCES

1. https://creativecommons.org/
2. http://www.cric.or.jp/english/clj/cl2.html#cl2_1+A47septies
3. https://www.crunchbase.com/organization/truelens
4. https://www.diffen.com/difference/Data_vs_Information
5. https://www.gnu.org/philosophy/free-sw.en.html
6. https://iplens.org/2017/12/22/first-results-of-the-public-consultation-on-revision-of-the-eu-database-directive/
7. http://libereurope.eu/resources/
8. https://en.oxforddictionaries.com/
9. https://www.tveyes.com/