INTERNATIONAL
HELLENIC
UNIVERSITY

TITLE: Text and Data Mining (focused on EU)

Vasiliki Karamousali

SCHOOL OF ECONOMICS, BUSINESS ADMINISTRATION & LEGAL STUDIES

A thesis submitted for the degree of
Master of Science (MSc) in Art, Law and Economy

Student Name:	KARAMOUSALI VASILIKI
SID:	287121868883
Supervisor:	Prof. IRINI STAMATOUDI

I hereby declare that the work submitted is mine and that where I have made use of another's work, I have attributed the source(s) according to the Regulations set in the Student's Handbook.

JANUARY 2016
Thessaloniki - Greece

CONTENTS:

ABSTRACT

The following dissertation refers to the Text and Data Mining (TDM) and certain basic issues about the related legal framework in the European Union. It is important to start our analysis from the definitions that describe better TDM and a small distinction in order to understand better the real meaning both of text and data mining. The present and the future are the information society and the use of all the available data. Information may be available in various ways, from freely available on the Web and largely available on social networks, to available from publishers to potential users after the acceptance of special terms and conditions.

Through a short mention to the history of TDM it would be understood that data analysis-as some referred to TDM- is a new, innovative development in the technological world. Information about Big data and the levels of access to information for analysis purposes are a significant step before moving on to the main analysis of this paper. In order for us to understand the legal issues that arise from text and data mining process, we have to know the way of doing TDM: the available content and the extraction of the new information.

The main analysis will focus on the InfoSoc Directive (2001/29/EC) and Database Directive (96/9/EC) as the basic legal framework behind TDM. The exceptions that are referred to the paper are of major importance. Licensing is another option to the regulation of relations between stakeholders, both with positive and negative options, as it will be referred in the text below.

Finally, it is expected that all the abovementioned affect the influence of TDM in our world. Europe seems to be in a much lower level of using TDM than USA and the reasons have to do -in a big part- with the existing legal protection.

INTRODUCTION

Definition

It is said that data and information is money in today's world. The need for structured information and for new knowledge extracted from the existent data and texts is a demand of our century. Text and Data mining is a method to extract useful information from the data and the texts and transform it into a more understandable and usable form.

Trying to find a definition for Text and Data mining, the researchers have come to many theories, some of them will be referred in the following text. "Text mining is the process that turns text into data that can be analyzed […] [while] data mining is an analytical process that looks for trends and patterns in data sets that reveal new insights" by Jonathan Clark. The author analyses differently the terms of "text mining" and "data mining". It is interesting that although in literature the definitions of data mining are much more extensive and precise, the phrase "text mining" as a research term appears 17 times more often than "text data mining" on the Web[1].

Giving some more definitions, the following definition gives a very clear picture of Text mining. "Text Mining may be loosely characterized as the process of analyzing text to extract information that is useful for particular purposes. Compared with the kind of data stored in databases, text is unstructured, amorphous, and difficult to deal with algorithmically. Nevertheless, in modern culture, text is the most common vehicle for the formal exchange of information. The field of text mining usually deals with texts whose function is the communication of factual information or opinions, and the motivation for trying to extract information from such text automatically is compelling—even if success is only partial." As Ian H. from University of Waikato defines[2].

In order to define Data mining we can use some words of the Database Directive. The Directive defines "database" as "a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means". In other part of the directive the lawmaker

[1] Clark J., "Text Mining and Scholarly Publishing", Report Commissioned by the Publishing Research Consortium (PRC), Amsterdam, 2013, p.5.

[2] Text mining, Ian H. Witten , Computer Science, University of Waikato, Hamilton, New Zealand

states that "the term 'database' should be understood to include literary, artistic, musical or other collections of works or collections of other material such as texts, sound, images, numbers, facts, and data". According to European Commission's explanation in various texts "materials" should be interpreted in the broadest sense[3]. Text mining is much narrower than data mining, as it excludes anything other than text. When we use the term Text and Data mining we refer to all types of contents such as text, images, videos, photographs.

A proper definition for TDM should not be closely connected with specific technologies and should be more general in order to be able to adapt to the process: new technologies and techniques arise, the mining methods change with the years in a more sophisticated way, so a definition which describes specific technological processes, quickly will be out of meaning. This is the reason the definition of the International Association of Scientific, Technical and Medical Publishers (STM) is a proper one. STM defines TDM as "Text and Data Mining means to perform extensive automated searches of Publisher's Content, the sorting, parsing, addition or removal of linguistic structures, and the selection and inclusion of content into an index or database for purposes of classification or recognition of relations and associations". The term "automated search" that is used here, is more convenient than other technical terms which are much more specific (ex. Computers, Programmes etc). The definition refers to a number of techniques ("search, sorting, parsing, addition or removal of linguistic structures, and the selection and inclusion of content into an index or database") and scopes without being exhaustive and leaving space for more techniques that may will be used[4].

Another interesting option is the definition of the Report from the Expert Group of European Commission "Text and data mining involves the deployment of a set of continuously evolving research techniques which have become available as a result of widely distributed access to massive, networked computing power and exponentially increasing digital data sets, enabling almost anyone who has the right level of skills and access to assemble vast quantities of data, whether as text, numbers, images or in any other form, and to explore that data in search of new insights and knowledge"[5] .

[3] A. STROWEL & J.P. TRIAILLE, *Le droit d'auteur, du logiciel au multimédia : droit belge, droit européen, droit comparé*, Cahiers du Centre de Recherches Informatique et Droit (CRID), Bruylant, Bruxelles, 1997, p. 260 "La Commission a, à diverses reprises, expliqué que le terme « matière » devait être compris au sens le plus large".

[4]Study on the legal framework of text and data mining (TDM), March 2014 , Jean-Paul Triaille, partner, De Wolf & Partners, lecturer, University of Namur, p.16

[5] European Commission, "Report from the Expert Group, Standardisation in the area of innovation and

Many researchers refer to text and data mining as Content Mining or Data analysis. Data mining is often associated with the term Digital humanities.

History of TDM

The first serious attempts for text mining were made in the mid-1980s; the great development of mining took place the last decades thanks to the increasing innovation to technology.

However, it was much earlier that the interest for mining was obvious to the stakeholders. For many years, professionals and businesses had to face the problem of unstructured data. It is characteristic that in 1958, IBM Journal trying to give a first definition of business intelligence (BI) described a situation where "…Both incoming and internally generated documents are automatically abstracted, characterized by a word pattern, and sent automatically to appropriate action points." It is obvious that the need for organization, extraction and classification of information was a tool that the business world wanted.

As Business Intelligence appeared in 1980-1990, the first attempts focused on numerical data stored in relational databases[6]. Businesses followed the idea that their processes should be measurable and any data worth collecting should be analyzed. It is expected because text as "unstructured" material is hard for analysis and process. Business intelligence focused firstly, on data mining and other techniques of the same type like OLAP, ETL and data warehousing.

As the years passed, market attention has turned to the difficult part of analysis, the text. In late 90's the attention turned from algorithm development to application. Characteristic is how Prof. Mart A. Hearst described the state: "For almost a decade the computational linguistics community has viewed large text collections as a resource to be tapped in order to produce better text analysis algorithms."[7]

As text analytics first appeared in 1990s[8] as "text data mining" or just "text mining", the first text sources were treated as "bags of words." Only basic linguistics was used in order to analyze word formations called technically n-grams. Although only with lexical analysis many functions could be done, like classification of similar

technological development, notably in the field of Text and data mining", 2014.

[6] A Brief History of text Analytics, Seth Grimes, 2007, http://www.b-eye-network.com/view/6311
[7] Untagling Text Data Mining, Marti A.Hearst, School of Information Management and Systems, University of California, Berkley
[8] A Brief History of text Analytics, Seth Grimes, 2007, http://www.b-eye-network.com/view/6311

texts or words, a conceptual analysis was not possible. The first text-mining users were investigators such as intelligence analysts and biomedical researchers trying hard to decode and understand all the offered information making connections among the numerous results.

Concluding, the history of TDM has long way ahead. We are at first steps and much more would be done. The technological innovations give a long perspective for the possibilities of mining.

BIG DATA

The "Big Data" phenomenon as Sergey Filippov refers, has three dimensions: volume, velocity and variety[9]. The quintillions of data produced every day as we abovementioned coming from different sources and with increasing speed day to day, make the management of all these data extremely difficult without the help of technology. Text and data mining plays a very important role as it enables users to classify, analyze data from many different dimensions and points, and manage their interconnection.

It is a worldwide need in our era the use of tools that could contribute to the set-up, analysis, controlling and sharing of information. The existing huge mass of data is in every aspect of life. Most of data are not included into databases and are without structure. It is noticed a great increase of big data that change rapidly provoking big expectations for the upcoming changes and underlining the need for superior management and analysis procedures. The big enterprises do not only need quick access to their information but their needs are focused to more structured data that they can comprehend easily because of the analysis that have been done to them.

ACCESS TO DATA FOR DATA ANALYSIS PURPOSES

Levels of access

Before our analysis go through the legal issues that arise from TDM, we should clarify another option of this topic. There is a distinction between freely accessible data and not freely accessible data. The accessibility to data is extremely important to characterize the applicable legal framework each time.

[9] Mapping text and data mining in Academic research Communities in Europe, the Lisbon Council, Sergey Filippov, 2014, p.2

The freely accessible data on the web are not restricted by any contract. Usually they are found in websites where there is no need for the user to give any consent to specific protective terms and conditions. So, only intellectual property laws in general will regulate the data (text, pictures etc) protection issues[10].

Concerning the restricted access to data, the study of Jean-Paul Triaille in which a triple distinction takes place is the most complete, in my opinion. More specifically, in this category are put social networks, contractual data and confidential data. Social networks data have to do with the private settings of users accounts and the terms of use of the specific platform. The contractual data, are protected by clauses included in a contract and can be used for mining only under specific terms and only the authorized users can access the content. Finally, the confidential data are those which belong to one person and can decide to offer it to another individual or company. Usually parties use Non Disclosure Agreements which contain terms for the use of data.

Public Sector Information

Public Sector Information is connected to TDM as there are many projects at mining text and data held by public administrations. PSI can be defined as the wide range of information that public sector bodies collect, produce, reproduce and disseminate in many areas of activity while accomplishing their institutional tasks. Public Sector Information includes geographical data, statistics, meteorological data, data from publicly funded research projects and digitized books from libraries[11].

The need to more and more reusable documents is highly connected to data mining as the content for mining is enlarged. Private sector tends to reuse public data and this strange relation is regulated by Directive 2003/98 ("PSI I Directive") and Directive 2013/37 ("PSI II Directive"). Reuse can be made for commercial and non-commercial purposes other than the initial purpose of their use, for free or for payment. PSI II Directive completed the first directive by binding public administrations on certain types of digital formats when they send the information and the costs that they ask for this availability, and adding more sectors to the right

[10] Study on the legal framework of text and data mining (TDM), March 2014 , Jean-Paul Triaille, partner, De Wolf & Partners, lecturer, University of Namur, p.20

[11] European Commission, "Digital Agenda: Commission's Open Data Strategy, Questions &Answers" (IP/11/1524), MEMO/11/891, Brussels, 12th December 2011, http://europa.eu/rapid/press-release_MEMO-11-891_en.htm?locale=en

of reuse[12].

Open Access

Text and Data mining would be incomplete if we ignore other underlying trends in scientific communication. Open Access is defined as a comprehensive source of human knowledge and cultural heritage that has been approved by the scientific community. As Berlin Declaration on Open Access defines "Open access contributions include original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material".

Open Access is characterized as very important for TDM procedure as it can free up databases and other resources to analytical exercises[13]. However the use of Open Access techniques is not widespread because of the complexity of making publications available for TDM and of the high cost of the procedure.

The way to do TDM

In order to understand how TDM system works, which part of this system needs protection and with which part there are conflicts we should first identify the basic steps/parts of the TDM procedure. The following text will determine the steps that should be done in order to do TDM. We will describe in a short, easy way these steps as it is impossible to cover all the different technical procedures[14]. The main parts are two: **1.** the existence of a content, which is placed into a data set, repository or collection and **2**. the extraction of information, after the access of the miner to the data and the proper procedure of mining through mining tools. These procedure leads to the final result, to the new information.

1. The existence of a content

[12] Directive 2013/37 of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information, *JO* L 175/1.

[13] Data Mining in the UK Higher Education Institutions: Law and Policy, Andres Guamuz and Diane Cabell, p.14

[14] Diagram from "Data Mining in UK Higher Education Institutions:Law and Policy", Andres Guadamuza and Dianne Cabell

The first prerequisite of mining is the existence of a content which will be mined. The content can be a text, an image, data or other material arranged in a systematic way. Which content and how much of it will be used depends on the reason for doing mining. There are different kinds of miners and different purposes that should be achieved.

Publishers mining supposed to be easier as most of them store all their content in a data repository. Publishers try to keep this content updated in order to be the results as more up-to-date as possible.

"**Knowledge discovery**" is another type of mining. Researchers use this term to define the mining for discovery of completely new, previously unknown and useful information. This procedure needs a wide range of content sources so that it could be as much as possible to find something valuable.

Computational linguistics[15] researchers usually deal with results of text mining and others are searching for new applications of TDM. The first ones need a stable body of content that should be also as large as possible in order to be proved that the tools can be effective and give results from wide range of sources. The second ones need great variety of sources but the important is that new undiscovered content is needed.

Technically speaking, in order to be achieved the mining, the text must be structured in order to be machine-readable. The structure of the text facilitates the extraction and helps to more efficient correlation between the different texts and finally to more and more new information. If the text is in a type that does not fit to the extraction software then it takes more time to conform it to a more convenient type. If the text is in XM format, then it must be conformed to the same Document Type Definition . The Document Type Definition contains explanations for tags used in the XML and other structural elements. The computer program must be able to lead to the meaning of XML.

It is very common that due to the lack of availability of XML content many researchers use PDF documents. The problem with PDFs is that as they need conversion, mistakes can be done during the extraction. However if texts are available in a common format then things are easier. If access is given through an Application

[15] Clark Jonathan, Text mining and Scholarly Publishing, Publishing Research Consortium 2012, p.10

Programming Interface (API)[16] then text mining can be done directly from publisher's database without being important to copy the text to a separate location. It is obvious that the problems of origin of the sources make the procedure more complicated. The more content providers there are the more time consuming this is in order to make all the types of documents available for mining and to make clear the rights and permissions according to copyright.

2. Extraction as a result of analysis

Before anything else we should make a small reference to the most important mining tools which play the major role to the extraction phase. Plenty of tools are available for mining tasks, some of them using artificial intelligence, machine learning and other techniques in order to extract data. It is characteristic that there are also a high number of tutorials and tools explaining how to use the mining tools.

The majority of tools have some basic functions. Some examples of text mining function are the following. Some recognize words as beings in dictionaries, some make the distinction between the different meanings of a word and find the better one, some POS (Part-of-Speech) searchers recognize types of words making connections between them, some use stemming in order to remove common endings to words reducing them to the same stem throughout the text e.g.: "gene" and "genes". Sentiment analysis tools finds notions and feelings words and enlist them as positive, negative or neutral. An example is a Bag-of-Words method with which the text is scanned to be identified the frequency of positive and negative words. The National Centre for Text Mining (NaCTeM) has a freely accessible to the public sentiment analysis test[17].

[16] In computer programming, an application programming interface (API) is a set of routines, protocols, and tools for building software applications. An API expresses a software component in terms of its operations, inputs, outputs, and underlying types. An API defines functionalities that are independent of their respective implementations, which allows definitions and implementations to vary without compromising the interface. A good API makes it easier to develop a program by providing all the building blocks. A programmer then puts the blocks together.

[17] Clark, Text Mining and Scholarly Publishing, 2012, p. 12

Some of the most famous open source TDM tools are (according to famous users sites and researchers)[18]: RAPIDMINER, WEKA, R-PROGRAMMING, NLTK and KNIME.

All the above mentioned tools and generally most of the mining tools follow some basic steps in order to extract information. The first step is when the text or data are reconstructed in order to be read from the software. The mining software recognizes words, set of words, sentences, verbs that create conceptual relations, and morphological variants of words. In this step, the text is cut into pieces and separated in to pieces, ready for the next step.

The second step is the step of extraction. The products of extraction are collected in a database. Usually we start from a template that describes a generalized form of the information to be extracted. The important is the template to be as efficient as possible in extracting meaning. There are also templates that are based on meaning rather than keywords.

APPLICABLE LAW

A. The InfoSoc Directive 2001/29/EC

The InfoSoc Directive is the result of the need of the European Union to create a legal framework able to harmonize and develop the information society issues about copyright and related rights. Technological developments have made more complicated the intellectual property issues and existing legal framework sometimes are unaffected to offer the proper protection. The Directive refers to the legal protection of copyright and related rights in the framework of the internal market, with particular emphasis on the information society.

Article 2 of the Directive is the article describing the reproduction right, as the obligation of Member states to provide the exclusive right to the authors to authorize or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, whole or part of their works[19]. The stakeholders claim that the point of

[18] http://thenewstack.io/six-of-the-best-open-source-data-mining-tools/, http://www.siliconafrica.com/the-best-data-minning-tools-you-can-use-for-free-in-your-company/

[19] Directive 2001/29, art.2 "Member States shall provide for 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

infringement is that of creating copies (from the software) of the content in order to mine. That is the reason that the reproduction right is infringed. Because of this provision, the researchers who need to mine text and data, have to obtain either a license from the rightholders or should be based in a copyright exception. The interpretation of the article should be broad and the definition of reproduction should include every act of reproduction as it is needed in order to be achieved legal stability within the internal market[20]. The European Court of Justice in its Infopaq Judgement supports the broad interpretation of the article and adds that the notion of reproduction is determined technically rather than functionally and moreover that the broad interpretation does not have to do if the act is "transient or irrelevant from an economic or functional perspective"[21]. This broad interpretation includes cases like the transformation of a text from one format to another-usually to XML- and the translation from one language to another.

Although adaptation and translation rights are not clearly referred to this directive, the broad interpretation of the present article permits to be included in the circle of reproduction right all the technical and contextual processes. Concerning translation some make a distinction between the human and the automated. The first one is covered by copyright as the translator has rights on his own work. In the contrary, translation made by a machine is supposed to be a mere reproduction as defined in the InfoSoc Directive[22].

In certain circumstances as in reality there is no copying there is no infringement and application of the directive. Some software technologies, during the TDM do not need to copy the content. They "crawl" through the content and extract the necessary information from one word/data each time.

B. The Database Directive 96/9/EC

It is very often that the content of mining is available only through a database. In the EU, databases are protected by the Directive 1996/9/EC, whose provisions include rules that only some small parts of the content of databases can be used without the permission of the database owner. This provision applies independently of possible

part: (a) for authors, of their works;[…]"

[20] Directive 2001/29, recital 21
[21] European Copyright Law : A Commentary, WALTER, M. W., and VON LEWINSKY, Oxford University Press, 2010, p. 967
[22] Study on the legal framework of text and data mining (TDM), March 2014 , Jean-Paul Triaille, partner, De Wolf & Partners, lecturer, University of Namur, p.32

copyright protection of the database content. According to this Directive, even if the whole content of the database is not under copyright protection, permission for the access to the content should be given by the database owner[23].

The 1996 EU Database Directive grants so-called sui generis rights to those who make a substantial investment in a database through collecting, verifying or presenting the contents. This directive tries to put in order the complex issues of rights associated with collections of data. TDM analysis has a place in this technological part.

The Database Directive is applicable when the data mining process concerns databases as not every article and every book is a database. For the TDM procedure large amounts of text, data, facts, images etc. will be needed. In certain cases the content which will be subject to data mining will itself collect a number of databases such as for example collections of medical publications. So, the database can be protected by copyright or its content may be protected by the sui generis right.

In the first case, when the database is protected by copyright, the directive provides protection to the databases which selection or arrangement of their content constitute the author's own intellectual creation[24]. This part of the directive applies to the act of copying to the selection or arrangement of information in databases and not to the content itself.

We start with **art.5** which provides: "In respect of the expression of the database which is protectable by copyright, the author of a database shall have the exclusive right to carry out or to authorize: (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 ; (c) any form of distribution to the public of the database or of copies thereof. The first sale in the Community of a copy of the database by the rightholder or with his consent shall exhaust the right to control resale of that copy within the Community; (d) any communication, display or performance to the public; (e) any reproduction, distribution, communication, display or performance to the public of the results of the acts referred to in (b)."

In the first paragraph of art. 5 there is the provision about the reproduction of a database temporary or permanently by any means and form in whole or part. Concerning TDM, as we already mentioned, there is difference between TDM process

[23] Text and Data Mining and the Need for a Science-friendly EU Copyright Reform, April 2015, Science Europe Working Group on Research Data Editor: Christoph Bruch, p.6

[24] Database Directive, art.3§1

and TDM output. Concerning the process, if the selection or the arrangement of the data base are copied it means that the whole data included there are copied too. However important role plays the technology and the methods of coping or "crawling" that are used. Sometimes it is possible that the part of the adopted databases will be translated or transformed to other formats in order to be capable for analysis and mining. It is also possible that the structure of the database will change. Even in these cases it is supposed that there is reproduction of the database during the stage of TDM process. One the other hand, data mining output is considered generally as an independent creation, completely different from the initial database and its material and a new original creation.

The communication to the public right exists when the database is communicated to a number of persons, not when it is mined only by a group of researchers or individuals. Concerning the output of TDM, thinking of the output as an independent, new creation, the initial data are not communicated to the public in the same as the output. It is a common sense that the final work is much more complicated and with new patterns, statistics and outputs that are not visible in the initial data[25]. Finally, data mining does not usually permits access to the initial selection or arrangement of the database.

The sui generis right is referred to **art. 7** of the Database directive ("Member States shall provide for a right for 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 to prevent extraction and/or re-utilization of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database.") and exists even if there is or not copyright protection in the database or its content. The sui generis right consists of two parts: the extraction right and the reutilization right.

Art. 7§2 refers to the extraction right and refers "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". Data mining process usually leads to the extraction of all or part of data which are included in a database and this is the reason that the authorization of the rightholder is needed. However, because of different technologies used, can be argued that in some cases the software extracts some parts that are so

[25] Study on the legal framework of text and data mining (TDM), Jean-Paul Traille, De Wolf and Partners, March 2014, p.37

small in number that should be considered as unsubstantial[26]. Moreover, software may only crawls through data, not copying, but counting occurrences or registering links between them. (Court of Justice of the European Union, 9 November 2004, case C-203/02, The British Horseracing Board Ltd and Others v William Hill Organization Ltd, par. 65)

At art. 7§2 we regard the re-utilization right: "any form of making available to the public of 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". The European legislator wanted to give a wide definition of both extraction and re-utilization. The re-utilization right could be associated with the right of communication to the public. Concerning the process of mining there is no communication to the public, so no re-utilisation, if the mining is making by specific researchers, or individuals or a private company. Concerning the output, as we mentioned above, for many cases the output supposed to be a new independent creation.

EXCEPTIONS

Concerning exceptions, in certain countries like USA, the existence of fair use or fair dealing provisions accompanied with relevant case law give solution to many problems. In Europe, however, the situation is more complicated as there is no harmonized provisions for exceptions. the UK remains the only EU member whose copyright law includes a clear exception for TDM. The Copyright Directive still does little things for the harmonization of national copyright provisions[27].

A. As we mentioned above, the InfoSoc directive (2001/29/EC) protects the reproduction right. The directive includes in addition a mandatory exception about the acts of reproduction in temporary way. The EU Member States have to impement this exception in their national copyright legislation. We will start with the **art. 5 (1)** with the general exception from copyright infringement for transient/temporary copies. All the rest exception in art.5 are not mandatory and they do not have an immediate relation to TDM. It is interesting that Israel, New Zealand and Switzerland have based their exception on the EU model, and their exceptions include very similar points. Certain criteria should be met and the law refers to them specifically: the

[26] Study on the legal framework of text and data mining (TDM), Jean-Paul Traille, De Wolf and Partners, March 2014, p.38

[27] Text and Data Mining and the Need for a Science-friendly EU Copyright Reform, April 2015, Science Europe Working Group on Research Data Editor: Christoph Bruch, p.7

temporary copy must be transient or incidental. More specifically transient refers to the limited duration of the act concerning the proper completion of the technological process, which does not include any human intervention[28]. "Incidental" does not focus on time limit but at the fact that the act should be some kind of random to the whole process, like a pop up window that appears for a while including terms of use.

Moreover, the temporary copy has to be an integral and essential part of a technological process, necessary for the correct and efficient function of the process. You should add that human intervention to the process does not affect the fact that the copy is integral and essential part. The copy can be also either to the first or to the last stages of process, as the law does not specify the stage.

In addition copy's sole scope must be: to enable a transmission in a network between third parties by an intermediary or a lawful use of a work or protected subject-matter. Lawful is the use when it is not restricted by the law or authorized by the rightholder. In many cases the courts have decided in favor of the lawful use of information being helped by technological progress[29].

In conclusion, the act must have no independent economic significance[30]. That means that the acts which facilitate the process should not create new economic profits. The Court of Justice supported that if the creator of the reproduction makes a profit of the temporary reproduction or if the temporary reproduction change the reproduced subject matter, then there is an "independent economic significance"[31]. The exception for temporary acts of reproduction is the only mandatory exception in the InfoSoc Directive. As a result, most of the EU countries have adopted the provision, with only exceptions the Netherlands where the law does not include temporary copies and as a result it does not treat them as exceptions and the Belgium where the directive was implemented with some differentiations.

Moreover, we continue with the **art.5§3a** of the Infosoc Directive**:** "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". The exception of article 5 provides an exception to the right of reproduction. As we mentioned above

[28] Case C-5/08, Infopaq I, par.64, European Court of Justice, 2009

[29] Infopaq II, C-302/10, ECJ, par.42-43 and Meltwater case, UKSC 18, UK Supreme Court, 2013

[30] Study on the legal framework of text and data mining (TDM), Jean-Paul Traille, De Wolf and Partners, March 2014, p.42

[31] Study on the legal framework of text and data mining (TDM), Jean-Paul Traille, De Wolf and Partners, March 2014, p.45

TDM almost always include copying of data. The exception for scientific research is not mandatory and has not been adopted in the same way from all the states. In certain states, although the exception was not implemented completely, the case law integrated the exception. The non common legal framework for the exception, puts barriers to the users of TDM. The implementation of the exception usually combines in one provision the teaching and research issues without treating them differently. As we can imagine, in order to "illustrate" an educational work we need specific parts of the content but for the scientific research it is not the same. Some countries have a clear distinction for the research and the education. A characteristic example is UK.

Another issue about this exception is that there are no mentioned specific groups of people that can be benefited from it. As a result, anyone could allege the profit of the exception, even if he/she comes from irrelevant professional background, or has no connection with a research center/institution. However some countries have more specific legislation: for example in Poland the exception does not apply to individual researchers but only to to research and educational institutions[32]. Moreover, each state has differentiations concerning the matter of exception. The countries have different content concerning the "works" and there is a tendency for limitation of the number of works used, something that provoques problems to mining as a large number of works is needed.

One of exceptions prerequisites is that the user can prove that the source, and the author's name, are included, unless this is not possible. Concerning TDM, the issue is more complicated as the research process include copies of the works but the final output is a total of new outputs.

We should add the fact that the art.5§3a includes only these acts that have a non-commercial purpose and are justified by this purpose. In order to define the meaning of non-commercial in this article we can refer to recital 42 of the InfoSoc Directive : "the non-commercial nature of the activity in question should be determined by that activity as such. The organizational structure and the means of funding of the establishment concerned are not the decisive factor". In order to understand better the distinction between the "decisive factor" and "the activity as such" we will keep in mind the characteristic example of a private company financing the research for a philanthropic purpose or a company financing a research the results of which will be

[32] Study on the legal framework of text and data mining (TDM), Jean-Paul Traille, De Wolf and Partners, March 2014, p.55

public or published in open access. It has been stated in many cases that the criterion of commercial-non commercial is often hard to apply.

B. Concerning the Database Directive, there is an exception to art. **6§1**: "The performance by the lawful user of a database or of a copy thereof of any of the acts listed in Article 5 which is necessary for the purposes of access to the contents of the databases and normal use of the contents by the lawful user shall not require the authorization of the author of the database. Where the lawful user is authorized to use only part of the database, this provision shall apply only to that part." The present exception is the only one in database directive which is compulsory for member states and also cannot be waived by a contract. In this directive we cannot find any definition for the term "lawful" user although it is a significant issue. The Explanatory Memorandum gives a definition of the lawful user as a person having acquired the right to use a database[33]. To the lawful users can be included the users who have access to content (see above analysis) relying on statutory or contractual exceptions provided by law or by contract, the licensees, and the lawful acquirer. Concerning the meaning of "access" and "normal use", in order to better understand the meaning of the provision we will refer to paragraph 34 of the Explanatory Memorandum which underlines that "Whereas, nevertheless, once the rightholder has chosen to make available a copy of the database to a user, whether by an on-line service or by other means of distribution, that lawful user must be able to access and use the database for the purposes and in the way set out in the agreement with the rightholder, even if such access and use necessitate performance of otherwise restricted acts."[34] In simpler way, normal use includes the provisions of the agreement between the rightholder and the user concerning the scope, the use and the access to the content.

Moreover there is also, the exception under **art.8 par.1**: "The maker of a database which is made available to the public in whatever manner may not prevent a lawful user of the database from extracting and/or re-utilizing insubstantial parts of its contents, evaluated qualitatively and/or quantitatively, for any purposes whatsoever. Where the lawful user is authorized to extract and/or re-utilize only part of the database, this paragraph shall apply only to that part". This paragraph has an illogical provision. It is not a necessary provision as it is clear that a user may use insubstantial parts of a database without permission of the database maker. This is because the rights of extraction and re-utilisation only apply when the user extracts or reuses substantial parts[35].

[33] DERCLAYE, E., "The Legal Protection of Databases: A Comparative Analysis", Edward Elgar, 2008, p. 120

[34] Database Directive (96/9/EC), Explanatory Memorandum, par.34

[35] EU Copyright Law: A Commentary, Irini Stamatoudi, Paul Torremans, p. 335

Art 8 is different from **6.1**. The last one is more restricted than art.8 as it refers to "normal use" and not to "any purpose". The prerequisite of the article 8 of making available to the public "in whatever manner" shows that the provision does not apply if the database has not been made public. The lawful user has the same meaning as we mentioned above. The terms "insubstantial part" and "quantitative" and "qualitative" are not defined in the Database Directive. Following case law, we are borrowing the interpretation that European Court of Justice tried to give to that terminology[36]. In a quantitative point of view, a part can be characterized substantial assessing the volume of data extracted from the database and/or re-utilized in relation to the volume of the contents of the whole of that database. In a qualitative point of view, a substantial part of the content of a database may in fact represent, in terms of obtaining, verification or presentation, significant human, technical or financial investment.

We continue with the exception of **art. 7§5:** "The repeated and systematic extraction and/or re-utilization of insubstantial parts of the contents of the database implying acts which conflict with a normal exploitation of that database or which unreasonably prejudice the legitimate interests of the maker of the database shall not be permitted."

As data analysis usually includes successive and systematic extractions of a database article 7§5 in reality is an exception to art.8§1. More particularly, the article refers to the users' acts which reconstitute through extraction/reutilization, the whole or substantial part of the content of a database protected by sui generis right or making available to the public and effect the investment made by the maker of the database. The accumulation of insubstantial parts is equivalent to a substantial part. Moreover, as explained by E. Derclaye, the harm caused cannot be hypothetical, it must exist. But we should assess each case separately as there are doubts about the extend of damage to the database maker.

Art. 9b of the Database Directive includes one more exception for scientific research to the sui generis right : "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: [...] (b) in the case of extraction for the purposes of illustration for teaching or scientific

[36] ECJ, 9 November 2004, case C-203/02, *The British Horseracing Board Ltd and Others v William Hill Organization Ltd., par.* 68

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

Contrary to the exceptions to copyright abovementioned in the Article 5(3)(a) of the Infosoc Directive and in Article 6(2) of the Database Directive, Article 9(b) does not have as the prerequisite of "sole" purpose of illustration for scientific research. This means that the exception covers more cases. From the definition it is not clear if illustration refers only to teaching or scientific research also. If "illustration" relates to both teaching and scientific research the exception is narrower than if it relates only to teaching. Concerning the source that should be indicated, contrary to abovementioned exceptions, in art.9 there is no condition of "this turns out to be impossible". This does not let any choices to the users. Except the identity of the maker of the database as source could also be the URL address or the name of the database[37]. Finally, the criterion of non-commercial has the same meaning as at the other exceptions.

LICENSING

Another substantial part in the protection of right holders in TDM is the license agreements. It is very common that the users ask for a license in order to take permission and have access to the content for the aims of text and data mining. Those wishing to text and data mine within the rules must agree contracts with the rightholders, and sometimes pay a fee.

Rightholders like publishers, have in their hand licenses controlling the mining of subscribed content in a more detailed way. The licenses include permissions and prohibitions which regulate the relation between rightholder and miner. Researchers must find the rightholders and then find the way to ask and obtain the permission. The majority of rightholders usually judge each case separately and take the decision case by case. The obtaining of a license is influenced by information about the protection of the content from copyright and the identity of the copyright owner. In their study Smit and van der Graaf report that over 90% of the publishers give permission for the research-focused mining requests they receive. Moreover, 32% of the publishers give permission for all types of mining , usually under their Open Access policy[38].

[37] Study on the legal framework of text and data mining (TDM), Jean-Paul Traille, De Wolf and Partners, March 2014, p.82.

[38] Journal Article Mining, A research study into Practices, Policies, Plans. and Promises, Commissioned by the Publishing Research Consortium by Eefke Smit and Maurits van de Graaf,

There are however some problems concerning the procedure of licensing. Sometimes obtaining a license have additional costs and also there is the risk that licenses will not be granted and the use of the output of mining will be restricted. Moreover, in some cases the acceptance of a mining license hides the risk of leaving part of the rights that the law offers[39]. It is characteristic that in some cases licenses, coupled with subscription contracts, limit TDM to the subscribed content. Licenses may put restrictions on the user that either expressly or inferentially bar TDM without permission. Even Creative Commons licenses may put burdens for researchers if the license does not permit derivative works[40]and the procedure to find all the rightholders is complex. In many cases the whole licensing procedure is too time-consuming and not able to cover practical needs. As the researchers usually are focusing on content that is owned by different rightholders, they have to overcome a difficult situation of discussing licensing terms with many people; usually almost an impossible achievement.

Concerning text mining more precisely, it is common certain publishers to oblige through the licenses the researchers to use for TDM, servers controlled by them and special software also installed by them. This approach burdens the way of mining of the researcher and it also exposes the interests and algorithms to the publisher. The publisher may regulate how the researcher can share and publish the results of that mining.

The International STM Association have made a model[41] including provisions for the licensed uses in TDM, prohibited uses relating to subscribed content, TDM output, security, grant of access rights, formats and delivery mechanisms[42]. STM is the leading global trade association for academic and professional publishers focusing to licensing as "Licensing (individually and collectively) is the 21st century's answer to legal access to copyright-protected works."[43]. According to STM policies, licensing is important

Amsterdam,2011,http://www.publishingresearch.net/documents/PRCSmitJAMreport20June2011VersionofRecord.pdf.

[39] Text and Data Mining and the Need for a Science-friendly EU Copyright Reform, April 2015, Science Europe Working Group on Research Data Editor: Christoph Bruch, p.6

[40] Securing Tekst Mining Rights for researchers in academic Libraries, Hillary K. Miller,p.21

[41] STM Statement on Text and Data Mining and Sample License http://www.stm-assoc.org/text-and-data-mining-stm-statement-sample- licence/.

[42] Licenses For Europe – Working Group 4 – TEXT AND DATA MINING, 17 April 2013, STM

[43] Submission on the Issues Paper "Copyright and the Digital Economy", STM, www.stm-assoc.org

to overcome copyright obstacles in TDM. It is expected that if copyright protection is restricted, third parties take the economic advantages with the result of lower reinvestment in good quality content and access.

Moreover, some other publishers have established a process for individual researchers in order to have the permission to text mine with some restrictions. Characteristic example is the SpringerOpen Access license agreement, as a creative common public license[44], which is a detailed license with specific terms and conditions for the use. Publishers Licensing Society (PLS) in UK has made an effort to concentrate the widest possible variety of content in order to be the common place where rightholders and potential users will be in contact for acquiring a permission[45]. Moreover, CrossRef has introduced a Common Content Server as a way to make easier the access to content. Publishers have to send XML of the content to a specific warehouse and from there researchers could search the data they want to mine with the help of simple information retrieval tools. This technique gives the chance to less powerful publishers to make easily their content available for TDM. PLOS also, decided that it will give to creators the right to sign a data availability statement that will make sure that when the creation will be published, all the content will be available to anyone publicly, except specific exceptions. Moreover, researchers at academic institutions were given the possibility to download documents in computer readable format (XML) in batches of up to 10000 articles per months; this was announced by Reed Elsevier.

We should add that although there are article abstracts freely available, which are open for TDM, text and data behind paywalls are not, even when institutions have paid for a site licence. "The licence is oriented towards permitting the human to download and read an article, but not to text-mine it," says John McNaught, the director of the National Centre for Text Mining at the University of Manchester. It is characteristic that even freely accessible papers may not come with permissive licences: of the 2.4 million abstracts listed by PubMedCentral, only 400,000 (17%) are licensed for text-mining[46].

Some scientific publishers have made suggestions on the clauses of licenses in order to allow access for mining reasons, in a more convenient way. However, researchers

[44] http://www.springeropen.com/pdf/Creative_Commons_Attribution_4.0_International_CC_BY_4.0.pdf

[45] http://www.pls.org.uk/about/vision,-aims-and-values/

[46] Trouble at the text mine, Richard Van Noorden, NATURE, VOL 483, 8 March 2012, p. 135

remain suspicious, demanding complete access to public domain databases, and supporting that 'the right to read is the right to mine'. Researchers support that efficient research needs the freedom of access. That is the reason that 'Open access' movement is becoming more and more popular. Two countries that have made progress having special copyright legislation, are UK and Ireland. Finally, the Open access movement ideas were discussed at the EU`s Horizon 2020 about innovative legal framework.

THE EFFECT TO THE USE TDM

The use of TDM in Europe is much lower than the use in USA[47]. Lisbon Council after a study, came to the conclusion that very few European researchers know or use TDM[48]. The European legal framework in comparison with other parameters play their own role. It is obvious that the copyright legislation in Europe, with the variety of exceptions, is not as efficient as the "fair use" system of USA, which presents more security to rightholders.

On the other hand, in Europe, all the legislation provided has not reached to harmonize legislation relevant to TDM in all the EU member states and the complex requirements for the exceptions make the legal TDM more difficult. Moreover, for groups and companies that are not interested in scientific research there are more barriers. In the exceptions is not included text and data mining for marketing purposes. More particularly, mining for marketing purposes will fall into the exception only if research institutions mine and not focusing on commercial marketing of new products.

In the InfoSoc directive, the interpretation of the "sole" makes narrower the group of stakeholders. Concluding, as for some cases the mentioning of sources does not have the excuse of "impossible", if the sources cannot be found the mining project has to stop. Another problematic effect is that the clauses of a contract can put aside the exceptions for scientific research. As a result, the fear of researchers for the legal consequences has led to a very careful and conservative approach to mining.

Except the legal issues there are also a number of technical issues that arise questions for the stakeholders in order to be improved the conditions for TDM and

[47] European Commission, "Report from the Expert Group, Standardisation in the area of innovation and technological development, notably in the field of Text and data mining", 2014, p. 35

[48] Mapping the Use of Text and Data mining in Academic and Research Communities in Europe. Lisbon Council, Filoppov

the costs. Certain publishers and businesses focus problems because of the ongoing technological innovations which provoke serious changes to the functional way of their company`s plans[49].

For all these reasons European Commission announced plans for new legislation specially for Text and Data mining. This plan includes a mandatory copyright exception for text and data mining. Researchers in Europe will have permission to use computer programs to search journals, a practice which is tightly controlled by publishers. The last years Research Commissioner, Carlos Moedas has been pushing for the right for researchers to mine papers unhindered, and for the opening to access to the outputs of publicly-funded research[50]. However, there are some opponents of this exception who support that the efforts of EU Commission are unsufficient. The League of European Research Universities (LERU) supports that the legislative program is not addressed in a "more convincing and coherent way" and the commission had to think more carefully about the new proposals.

As we already mentioned, text and data mining is a new entry that can be ameliorated through more and more factual and market information, as the current situation of TDM does not match to the needs of rightholders. For many writers the solution would be a new, clear exception in copyright and database laws specifically for text and data mining, including also a clear definition of TDM and having its own characteristics. Moreover, many other changes should be done like a control of licenses and the way of giving them, a normative interpretation of the so called reproduction right and of course more serious thoughts about the "open norm", which will help the justice to be more flexible with the miners.

CONCLUSION

Nowadays, there is a great increase of information coming from different sources and made available in different ways. Digital information nowadays is huge and the digitized traditional archives are added to these great mass of information. The need of a way to read, organize and analyze all this information has lead to technological innovations that make the life of researchers, institutions and individuals easier. TDM is a recent method of extracting information from text and data. Thanks to TDM

[49] European Commission, "Report from the Expert Group, Standardisation in the area of innovation and technological development, notably in the field of Text and data mining", 2014, p.5

[50] EU Commission sets out plan to allow free data mining, Éanna Kelly, Science-Business, 10 December 2015

complex content which the human eye cannot read and analyze efficiently, can now be analyzed on time and by using resources from a wide range of data and, articles, books etc. Text and Data mining gives also the change to humans, to reveal new connections between information and data, which researchers were not able to discover themselves. As we mentioned there are many different definitions on TDM, some of them cover all the technological functions of mining, while others are more related to the processes used to mine.

This paper focuses on how the copyright and sui generis database provisions function concerning TDM. The legal framework is completed in several different ways, through licensing process, or through exceptions. The abovementioned analysis gave us knowledge about the basic legal framework which protects and restricts stakeholders' rights and actions and made us conclude that many efforts have to be done as the needs are increasing and the technological and economical changes make things more complicated.

Concluding, difficulties concerning the access to the content, time-consuming procedures, problems with the legal framework and lack of skilled miners may be the reasons for the low rates of use of TDM in Europe. As a solution in order to be encouraged TDM, many researchers propose the adoption of a new copyright and database exception especially for TDM as the existing legal framework has not been implemented in the right way in certain EU countries. This has created uncomfortable situations in EU. Recently the EU Commission set out plan to allow free text and data mining. TDM is a sector with increasing rates and promising innovations. Mining applications are full of potentialities in many domains and this is the reason that investments in the legal and technological sector have to be done for sustainable TDM in the future.

BIBLIOGRAPHY

AIPPI Japanese Group, Exceptions to Copyright protection and the permitted Uses of Copyright works in the hi-tech and digital sectors, Question Q216B, https://www.aippi.org/download/commitees/216B/GR216Bjapan.pdf (2011).

Ananiadou, S., "Text Mining, IPR, Derived Data and Licensing", on behalf of NaCTeM, http://ec.europa.eu/licenses-for-europe-dialogue/node/7.

Ananiadou S., "The National Centre for Text Mining: A Vision for the Future", October 2007, http://www.ariadne.ac.uk/issue53/ananiadou.

Battisti M., "Le datamining, prochaine exception au droit d'auteur", Paralipomenes 2011, http://www.paralipomenes.net/archives/6490.

Borghi M., & Karapapa S., "Copyright and Mass Digitization: a Cross-Jurisdictional Perspective", Oxford University Press, 63 (2013).

Case : British Horseracing Board Ltd v. William Hill Organization Ltd (BHB decision), C-203/02, [2004] ECR I-10415

Case: Fixtures Marketing Ltd v. Sverka AB, C-338/02, [2004] ECR I-10497

Clark J., Text Mining and Scholarly Publishing, PRC (2012)

Clark J., "Text Mining and Scholarly Publishing", Report Commissioned by the Publishing Research Consortium, 3 (2013).

Colcanap G. & Perales C., "CSPLA – Mission relative au data mining (exploration de données): l'analyse de Couperin et de l'ADBU" http://adbu.fr/wp-content/uploads/2014/04/Audition_CSPLA_TDM_2014_04_04_final.pdf (2014).

Copyright Clearance Center, TDM Pilot Program, User Guide for CCC's Text and Data Mining Service (2013).

Copyright Review Committee, "Copyright and Innovation: a Consultation Paper", for the Minister for Jobs, Enterprise and Innovation, Dublin http://www.djei.ie/science/ipr/crc_consultation_paper.pdf (2012).

Commission of the European Communities. DG INTERNAL MARKET AND SERVICES WORKING PAPER: First evaluation of Directive 96/9/EC on the legal protection of databases (2015)

Davison J. M. & Hugenholtz B. P., Football fixtures, horse races and spin-offs: the ECJ domesticates the database right, EIPR 2005, Issue no. 3.

Depreeuw S. & Hubin J.B., J.-P., TRIAILLE (ed.), Study on the territoriality of the making available right. Localisation of the act of making available to the public and its consequence in the Application Of Directive 2001/29/EC On Copyright And Related Rights In The Information Society.

Depreeuw S., De uitzondering voor « tijdelijke technische reproductiehandelingen » na Infopaq I en II en Premier League, A&M, 76-85 (2013).

Derclaye E., The Legal Protection of Databases: A Comparative Analysis, Edward Elgar (2008).

European Commission, "Report from the Expert Group, Standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining" http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf (2014).

Europe PubMed Central labs, Copyright, http://europepmc.org/Copyright .

Fan W., Wallace L., Rich S. & Zhang Z., Tapping into the Power of Text Mining, (2005).

Filippov S., "Mapping the Use of Text and Data mining in Academic and Research Communities in Europe", Lisbon Council, Brussels, file:///C:/Users/Chris/Downloads/LISBON_COUNCIL_Mapping_Text_and_Data_Mining.pdf .

Frawley J.W., Piatetsky-Shapiro G. & Matheus J.C., Knowledge Discovery in Databases: An Overview (1992).

Guadamuz A. & Cabell D., "Data Mining in UK Higher Education Institutions: Law and Policy", 2014, Queen Mary Intellectual Property Review 4:1, 3-29 http://ssrn.com/abstract=2446447 (2014).

Guadamuz A. & Cabell D., "Analysis of UK/EU Law on Data Mining in Higher Education Institutions" http://ssrn.com/abstract=2254481 (2013).

Guernsey L., Digging For Nuggets Of Wisdom (2013).

Guibault L., "Trouver le diamant dans la mine de données ou les implications juridiques de l'exploration de données", Documentaliste – Sciences de l'information, vol. 51, n. 2, http://www.ivir.nl/publicaties/download/1404 (2014).

Guibault L., "Intellectual property rights' obstructions to text and data mining", http://www.youtube.com/watch?v=hfpkJs6GJUgwatch?v=hfpkJs6GJUg.

Hargreaves I., "Digital Opportunity, A Review of Intellectual Property and Growth" http://www.ipo.gov.uk/ipreview-finalreport.pdf (2011).

Hearst A. M, What is Text Mining?,SIMS, UC Berkeley (2003).

Hearst A. M., Untangling Text Data Mining (2012).

Hellwig F., "Change in Copyright Law as a Market Intervention to Realize the Welfare Potential of Text Mining in Scientific Research" http://ssrn.com/abstract=2386238 (2013).

Hillary K. Miller. Securing Text and Data Mining Rights for Researchers in Academic Libraries

IFLA, "Statement on data mining", http://www.ifla.org/publications/ifla-statement-on-text-and-data-mining-2013.

International Council for Scientific Information (ICSTI), Text and Data Mining (2009).

JISC Collections, http://www.jisc-collections.ac.uk/nesli2/NESLi2-Model-License

JISC, Designing a licensing strategy for sharing and re-use of geospatial data in the academic sector (2007).

JISC, Use Case Compendium of Derived Geospatial Data (2005).

Jusoh S. & Alfawareh M. H., Techniques, Applications and Challenging Issue in Text Mining (2012).

Kretschmer M., Deazley R., Edwards L., Erickson K., Schafer B., Zizzo D. J., "The European Commission's Public Consultation on the Review of EU Copyright Rules: A Response by the CREATe Centre" European Intellectual Property Review, 36(9), 547-553, http://ssrn.com/abstract=2509186 (2014).

Langlais P. C. & Maurel L. (Savoirscom1), "Quel statut légal pour le content mining? – synthèse de Savoirscom1, consécutive à l'audition du 15 janvier 2013 par la Conseil Supérieur de la Propriété Littéraire et Artistique", http://www.savoirscom1.info/wp-content/uploads/2014/01/Synthe%CC%80se-sur-le-statut-le%CC%81gal-du-content-mining.pdf .

Liber Europe, "The Perfect Swell: defining the ideal conditions for the growth of text and data mining in Europe. A report from a workshop held at the British Library" http://libereurope.eu/wp-content/uploads/TDM%20Workshop%20Report%5B1%5D.pdf (2013).

Linder B. & Shapiro T., Copyright in the Information Society, Edward Elgar, Cheltenham (2011).

Martin J. - CSPLA, "Mission sur l'exploration de données ("Text and Data mining")" (2014), http://www.savoirscom1.info/wp-content/uploads/2014/10/Rapport_Text_and_Data_Mining_exploration_de_donn%C3%A9es.pdf

Maximilian Haeussler, Jennifer Molloy,Peter Murray-Rust and Charles Oppenheim Responsible Content Mining, June 16, 2015

Michaux B., Droit des bases de données, Kluwer (2005).

Mooney J. R. & Bunescu R., Mining Knowledge from Text Using Information Extraction, http://www.cs.utexas.edu/~ai-lab/pub-view.php?PubID=51469 .

Miner G., D. Delen, J. Elder, A. Fast, T. Hill, and R. Nisbet Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications, Elsevier, January 2012

OECD, "Exploring data-driven innovation and new sources of growth: mapping the policy issues raised by Big Data", OECD Digital Economy Paper No 222 (2013).

Okerson A., Text & Data Mining - A Librarian Overview (2013).

Reichman H. J. & Okedigl L. R., When Copyright Law and Science Collide: Empowering Digitally Integrated Research Methods on a Global Scale, Minn. L. Rev., I.I.C., vol.43 http://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=5351&context=faculty_scholarship (2012).

Rust M. P, The Right to Read Is the Right to Mine, Open Knowledge Foundation Blog http://bit.ly/O75Rwd (2012)

Reilly F.Bernard, Jr. Text Mining and Libraries: Summary of a conversation with Publishers

Science Europe Working Group. Text and Data Mining and the Need for a Science-friendly EU Copyright Reform, April 2015

Smit E. & Van Der Graaf M., Journal article mining: A research study into Practices, Policies, Plans…..and Promises (2011).

Smit E. & Van der Graaf M., "Journal article mining: the scholarly publishers' perspective", Learned Publishing vol. 25 no. 1, 36 (2012).

Stamatoudi I., Cultural Property Law and Restitution. A Commentary to International Conventions and European Union Law, Edward Elgar Publishing, (2011).

Stamatoudi I. (ed), Copyright and the Digital Agenda for Europe: Current Regulations and Challenges for the Future, Sakkoulas Publications (2015).

Stamatoudi I. & Torremans P., (eds) Copyright in the New Digital Environment: The Need to Redesign Copyright, Sweet & Maxwell (2000).

Stamatoudi I. & Torremans P., European Union Copyright Law, Edward Elgar Publishing (2014).

STM, Statement Sample License Text Data Mining, http://www.stm-assoc.org/text-and-data-mininghttp://www.stm-assoc.org/text-and-data-mining-stm-statement-sample-licencestm-statement-sample-license.

STM, Text and Data Mining Sample Subscription (2012).

STM, Sample License for Text and Data Mining of subscribed copyright-protected works and materials (2013).

STM, "Submission on the Issues Paper, Copyright and the Digital Economy (Australia) (2012).

STM, Text and Data Mining Sample Subscription, http://www.stmhttp://www.stm-assoc.org/2012_03_15_Sample_Licence_Text_Data_Mining.pdfassoc.org/2012_03_15_Sample_License_Text_Data_Mining.pdf (2012).

Strowel A. & Trialle P.J, Le droit d'auteur, du logiciel au multimédia : droit belge, droit européen, droit comparé, Cahiers du Centre de Recherches Informatique et Droit (CRID) (1997).

Triaille J. (de Woolf partner), with de Meeus d'Argenteuil J. and with the collaboration of de Francquen A., "Study on the legal framework of text and data mining (TDM)", http://ec.europa.eu/internal_market/copyright/docs/studies/1403_study2_en.pdf.

Triaille J., Study on the Application of D. 2001/29/EC on Copyright and Related Rights in the InfoSoc Directive, 2013.

Truyens M. & Van Eecke P., "Legal aspects of text mining", to be published in CLSR http://www.lrec-conf.org/proceedings/lrec2014/pdf/452_Paper.pdf (2014)

Universities UK, UK Higher Educational International Unit, "European Commission's Stakeholder Dialogue "Licenses for Europe and Text and Data Mining"", http://www.international.ac.uk/media/2243028/Briefing%20-%20Licenses%20for%20Europe%20and%20Text%20and%20Data%20MiningREVISED.pdf.

Van Noorden R., Trouble at the text mine (2012).

Walter W. & Von Lewinsky S. V., European Copyright Law: A Commentary, Oxford University Press (2010).

Weiss M.S, Indurkhya N. & Zhang T., Fundamentals of Predictive Text Mining, Texts in Computer Sciences, Springer (2010).

X., Questions, What are the arguments in favor of using ELT process over ETL?, Stackexchange.com, http://dba.stackexchange.com/questions/19242/what-are-the-arguments-inhttp://dba.stackexchange.com/questions/19242/what-are-the-

arguments-in-favor-of-using-elt-process-over-etlfavor-of-using-elt-process-over-etl

Irini
Stamatou
di

Digitally signed by Irini
Stamatoudi
DN: c=GR, st=Thessaloniki,
l=Thermi, o=International Hellenic
University, ou=School of
Economics & Business
Administration, cn=Irini
Stamatoudi,
email=stamatoudi@syrigos.gr
Date: 2016.02.15 11:04:37 +02'00'